



Building Bioinformatics Solutions for Biomarker Identification

DARREN OAKLEY

PhD Thesis

August 2008

Supervisor: Dr. C. Bessant

Academic Years: 2004 – 2008

Cranfield Health, Cranfield University

This thesis is submitted for the degree of Doctor of Philosophy

Abstract

This thesis describes the design, implementation and application of bioinformatics systems to aid work in the field of biomarker discovery and diagnostic test development. The aim of the work was to develop a flexible data storage and analysis platform that would be capable of housing and working with data from a variety of modern biomarker analysis techniques.

In order to achieve this aim, several tools were developed: a flexible database schema, taking ideas from the field of systems biology, was developed with the goal of being flexible enough to house information about experiments looking at targets such as genes, proteins and metabolites; an API was created to allow easy programmatic interaction with the database; and multivariate data analysis routines were prepared so that data imported into the database could be investigated. Together this toolset was named XPA (for 'Cross Platform Analysis').

The XPA system was tested by using it to house and analyse data from two different medical studies, one using quantitative PCR (qPCR) to observe gene expression changes in prostate cancer, and the second using surface enhanced laser desorption/ionisation mass spectrometry (SELDI MS) to generate protein profiles in sufferers of pre-eclampsia. In both studies XPA was used to develop multivariate classification models using partial least squares discriminant analysis (PLS-DA) and support vector machines (SVMs), with the aim of evaluating the data acquired for potential diagnostic use. The results showed the benefit of a tool such as XPA to the field of biomarker discovery.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Conrad Bessant for his guidance during this work and for giving me the opportunity to carry out my PhD studies.

I would also like to thank all of the staff and students within Cranfield Health and the Bioinformatics group for their friendship and support over the course of my studies. Jeremy Austin, Dr. Luca Bianco, Dr. Michael Cauchi, Paul Knight, Dr. Judit Kumunthini, Dr. Lee Larcombe, Jennifer Mead, Fady Mohareb, Dr. Ian Shadforth, Andrew Spooner and Dr. Weibing Xu are all notable for their help with keeping my sanity over the last few years.

Thanks also to the staff at St. George's Medical School, University of London (SGUL) for supplying the project with data when it was in great need; particular thanks go to Dr. Gary Coulton, Dr. Christodoulos Pipinikas and Prof. Guy Whitely.

I would also like to thank the staff at Unilever R&D for their input to the project during its early stages; Yeyejide Adeleye, Dr. Guy Warner and Dr. Andrew White.

Finally, I would like to thank Catherine, my family, and Catherine's family, without whose never-ending patience and support I would never have finished the thesis – thank you.

Abbreviations

Abbreviation	Expansion
ANN	Artificial Neural Network
API	Application Programming Interface
CV	Cross Validation
DBI	Database Interaction
ELISA	Enzyme-Linked ImmunoSorbent Assay
MAD	Median of the Absolute Deviations
MS	Mass Spectrometry
M/Z	Mass / Charge Ratio
PCA	Principal Component Analysis
PERL	Practical Extraction and Reporting Language
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminate Analysis
RMSEC	Root Mean Square Error of Calculation
RMSECV	Root Mean Square Error of Cross Validation
RT-PCR	Real Time Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
SD	Standard Deviation
SELDI MS	Surface Enhanced Laser Desorption/Ionisation, Mass Spectrometry
SELDI-TOF MS	Surface Enhanced Laser Desorption/Ionisation, Time of Flight, Mass Spectrometry
SIMCA	Soft Independent Modelling of Class Analogy
SQL	Structured Query Language
SVM	Support Vector Machine
XPA	Cross-Platform Analysis

Contents

ABSTRACT.....	I
ACKNOWLEDGEMENTS.....	II
ABBREVIATIONS.....	III
CONTENTS.....	IV
FIGURES.....	VII
TABLES.....	XII
1 INTRODUCTION	1
1.1 BIOMARKER DISCOVERY.....	2
1.2 GENOMICS	3
1.3 TRANSCRIPTOMICS	4
1.4 GENOMICS AND TRANSCRIPTOMICS IN BIOMARKER DISCOVERY	5
1.4.1 <i>Genomic Copy Number Variation</i>	6
CNVs as Biomarkers	7
CGH and Array CGH	7
1.4.2 <i>Single-nucleotide Polymorphisms</i>	12
SNPs as Biomarkers.....	12
1.4.3 <i>Gene Expression Profiling</i>	14
Gene Expression Profiling as a Biomarker	14
Quantitative PCR.....	15
DNA Microarrays	20
1.5 PROTEOMICS.....	26
1.6 PROTEOMICS IN BIOMARKER DISCOVERY	26
1.6.1 <i>Mass Spectrometry</i>	27
Concepts of Mass Spectrometry	28
MALDI-TOF Mass Spectrometry.....	29
SELDI-TOF Mass Spectrometry	33
Data Obtained from Mass Spectrometry	36
1.7 DATA MANAGEMENT FOR MODERN LABORATORY TECHNIQUES	39
1.7.1 <i>Data Integration</i>	40
1.8 WORKING WITH DATA FROM MODERN LABORATORY TECHNIQUES	45
1.8.1 <i>Multivariate Statistical Methods</i>	46
Principle Components Analysis.....	46
Partial Least Squares Discriminate Analysis	49
1.8.2 <i>Machine Learning</i>	52
Artificial Neural Networks.....	53
Support Vector Machines	56

1.9	PROJECT AIMS	60
2	DEVELOPMENT OF A HETEROGENEOUS DATA MANAGEMENT SYSTEM AND ANALYSIS PLATFORM FOR BIOMARKER IDENTIFICATION AND ANALYSIS.....	62
2.1	A DIFFERENT APPROACH: XPA – CROSS-PLATFORM ANALYSIS.....	63
2.2	XPA SYSTEM OVERVIEW	65
2.2.1	<i>Class and Database Structure.....</i>	<i>67</i>
	Investigation	69
	Experiment.....	69
	Protocol.....	69
	Patient.....	70
	Dataset.....	70
	Molecule.....	71
	Ontology.....	72
	The XPA Database.....	73
2.3	DATA HANDLING.....	75
2.4	DATA ANNOTATION	78
2.4.1	<i>Annotator.....</i>	<i>78</i>
	Programming Specifics.....	81
2.4.2	<i>Annotator within XPA.....</i>	<i>81</i>
2.5	DATA ANALYSIS WITHIN XPA	82
2.5.1	<i>Exploratory Analysis.....</i>	<i>84</i>
2.5.2	<i>Developing Classification Systems</i>	<i>85</i>
2.6	CASE STUDIES.....	85
3	EARLY PREDICTION OF PROSTATE CANCER FROM QPCR DATA.....	86
3.1	TRADITIONAL METHODS OF DIAGNOSING PROSTATE CANCER	87
3.2	DEVELOPING DIAGNOSTIC TOOLS FOR PROSTATE CANCER WITH QPCR AND XPA	88
3.2.1	<i>Patient Recruitment and Groups.....</i>	<i>89</i>
3.2.2	<i>Laboratory Methodology</i>	<i>89</i>
3.2.3	<i>Data Handling and Analysis.....</i>	<i>91</i>
3.2.4	<i>Classification Models using qPCR Data in XPA.....</i>	<i>92</i>
	Statistical Methods.....	92
	Machine-Learning.....	95
	Performance Metrics.....	96
3.3	RESULTS	98
3.3.1	<i>Data Investigation with PCA</i>	<i>98</i>
3.3.2	<i>Classification – PSA Alone.....</i>	<i>103</i>
3.3.3	<i>Classification – PSA with Gene Expression Data.....</i>	<i>106</i>
3.4	DISCUSSION AND CONCLUSIONS.....	109
3.4.1	<i>Expected PSA Test Sensitivity vs. SVM and PLS-DA models</i>	<i>109</i>
3.4.2	<i>Comparison of the Two Classification Methods</i>	<i>112</i>
3.4.3	<i>Using XPA to Study Prostate Cancer</i>	<i>114</i>

4	EARLY PREDICTION OF PRE-ECLAMPSIA FROM SELDI MS DATA	115
4.1	ASSISTING EARLY DIAGNOSIS OF PRE-ECLAMPSIA WITH SELDI MS AND XPA	117
4.1.1	<i>Study Goals</i>	118
4.1.2	<i>Sample Preparation and Laboratory Methodology</i>	121
4.1.3	<i>Processing SELDI MS Data for Use with XPA</i>	123
	Storing SELDI MS Data in the XPA Database.....	128
4.1.4	<i>Classification Models Using SELDI MS Data in XPA</i>	130
	Statistical Methods	130
	Machine Learning	132
	Performance Metrics.....	134
4.1.5	<i>Using XPA to Look for Biomarkers</i>	135
4.2	RESULTS	135
4.2.1	<i>Data Investigation with PCA</i>	135
4.2.2	<i>Statistical Classifiers</i>	142
4.2.3	<i>Machine Learning Classification</i>	146
4.2.4	<i>Comparison of the Two Classification Approaches</i>	150
4.3	BIOMARKER IDENTIFICATION.....	154
4.3.1	<i>Testing the Important MS Peaks</i>	157
4.4	DISCUSSION AND CONCLUSIONS.....	161
4.4.1	<i>Diagnostic Platform Development</i>	161
4.4.2	<i>Biomarker Identification</i>	162
5	DISCUSSION AND CONCLUSIONS	165
5.1	DISCUSSION	166
5.1.1	<i>Generic Data Storage</i>	166
5.1.2	<i>Generic Data Analysis</i>	167
5.1.3	<i>Conclusions</i>	169
5.2	FUTURE WORK.....	170
5.2.1	<i>Further Data Integration</i>	171
5.2.2	<i>Integrated Automatic Data Analysis</i>	171
5.2.3	<i>Interfacing With XPA</i>	172
5.2.4	<i>Systems Biology</i>	173
	BIBLIOGRAPHY	174

Figures

Figure 1-1: A graphic representation of the CGH technique. Differentially labelled test DNA (i.e. from a tumour) and reference/control DNA (from normal individuals) are mixed and co-hybridized to normal metaphase chromosomes, and fluorescence ratios calculated along the length of the chromosomes help to build up a representation of the relative DNA copy number variation.10

Figure 1-2: A graphic representation of the Array CGH technique. As with 'normal' CGH, differentially labelled test DNA is mixed with DNA from a control subject. The DNA mixture is then hybridized to the BAC clones on the array surface. Fluorescence ratios calculated at the arrayed DNA elements (genes/loci of known identity or position in the genome) provide a locus-by-locus measure of copy number variation at a far higher resolution than can be achieved using 'normal' CGH.11

Figure 1-3: A basic overview of the polymerase chain reaction (PCR). This diagram represents a single cycle in a PCR reaction. (A) The reaction starts with a single double-stranded piece of DNA, (B) which is then denatured into single-strands using high temperatures (typically 94-98°C). (C) Primers specific for the region of DNA to be amplified are annealed to the single-stranded DNA, (D) followed by the binding of a DNA polymerase to these primers. (E) The final step of the reaction is elongation; dNTPs are added to the single-stranded DNA by the DNA polymerase resulting in the production of two new pieces of double-stranded DNA. (F) The second cycle of PCR commences with a denaturing step and will repeat the same processes as the first – the end result will be four pieces of double-stranded DNA. The PCR experiment continues for many cycles, doubling the number of double-stranded DNA pieces with each cycle.18

Figure 1-4: A simplified overview of the MALDI-TOF MS technique. The first step in MALDI MS is that the sample to be analysed is absorbed into the matrix. Secondly, a laser is used to bombard the matrix, thus causing an increase in excitation energy and the release of analyte and matrix molecules as charged ions. Finally, these charged ions are accelerated through a time of flight (TOF) analyser – an electric field that causes the ions to accelerate differently based on their charge and mass coupled with a detector to record the time taken for the ions to pass through the electric field – therefore allowing a mass-to-charge ratio (m/z) to be measured for each ion resulting from the sample and matrix.31

Figure 1-5: SELDI-TOF MS: Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. A serum sample is loaded onto a chemically selective protein chip. The chip is then washed with buffer to remove unbound proteins and is treated with an energy-absorbing molecule, such as synapic or cinnaminic acid. The bound proteins are then ionized with a laser and accelerated through an electric field where they are separated based upon their mass/charge ratio (m/z). The detector quantifies the proteins, and a spectrum is generated with analytical software. Figure and caption taken from Simpkins et al. (2005).35

Figure 1-6: Examples of data generated by SELDI MS taken from the study by Rai et al. (2002) where SELDI MS was used to identify and test potential biomarkers for ovarian cancer from blood plasma. (a) A representative spectrum obtained from SELDI analysis (relative intensity vs m/z) generated from a blood plasma sample run on an immobilized nickel (IMAC-Ni) ProteinChip array. The upper panel shows a small portion of the protein profile in a 'regular' spectrum view, while the lower panel shows the same profile portion in a pseudogel view often used with SELDI analysis. (b) Stacked pseudogel views (patient vs m/z) of SELDI data from the analysis of 67 blood plasma samples showing the relative abundance of perspective biomarkers – indicated by the asterisks at the top of the figure. Samples from patients without

cancer (n = 38) are shown above the solid line; samples from patients with cancer (n = 29) are shown below the solid line.....	37
Figure 1-7: An overview of the SBEAMS modular structure concept. This represents the proposed functionality that SBEAMS will eventually provide – modules for many different types of analysis including genomics/transcriptomics, proteomics and many more. However, work on the SBEAMS project is at a very early stage. Currently the only usable components of the system are the proteomics and microarray modules. Image adapted from the SBEAMS website (http://www.sbeams.org/)	43
Figure 1-8: An overview of the Bioinformatics Resource Manager and the potential connections that can be made between disparate data with it. BRM allows the combination of microarray and proteomic data from local repositories, with information from remote sources such as the NCBI (The National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/ a repository of gene, protein SNP and other data) and KEGG (The Kyoto Encyclopedia of Genes and Genomes - http://www.genome.jp/kegg/ which contains detailed information about biological systems and pathways) so that they can be analysed effectively with tools such as Cytoscape (http://cytoscape.org/). Image taken from the Pacific Northwest National Laboratory website (http://www.sysbio.org/).	44
Figure 1-9: Principal components analysis (PCA), image adapted from Brereton (2003). This figure shows the basic principle of PCA – the decomposition of the original data matrix X into the scores and loadings matrices (T and P respectively). The original data matrix represents the recorded values of experimental variables (J) for different samples (I). This is then transformed into the scores matrix (T – samples, I versus scores, A), and the loadings matrix (P – variables, J versus scores A).	47
Figure 1-10: A basic representation of a multi-layered artificial neural network. The interconnections between the neurons are weighted during training using error functions in order to map the inputs and outputs of the network correctly.....	55
Figure 1-11: An application of an SVM to a linearly separable problem. In the plot on the left there are two groups that are separable in linear space. The plot on the right shows how a SVM defines the relationship between the two groups. The two classes are separated by a central plane, and to either side of the dividing plane are parallel ‘support’ hyperplanes that define the maximum distance possible between each class and the central hyperplane – the space between the support hyperplanes and the dividing hyperplane contains no data points. The name support vector machine comes from the name given to the data points that lie on the support hyperplanes and define their position, these are known as the support vectors.....	58
Figure 1-12: An SVM solving a non-linear problem with a linear classifier. The approach the SVM uses to solve this problem is to map the data points to a higher dimensional space where linear separation of the data can be achieved.....	58
Figure 2-1: An Overview of the XPA vision and process flow. Patient samples and information about the patients, and samples themselves are combined with various types of experimental data from different techniques within the XPA database schema. From here different data analysis approaches and investigational techniques can be used on this data to develop classification models or identify points within the data that could be considered putative biomarkers.	64
Figure 2-2: A diagrammatic representation of the XPA system (from a software point of view). This shows a relatively common approach to software construction. The most important aspect of the system is the central database that houses all of the patient, sample and experimental data. This is then wrapped and interacted with via a software layer (often called a ‘middleware’ layer) as this allows interfaces and data analysis routines to be constructed easily without taking into account basic interactions with the database itself – this would be	

taken care of by the XPA API.	66
Figure 2-3: A flowchart giving a representation of the class structure used in the design of the XPA database. The Blue boxes represent the main classes of data, with the arrows defining the flow of data between them. The red boxes represent extra detail and data that are related to the classes that they sit next to.....	68
Figure 2-4: The XPA database schema. This diagram shows the entire XPA database schema: all of the classes, tables, and the links between the different tables.....	74
Figure 2-5: The structure of the XPA Perl module libraries. This diagram represents the layout and contents of the four main Perl modules that make up the XPA system at present (XPA, In, Out, and Annotator). The grey boxes represent the methods/subroutines that are currently with the given modules (for interacting with the XPA database).....	77
Figure 2-6: The Annotator process flow. This diagram represents the way in which Annotator visits various different databases and the information that is retrieved. This represents the steps taken to annotate genomic information within XPA.....	80
Figure 3-1: Results from PCA performed on the raw qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.....	100
Figure 3-2: Results from PCA performed on the normalised and mean centred qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.....	101
Figure 3-3: Results from PCA performed on the range scaled qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.....	102
Figure 3-4: A plot representing the data acquired during the grid search used for training and optimising of the SVM model using only the PSA measurements. The different coloured bars represent the accuracy of the SVM models produced with the values of C and γ represented by the axes.	105
Figure 3-5: A plot representing the data generated from the five-fold cross validation of the PLS-DA models using different numbers of LVs in order to optimise the final model. The blue line represents the root mean square error of calibration – the error when the classifier was tested against the training datasets, whereas the red line represents the root mean square error of cross validation.	107
Figure 3-6: A plot representing the data acquired during the training and optimising of the SVM model. The different coloured bars represent the accuracy of the SVM models produced with the values of C and γ represented by the axes.....	108
Figure 3-7: A comparison of the sensitivity results achieved using SVM and PLS-DA models against the published expected sensitivity figures for the PSA test when predicting the condition of prostate cancer. The results displayed in this figure indicate that the models produced using SVMs and PLS-DA are over twice as sensitive as the expected performance of the traditional PSA test indicating that they would be very suitable for use in a clinical situation (given further testing) as these models would correctly identify patients with prostate cancer more accurately than the expected performance of the traditional PSA test.	111

- Figure 3-8: A comparison of the results achieved from the two different classification approaches used in this study when working with both PSA alone and PSA combined with gene expression data. The results represented in this figure clearly demonstrate that SVMs were the better classifiers on these datasets as they produced not only more accurate predictions of patient outcome, but also had a higher sensitivity and specificity giving greater confidence in the predictions made by the SVM models..... 113
- Figure 4-1: An overview of the processes used to look at the SELDI MS data from pre-eclampsia sufferers. This details each individual step that was carried out within this study in order to use the SELDI MS data with the XPA system. 120
- Figure 4-2: A flowchart representing the processes put in place to convert the Ciphergen XML data files into raw spectra files that could be used with the R PROcess package. All of these steps were implemented as part of a simple program written using the Perl programming language. 124
- Figure 4-3: Spectrum plots demonstrating the various stages of processing with R. (a) A raw spectrum. (b) Spectrum baseline subtraction – the green line shows the raw spectrum; the red shows the estimated baseline of the raw spectrum; and the blue indicates the raw spectrum minus the estimated baseline. (c) A spectrum with the detected peaks marked by red circles..... 129
- Figure 4-4: Results from PCA performed on all of the SELDI data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the PE group, and green represents the control group. No clear disease-related separations could be seen in this data to lead to any more findings. 137
- Figure 4-5: Further results obtained by performing PCA on all of the SELDI MS data to see if we were observing differences in the SELDI chip types being used in Figure 4-4 instead of variances in the data due to whether the subject suffered from PE or not. These plots represent the first three PCs compared to each other to see if there was any noticeable separation between the chip classes. Red markers represent the CM10 chips, green the H50 and blue represents the IMAC chip type. As can be seen from the plots, some clear clustering of data points can be observed, suggesting that by performing PCA on all of the SELDI datasets together, most of the variance captured by the technique is due to the SELDI chip type used. 138
- Figure 4-6: Results from the PCA carried out on each chip type dataset separately. (a) Represents the CM10 SELDI chip, (b) the H50 chip, and (c) the IMAC chip type. As with Figure 4-4, the red markers indicate the PE group, and the green markers indicate the control group. As can be seen in the figures, no clear separation was observed between these groups (on any chip type) that could lead to further investigation with PCA..... 141
- Figure 4-7: Comparison of the results obtained from all of the spectrum peaks identified by the R PROcess package (the ‘peak lists’) when they were used to train and test classifiers based on SVMs (blue bars) and PLS-DA (red bars). It can be clearly observed that the SVM models were the better classification tools with this dataset..... 151
- Figure 4-8: Comparison of the results obtained from the common peaks (identified by the R PROcess package) when they were used to train and test classifiers based on SVMs (blue bars) and PLS-DA (red bars). Once again it can be clearly observed that the SVM models were the better classification tools when used with this dataset..... 152
- Figure 4-9: A comparison of the mean average classification accuracy, sensitivity and specificity achieved from the different SELDI MS protein chip types. Representing the data this way allows the clear observation that the most effective chip type to use for the classification of the pre-eclampsia data was the H50 chip (red bars) as it produced consistently better

classification models.....	153
----------------------------	-----

Figure 4-10: A comparison of the accuracies achieved by the PLS-DA models using the all, common and important peaks datasets. Looking at the datasets this way shows a trend in many of the chip/fraction/laser combinations that as the dataset becomes reduced (i.e. moving from all peaks to common peaks to important peaks) so too does the accuracy of the model produced. However, there are some exceptions to this rule – one particular example is the H50, fractions 1&2 datasets that consistently prove to build the best classification models. 160

Tables

Table 3-1: Results of the PLS-DA model trained and tested on the PSA data only.....	104
Table 3-2: Results of the SVM model trained and tested on PSA data only.....	105
Table 3-3: Results of the classifications made by the PLS-DA model on unseen test data containing both PSA measurements and gene expression values/ratios.....	107
Table 3-4: Results of the classifications made by the SVM model on unseen test data containing both PSA measurements and gene expression values/ratios.....	108
Table 4-1: pH level / solvents used to fractionate the antenatal blood serum samples.....	122
Table 4-2: Results for the PLS-DA models trained and tested against the peak lists generated by the R PROcess package. Reported values represent the percentage accuracy of the PLS-DA models created for each dataset when used on the test (30% of total samples) dataset.....	144
Table 4-3: A more in-depth look at the results obtained by the PLS-DA models created from the peak lists. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.....	144
Table 4-4: Results for the PLS-DA models trained and tested against the most common peaks identified by the R PROcess package. Reported values represent the percentage accuracy of the PLS-DA models for each dataset when used on the test (30% of total samples) dataset.....	145
Table 4-5: A more in-depth look at the results obtained by the PLS-DA models created from the most common peaks. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.....	145
Table 4-6: Results for the SVMs trained and tested against the peak lists generated by the R PROcess package. Reported values represent the percentage accuracy of the SVM created for each dataset when used on the test (30% of total samples) dataset.....	148
Table 4-7: A more in-depth look at the results obtained by the SVMs created from the peak lists. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.....	148
Table 4-8: Results for the SVMs trained and tested against the most common peaks identified by the R PROcess package. Reported values represent the percentage accuracy of the SVM created for each dataset when used on the test (30% of total samples) dataset.....	149
Table 4-9: A more in-depth look at the results obtained by the SVMs created from the most common peaks. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.....	149
Table 4-10: The m/z values (MS peaks) identified in the common peaks datasets as contributing the most information content to the PLS-DA models for the classification of patient samples.....	156
Table 4-11: Results of the PLS-DA models trained and tested using only data for peaks identified as contributing the most information to the previous PLS-DA models. The cells with no results represent chip/fraction/laser combinations that did not contain enough important peaks to generate a new PLS-DA classification model.....	159

Table 4-12: Performance metrics for the PLS-DA models trained and tested using only data for peaks identified as contributing the most information to the previous PLS-DA models. The cells with no results represent chip/fraction/laser combinations that did not contain enough important peaks to generate a new PLS-DA classification model. 159

Table 4-13: Results and performance metrics for the SVM model trained and tested using only data for the peaks identified as contributing the most information to the PLS-DA models for the H50 (fractions 1&2) datasets compared against the results achieved using the common peaks dataset. 160

1 Introduction

Biomarkers have been defined in many ways in the past, but a particularly good definition is that “a biomarker is a biological characteristic that is measured and evaluated objectively as an indicator of normal biological processes and pathogenic processes, or pharmacological responses to therapeutic intervention”, (De Gruttola et al. 2001; Atkinson et al. 2001).

This means that biomarkers may be considered as any parameter of a patient (or combination of parameters) that can be measured, for example; mRNA expression profiles, proteins, proteomic patterns, lipids, or features from imaging methods, or in electrical signals (Hewitt et al. 2004). As such, biomarkers are of great interest to the research community, pharmaceutical industry and the medical fields as they have applications in all of these areas, i.e. they can be used to indicate whether a patient is suffering from a particular disease; or report on a stage of a biological process; or even indicate whether an individual has ever been infected by a particular pathogen. The uses of biomarkers are numerous and widespread, but the best biomarkers are accurate, relatively non-invasive and are detectable using easy-to-perform tests that can be done at the bedside or in the outpatient setting (De Gruttola et al. 2001; Bao-Ling Adam et al. 2002; Frank & Hargreaves 2003; Hewitt et al. 2004).

Many commonly used blood tests in medicine are tests for the presence of biomarkers, one such example is the test used to diagnose diabetes – this measures

the level of glucose in the blood at a given time. If a recorded blood-glucose level were above what is considered 'normal' (>200 mg/dl) an individual would be recommended for further screening (American Diabetes Association 2007) as this is a clear indicator of diabetes.

Another such widely used biomarker test is the screening for prostate specific antigen (PSA) in the blood serum of men, which is used to detect the onset of prostate cancer (Potosky et al. 1995; Grönberg 2003). An elevated level of PSA in the blood serum is an indicator of cancer of the prostate and is a preferred test compared to the prostate biopsies that the test has replaced.

1.1 *Biomarker Discovery*

Biomarker discovery is the term coined to describe the process by which biomarkers are discovered, characterised and put to use; and is the main focus of this PhD thesis. Traditionally biomarker discovery has focused on looking for a single molecule or characteristic (such as glucose in diabetes) that could be used to predict or diagnose a condition. However, this approach to biomarker discovery is proving harder to follow as fewer and fewer successful individual biomarkers are being discovered.

The idea of finding a single biomarker for a disease or condition is highly appealing as it is analytically very simple. However, this makes little sense from a biological perspective as (for example) a condition such as cancer is caused by damaged and deranged cells of the host multiplying uncontrollably – not an external infectious

agent, and comes in hundreds of forms, types and stages (Liotta et al. 2003). We would not expect that there is a single unique characteristic that can be measured to accurately monitor a complex condition such as cancer, as the interplay of multiple genes and multiple non-genetic factors, not a single allele, usually dictates disease susceptibility and response to treatments (Collins et al. 2003). These are reasons why biomarker discovery studies must move on from this traditional approach and adopt new technologies in order to achieve significant new findings. Just as genomics researchers have moved beyond one-gene-at-a-time analysis, and are profiling thousands of gene transcripts to generate entire patterns of information, the same should be done with disease biomarkers (Liotta et al. 2003)

Modern biomarker discovery is doing just that. New molecular biology techniques promise to find relevant markers rapidly, without detailed insight into mechanisms of disease. By screening many possible bio-molecules at a time, a parallel approach can be tried. Genomics, transcriptomics and proteomics are some technologies that are being used in this process.

1.2 Genomics

The genome can be defined as a complete collection of an organisms' hereditary information encoded within its DNA (deoxyribonucleic acid) and includes both the genes and the non-coding sequences of the DNA (Ridley 2000).

Genomics is the very broad term used to encompass the study of the genome and the DNA that it is made up from, e.g.

- Efforts to determine the entire DNA sequence of an organism such as the Human Genome Project (Human Genome Sequencing Consortium 2001), and the sequencing of numerous other organisms (Nakabachi et al. 2006; Adams et al. 2000; Blattner et al. 1997; Fiers et al. 1978).
- The identification and analysis of the genes and other traits found within these sequences (Zhang 2002; Gelfand 1995).
- Comparing genomic data between organisms to reveal homologous sequences, (comparative genomics), for example within individual genes that reflect their common evolutionary origin and subsequent conservation (Gregory et al. 2002).

These are but a few examples of the works that are encompassed under the term genomics.

1.3 *Transcriptomics*

The transcriptome has been defined as the full set of mRNA molecules (transcripts) produced by a system under observation (Carninci 2007). Unlike the genome, which is essentially a static entity, the transcriptome can be affected by both external and internal factors (Velculescu et al. 1997) and serves as a link between an organism's genome and its proteome and therefore, ultimately its physical characteristics.

The term 'transcriptomics' is used to describe the study of the transcriptome, just as genomics is the study of the genome. It is the research that deals with the

analysis of large datasets on the level of cDNA and mRNA expression (Jansen & Schalkwijk 2003) and the up and down regulation of gene expression in response to differing situations and conditions.

1.4 Genomics and Transcriptomics in Biomarker Discovery

In recent years, there has been an explosion in the use of genomic information in science. This information has greatly expanded our understanding of disease, and advanced scientific research.

It is now accepted that in many common diseases, the interplay of multiple genes, not a single allele, usually dictates disease susceptibility and response to treatments. Deciphering the role of genes in human health and disease is a formidable problem for many reasons (Collins et al. 2003). Increasing the complexity of understanding these disorders is the growing recognition that the genetic risk factors likely interact not only with each other but also with poorly understood environmental factors. Ultimately, the prediction of an individual's risk for any disorder will be determined by their genotype (Rich 2008).

As such, studies using genomics and transcriptomics technologies have been undertaken to elucidate potential candidate biomarkers in many such conditions. Investigations by Helgadottir et al. (2004) and Hakonarson et al. (2005) identified the gene ALOX5AP, encoding 5-lipoxygenase activating protein (FLAP) as an indicator of an individual's risk of developing heart disease or stroke. Other

examples of work in this field have identified a number of genes that change their expression levels in human prostate tumours (Eder et al. 2004; Creighton 2008) by using microarrays to analyse and profile the expression characteristics of thousands of genes at the same time. These are but a few examples of the many studies recorded in this field of work.

The following sections describe several important areas in the use of genomics and transcriptomics in biomarker discovery.

1.4.1 Genomic Copy Number Variation

A copy number variation (CNV) is a section of DNA one kilobase (Kb) to several megabases (Mb) in size, for which copy-number differences have been revealed by comparison of two or more genomes (Feuk et al. 2006).

Ordinarily, humans have two copies of each chromosome as humans are described as diploid in traditional genetics terms, but this can vary in many regions (of the chromosomes) due to mutations such as deletions, duplications, inversions or translocations. Features that appear to render genomic regions susceptible to rearrangement (and therefore produce CNVs) include region-specific repeat sequences, or low-copy repeats (LCRs). LCRs can contain one or multiple genes, gene fragments, retroviral sequences, or regulatory regions (Lee & Lupski 2006; Lupski 1998).

CNVs can affect gene function in many ways. Deletion or disruption of one or more genes can cause functional loss, with CNVs functioning as dominant or recessive

alleles according to the function of the affected product (Cook & Scherer 2008), alternatively, CNVs can disrupt a regulatory element or even produce novel gene products (Lupski & Stankiewicz 2005).

CNVs as Biomarkers

DNA copy number variation represents a considerable source of human genetic diversity, and like other types of genetic variation some CNVs have been associated with susceptibility or resistance to disease. Known examples of CNVs identified as markers for disease include a higher than normal copy number for the gene EGFR in non-small cell lung cancer (Cappuzzo et al. 2005), and low copy numbers of the CCL3L1, FCGR3B and DEFB4 genes, are associated with increased susceptibility to AIDS, immunologically mediated glomerulonephritis and Crohn disease respectively (Kehrer-Sawatzki 2007).

CNVs can be discovered by cytogenetic techniques such as comparative genomic hybridization (CGH) and array comparative genomic hybridization (Array CGH, but also known as aCGH or a-CGH).

CGH and Array CGH

Comparative genomic hybridization (CGH) and array comparative genomic hybridization (Array CGH) are genomic techniques for detecting copy number changes (gains/losses) in the DNA content of a given subject's DNA.

CGH

A typical CGH experiment is carried out in the following steps:

1. DNA samples taken from a sample of interest (i.e. a tumour cell), and from normal DNA (the control/reference) are fluorescently labelled using different colours for each sample (frequently used labels include fluorescein, rhodamine or Texas Red).
2. These DNA samples are mixed along with unlabelled Cot-1 DNA (used to suppress repetitive DNA sequences).
3. The DNA mix is then hybridised to normal metaphase chromosomes fixed onto a glass slide.
4. Epifluorescence microscopy and quantitative image analysis is used to calculate fluorescence ratios along the chromosomes. This colour (fluorescence) ratio is used to evaluate regions of DNA gain or loss in the sample of interest.

This process is shown diagrammatically in Figure 1-1.

Array CGH

Array CGH is an evolution of CGH that allows for detection of CNVs at a much higher resolution. The basic principle of Array CGH is identical to CGH, the differentiating step is that instead of the DNA samples being hybridised to fixed metaphase chromosomes, they are hybridised to arrays of genomic BAC, P1,

cosmid or cDNA clones. Fluorescence ratios are then calculated at the arrayed DNA elements to provide a locus-by-locus measure of DNA copy-number variation. This allows variations in copy number to be detected in sequences as small as 5-10 kilobases – a much higher resolution compared to the 5-10 megabase resolution expected from chromosomal CGH.

A graphical representation of Array CGH is shown in Figure 1-2.

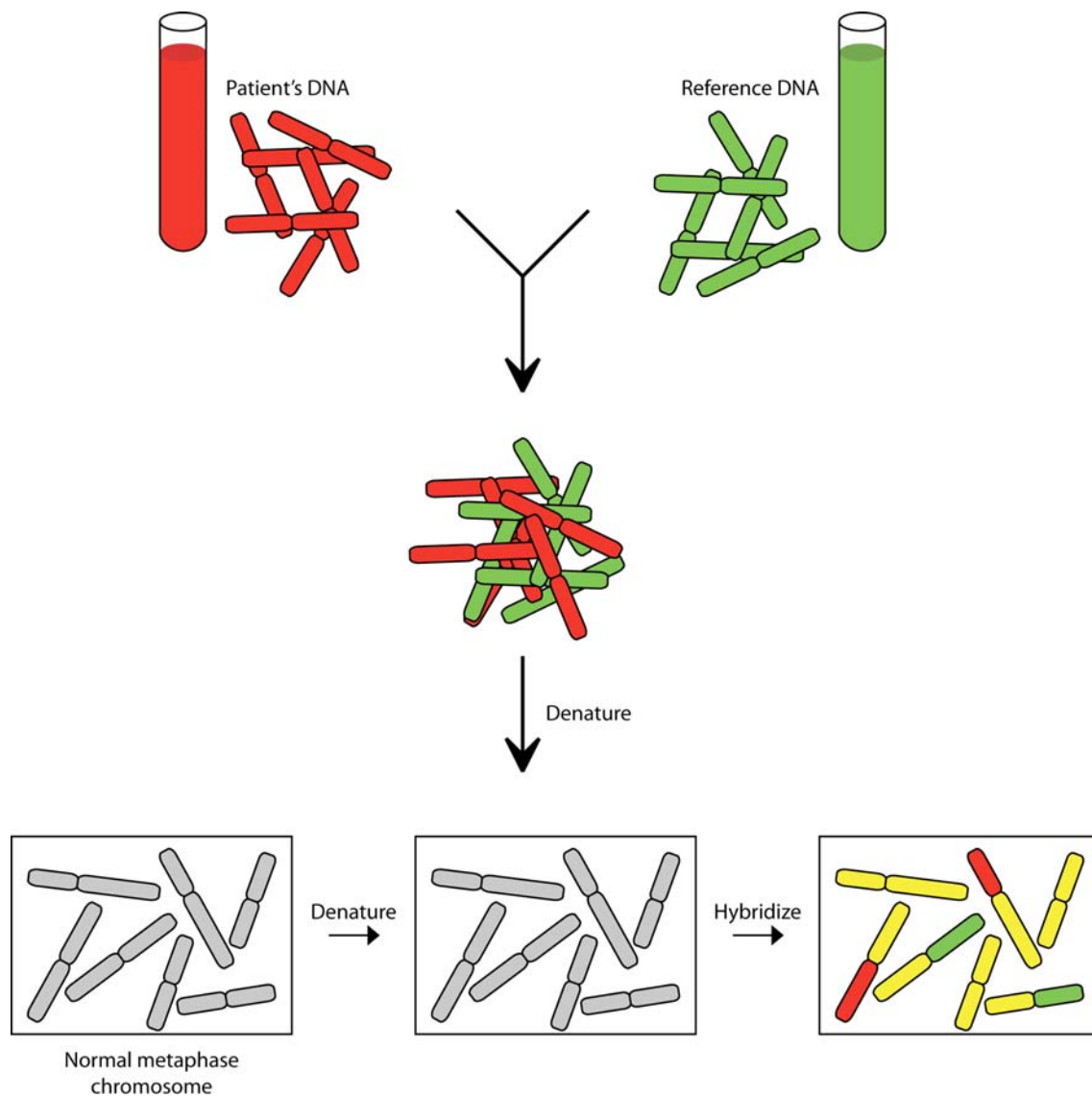


Figure 1-1: A graphic representation of the CGH technique. Differentially labelled test DNA (i.e. from a tumour) and reference/control DNA (from normal individuals) are mixed and co-hybridized to normal metaphase chromosomes, and fluorescence ratios calculated along the length of the chromosomes help to build up a representation of the relative DNA copy number variation.

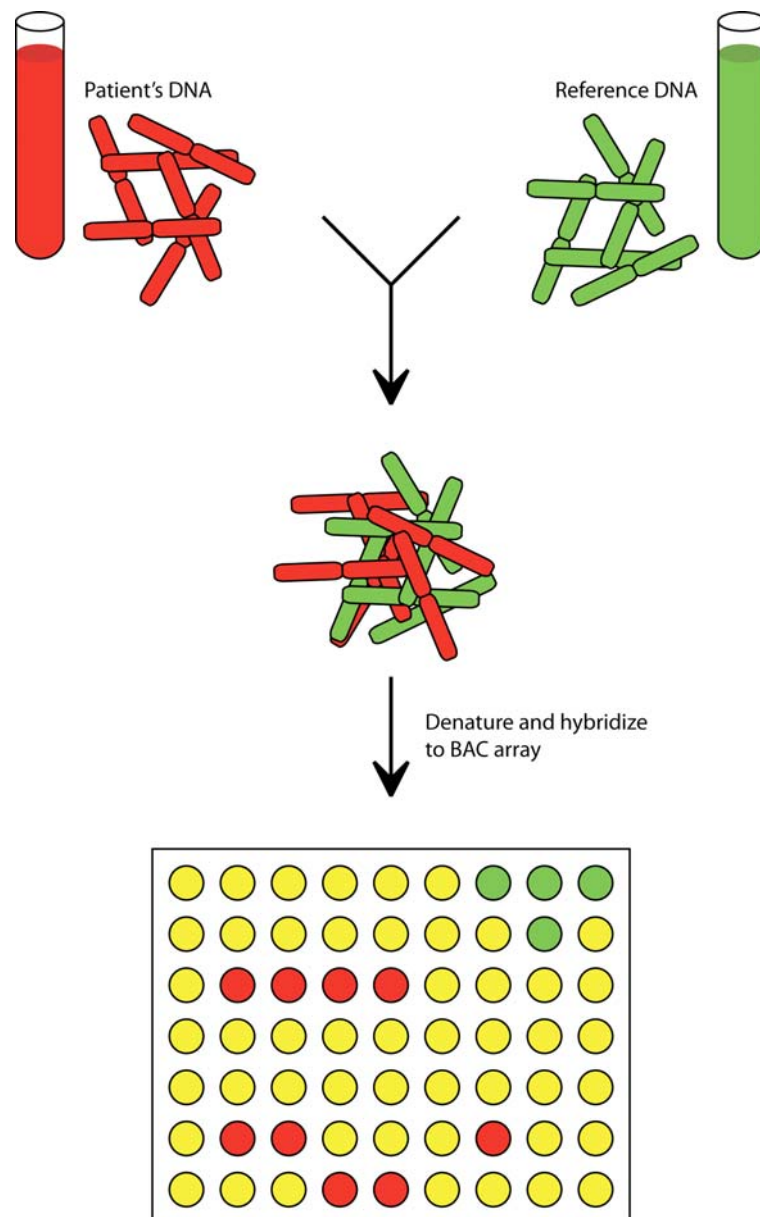


Figure 1-2: A graphic representation of the Array CGH technique. As with 'normal' CGH, differentially labelled test DNA is mixed with DNA from a control subject. The DNA mixture is then hybridized to the BAC clones on the array surface. Fluorescence ratios calculated at the arrayed DNA elements (genes/loci of known identity or position in the genome) provide a locus-by-locus measure of copy number variation at a far higher resolution than can be achieved using 'normal' CGH.

1.4.2 Single-nucleotide Polymorphisms

A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide – A, T, C, or G – in the genome differs between members of a species or between paired chromosomes in an individual. At present there are more than 9 million SNPs catalogued in online public databases (Rocha et al. 2006) such as Entrez SNP, formerly known as dbSNP – <http://tinyurl.com/c3ahmk> (Sherry et al. 2001) and the HapMap project – <http://www.hapmap.org/> (The International HapMap Consortium 2003).

SNPs may occur anywhere within the genome: introns and exons within genes, or intergenic regions between genes. SNPs occurring within a coding region of a gene (exon) will not necessarily change the resulting amino acid sequence of the protein that is produced, these types of SNPs are known as *synonymous* SNPs (also sometimes called a silent mutations). SNPs that cause a different amino acid sequence (and peptide) to be produced are known as *non-synonymous*. Non-synonymous mutations (SNPs) can cause either a *mis-sense* or a *non-sense* change, where a mis-sense change results in a different amino acid, and a non-sense change results in a premature stop codon. SNPs that occur outside exonic regions may however still cause mutations by affecting gene splicing or transcription factor binding.

SNPs as Biomarkers

As with copy number variation (described in the previous section) and all other

forms of genetic variation, SNPs have been associated with susceptibility or resistance to disease. Examples of such findings are:

- One such example is a recent genome-wide association study (GWAS) that found a SNP (rs5984894) on Xq21.3 in the gene PCDH11X was shown to be strongly associated with late-onset Alzheimer's disease (Carrasquillo et al. 2009).
- Another recent study by Shah et al. (2007) identified three SNPs (rs5987079, rs1936645 and rs4898352) located within the gene F8 on the X chromosome as effective markers for carriers of the hemophilia A locus in people of Indian origin.
- A study by Maraganore et al. (2005) looking at the genetic factors contributing to Parkinsons Disease found a SNP (rs7702187) within the semaphorin-5A gene on chromosome 5p. The protein encoded by this gene plays an important role in neurogenesis and in neuronal apoptosis, therefore suggesting an important marker for the condition.

There are many laboratory techniques and commercial products used to observe known SNPs in different clinical situations, common examples in modern biomarker studies are high throughput SNP microarrays such as the Affymetrix range of SNP microarrays (Affymetrix 2009) and the Infinium HD BeadChip range from Illumina (Illumina 2009b; Shen et al. 2005) – arrays capable of targeting and measuring the presence and copy number of around one million genomic markers in a single experiment. (The technology behind the Affymetrix arrays is discussed in more detail in section 1.4.3).

1.4.3 Gene Expression Profiling

The previous sections discussed the use of genomic variations and mutations such as CNVs and SNPs as markers for disease, and some of the technologies used to investigate and identify them. These sources of variation and mutation are not the only source of marker information to come from the genome; gene expression itself can be used as a biomarker for different conditions and processes.

Gene Expression Profiling as a Biomarker

Gene expression profiling (measuring levels of cDNA/RNA in a given sample) has been used in many areas of biological and medical research over the last decade; one area where these techniques have been put to use extensively is the area of cancer research (Lukes et al. 2009; Segal et al. 2005).

One such example is a recent study by Nimeus-Malmstrom et al. (2008) looking at cases where some patients with breast cancer develop local recurrence after breast-conservation surgery despite postoperative radiotherapy, whereas others remain free of local recurrence even in the absence of radiotherapy. Their work identified a distinct gene expression profile for patients who developed local recurrences of breast cancer vs. those that did not, that if verified in further studies, might prove to be an important tool in the decision making for surgery and follow-up therapy.

This work is not however limited to the confines of cancer research, a study by Mootha et al. (2003) analysed biologically linked sets of genes (e.g., pathways) for

association with disease. They applied this method to a data set of human diabetic muscle (Type 2 diabetes mellitus), with the goal of identifying processes that were systematically altered in diabetic muscle. Their analysis identified a set genes involved in the oxidative phosphorylation pathway whose expression is coordinately decreased in human diabetic muscle and together might represent a target for the monitoring, prevention and treatment of diabetes.

There are many laboratory techniques used to study and investigate the transcription machinery of the genome in biomarker studies, but two of the more common techniques used are the Polymerase Chain Reaction (PCR) – in particular, quantitative PCR, and microarrays.

Quantitative PCR

The Quantitative Polymerase Chain Reaction (qPCR) is a laboratory technique for amplifying and quantifying targeted DNA molecules. qPCR is also known and referred to as Real-Time PCR (RT-PCR), but is not to be mistaken for Reverse-Transcription PCR (confusingly, also known as RT-PCR) as this is a method for the amplification of mRNA and does not produce quantitative measurements, whereas qPCR allows for the accurate measurement of DNA transcription.

In its most basic description qPCR is analogous to ‘regular’ PCR, the only difference is the incorporation of a DNA binding fluorescent dye into the laboratory process, and the measuring of fluorescence (and therefore DNA quantity) at the end of each round of amplification. As more DNA is produced, more of the fluorescent dye is also incorporated into the resulting amplified DNA, thus making the fluorescence

greater. This allows the quantity of DNA in the original sample to be recorded, in a very high throughput fashion compared to traditional laboratory methods, but more importantly it allows the quantification of difficult samples that without the amplification steps would not be detectable by other means.

Concepts of PCR

A typical PCR procedure is made up of a series of 20 to 40 repeated cycles; each PCR cycle consists of three steps: denaturation, annealing and elongation, a basic overview of the PCR cycle is presented in Figure 1-3.

- **Denaturation** – This is the first stage in a PCR cycle, and consists of heating the target DNA, two specific DNA primers (flanking the section of DNA that is to be amplified), dNTPs and DNA polymerase to 94-98°C for a given duration of time. This causes the target DNA to disassociate into two single strands.
- **Annealing** – In the second stage of the PCR cycle the temperature is lowered to between 50-65°C. This allows the primers to anneal to the target DNA strands. The DNA polymerase will now also bind to the target DNA-primer complex to begin DNA synthesis.
- **Elongation** – The final step in the PCR cycle has a temperature dependant on the DNA polymerase in use; if the DNA polymerase is Taq polymerase, 72°C is commonly used. This step will then allow the combining of the dNTPs (added in the first step) to the single strands of DNA and if there are no limiting factors (such as running out of reagents), will double the

amount of DNA in the procedure.

These three steps may also be joined by a number of extra steps at the beginning and end of a PCR experiment, these are typically extended elongation steps to ensure that all remaining single stranded DNA is fully extended, and holding steps at a reduced temperature (4-15°C) to allow for short-term storage or pausing of the reaction.

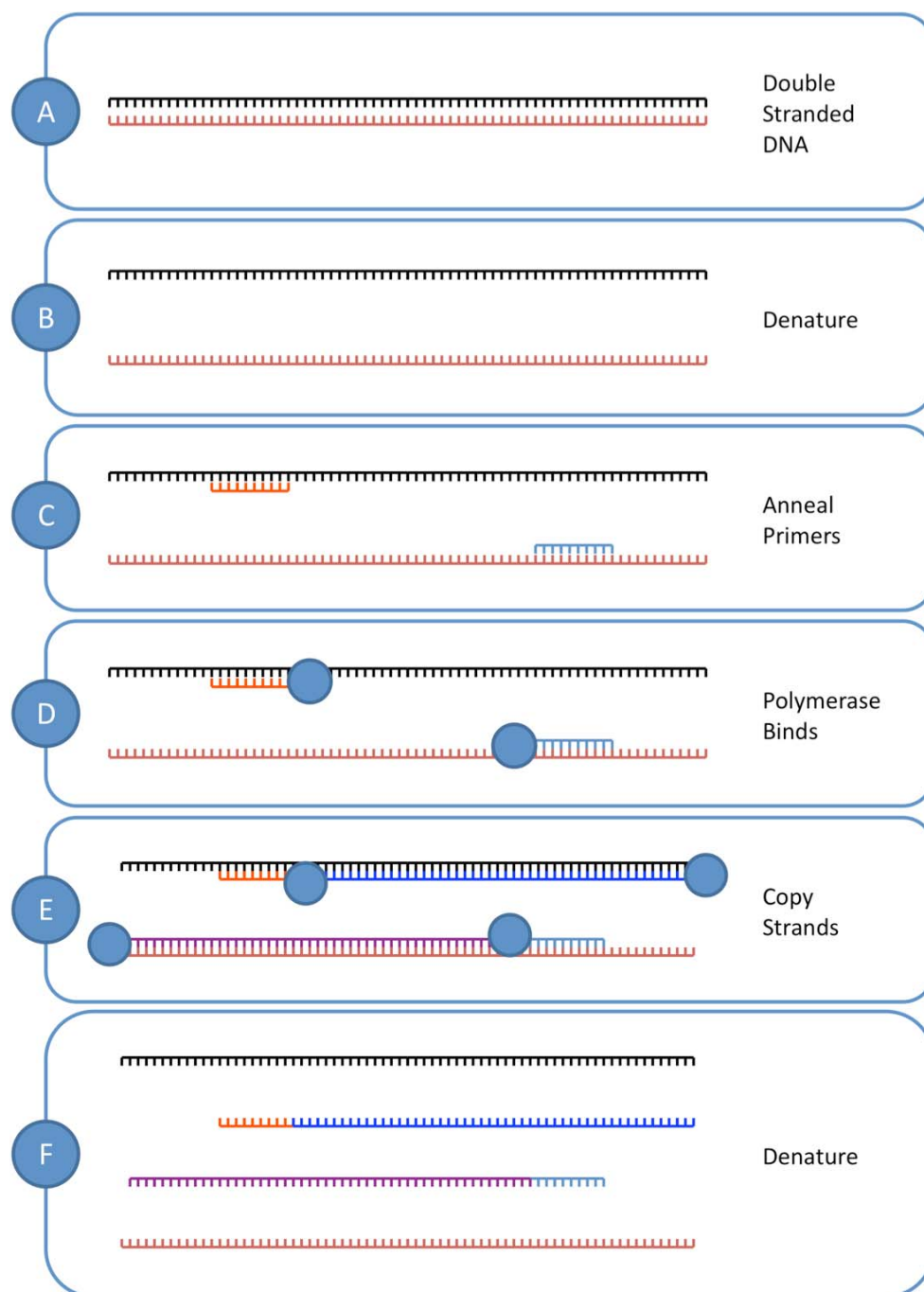


Figure 1-3: A basic overview of the polymerase chain reaction (PCR). This diagram represents a single cycle in a PCR reaction. (A) The reaction starts with a single double-stranded piece of DNA, (B) which is then denatured into single-strands using high temperatures (typically 94-98°C). (C) Primers specific for the region of DNA to be amplified are annealed to the single-stranded DNA, (D) followed by the binding of a DNA polymerase to these primers. (E) The final step of the reaction is elongation; dNTPs are added to the single-stranded DNA by the DNA polymerase resulting in the production of two new pieces of double-stranded DNA. (F) The second cycle of PCR commences with a denaturing step and will repeat the same processes as the first – the end result will be four pieces of double-stranded DNA. The PCR experiment continues for many cycles, doubling the number of double-stranded DNA pieces with each cycle.

Concepts of qPCR

As stated previously, a basic description of qPCR is that it is analogous to 'regular' PCR, with the only change being the addition of fluorescently labelled molecules to the DNA, and fluorescence being measured at the end of each elongation step to record the quantity of DNA in the sample mixture.

The two common methods of quantification used are either fluorescent dyes that bind to double-stranded DNA (dsDNA) or the use of fluorescent reporter probes specific for the targeted sequence of DNA for amplification. Both methods have good and bad points:

- The main points for the DNA dyes are that they are inexpensive (comparatively) but they report on the total amount of double-stranded DNA in the sample mixture, so if there is any non-specific (non-targeted DNA) dsDNA in the sample, this will adversely affect the recorded measurements.
- The arguments for and against the use of fluorescent probes are that on the negative side, they are more expensive than the DNA dyes as a specific probe for the targeted region of DNA needs to be constructed for each experiment. But the positive argument is that they give the most accurate measurements, as through the use of a specific probe, non-specific dsDNA does not affect the fluorescence measured.

Data Obtained from qPCR

As described previously, qPCR is a quantitative variant of the polymerase chain reaction. The data that is generated from a qPCR experiment is very simple as it is a single measurement of the fluorescence observed at the end of each amplification cycle that is directly proportional to the amount of DNA present. The final dataset produced from an qPCR experiment is a simple text file that details the amount of fluorescence/DNA present at the end of each amplification cycle – these files are generally very small in size (less than a megabyte) and easy to handle from a bioinformatics perspective as each individual qPCR experiment only looks at the expression levels of a single target unless fluorescent probes are used, then multiple targets can be quantified at the same time.

DNA Microarrays

Like qPCR, microarrays are a laboratory technology that allows the quantification of DNA transcription to be quickly recorded in the laboratory. However, the scale of microarrays and qPCR platforms are quite different in that qPCR can only measure the quantity of expression of a single, or small number of loci in an individual laboratory run, (the targeted segment of DNA for amplification), microarrays on the other hand are routinely used to interrogate thousands of genes and entire genomes in a single experiment (Hardiman 2004).

Only a decade ago, the concept of being able to simultaneously measure the concentrations of every transcript in the cell in a single experiment would have seemed like science fiction to most researchers. However, the advent of these

technologies has empowered researchers to do just that, and has also added new dimensions to our ability to leverage information from sequencing projects into a more comprehensive and holistic understanding of cell physiology (D E Bassett et al. 1999). DNA microarrays provide a simple and natural vehicle for exploring the genome in a way that is both systematic and comprehensive (Brown & Botstein 1999).

Concepts of Microarrays

Microarrays are based on one simple principle, the principle of base pairing. Every nucleic acid strand carries the capacity to recognise complementary sequences through base pairing (Lipshutz et al. 1999). The main differentiating feature between microarrays and PCR is that in PCR the entire transcription process is mimicked, enabling the quantification of very small concentrations of DNA; microarrays on the other hand only quantify DNA that is already present in samples.

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, inkjet printing, photolithography (Pease et al. 1994) or microscopic silica beads (Illumina 2009a; Steemers & Gunderson 2007).

Spotted Arrays

In spotted arrays (using pins or inkjet printing) nucleic acid molecules of interest (be they cDNA or artificially synthesised oligonucleotides) are immobilised on a surface by laboratory robots as 'spots' of DNA, in an ordered fashion so that the

identity of each immobilised spot is known (or at least a given ID for the sequence is known). Each array can contain thousands of these spots (Schena et al. 1996), and in some cases an array can be made up of many other miniaturised sub-arrays (Zarrinkar et al. 2001). Spotted arrays can be both manufactured in the laboratory (if a laboratory has the appropriate equipment) or purchased commercially.

Photolithography Arrays

Oligonucleotide arrays constructed by photolithographic synthesis were first introduced by the company Affymetrix (<http://www.affymetrix.com>). These arrays are manufactured directly on the surface of the microarray (in the case of Affymetrix GeneChips, a quartz based silicon chip is used) using the following process, adapted from Dalma-Weiszhausz et al. (2006):

- Near-ultraviolet light is passed through a mask containing open windows over the surface of the entire array; the size and the location of each open window indicates the surface on the array that will be activated for chemical synthesis.
- A single nucleotide is washed over the array and couples to the unprotected areas. This process is then repeated for the different nucleotides.
- Through successive steps, any oligonucleotide sequence can be built on each feature of the array.

The length of oligonucleotide sequences constructed may be longer or shorter depending on their desired purpose; longer probes increase the specificity to individual target genes, whereas shorter probes allow for the construction of

higher density arrays and are cheaper to manufacture.

These types of arrays generally offer a much higher number of probes per-chip compared to spotted arrays and could help to reduce the experimental variability that would be introduced by producing microarrays in the laboratory.

Photolithography based arrays are available from a number of manufacturers including Affymetrix and Agilent Technologies.

Bead Arrays

Bead based microarrays, (also known as BeadChips or BeadArrays), are a proprietary platform manufactured and distributed by Illumina Inc. (<http://www.illumina.com>).

As previously described, other types of microarrays are manufactured by spotting or synthesizing probes onto two-dimensional substrates or surfaces at known locations. In contrast, bead arrays are based on the random self-assembly of a bead pool onto a patterned substrate (Fan et al. 2006). Silica beads, each 3µm in diameter with unique (per bead) 50-mer gene-specific probe sequences immobilized on their surface, are assembled randomly into wells on the surface of the chip. The final step in the manufacturing of these chips is a decoding step (Gunderson et al. 2004), to identify the probes at each location in the array.

As with photolithography based arrays, bead arrays generally offer a much higher number of probes per-chip compared to spotted arrays and could help to reduce the experimental variability that would be introduced by producing microarrays in the laboratory. Bead array microarrays are a proprietary microarray platform

only available from Illumina.

Using Microarrays

Expression data is generated from microarrays through the hybridisation of fluorescently labelled nucleic acid from laboratory samples (i.e. diseased tissue, or yeast strains in certain stages of the cell cycle) to the spots/probes on the microarray. DNA molecules that are complementary to the probes (spots) on the microarray will hybridise and 'stick' to the probe. After a given length of time, the microarray is then washed, removing any non-hybridised nucleic acid. An image is then generated from the finished microarray using a fluorescence scanner that then inputs the image onto a computer. Expression ratios can then be calculated by comparing the fluorescence produced by the samples against controls (that will emit a known fluorescence) built onto the microarray at production.

Two-Channel Arrays

The use-case given above can be described as a 'single-channel' microarray – it is only capable of measuring expression data for one sample at once. Examples of single-channel arrays are the Affymetrix GeneChip arrays and several members of the Illumina BeadArray family.

Two-channel microarrays on the other hand allow the hybridisation of two differentially labelled samples (i.e. disease and control samples) onto a single array, which is then scanned twice. An individual image is generated for each type of fluorescent dye in use; the single compound images depicting red and green spots normally seen in publications is generated at a later stage in software by

combining these images.

Examples of two-channel microarrays are spotted arrays and certain variants of Illumina's BeadArray family.

Data Obtained from Microarrays

As described above, the raw data output generated from a microarray experiment is in the form of several large high-resolution images (up to several megabytes in size dependant on the quality of the scanner) acquired via a fluorescence scanner on a microarray. Once the raw images are generated, specialist software is used to identify individual spots on the array, quantify the intensity of colour in each spot and calculate a relative expression value for each gene covered on the microarray as this is directly proportional to the amount of labelled DNA hybridised on each spot, thus giving an accurate picture of the level of expression for each DNA molecule found in the array. Image analysis software for this purpose is typically supplied with the microarray scanning hardware itself, in the form of packages such as the proprietary GeneChip support software from Affymetrix, or GenePix from MDS Analytical Technologies (<http://www.moleculardevices.com/>) but alternative packages are also available for this purpose; such as BlueFuse from the Cambridge based company BlueGnome (<http://www.cambridgebluegnome.com>) or the microarray image analysis toolboxes for Matlab (White et al. 2005).

The output from these software packages is usually in the form of a rather large text or binary file that details relative expression values measured for each spot/gene on the array – the size of these files is dependant on the complexity of

the original array, as a more complex microarray will contain many more data points and therefore have a much larger final output file; but these files can vary from less than one megabyte in size up to several tens of megabytes in size for a single experiment. Following the interpretation of the fluorescence data into quantitative expression values all subsequent data analysis on microarray experiments is performed on the numerical gene expression values.

1.5 Proteomics

The proteome is the full complement of proteins expressed by a genome, cell, or organism. More specifically, it is the set of proteins expressed at a given time under defined conditions. Proteomics is the large-scale study of the proteome and all the individual proteins that it encompasses.

1.6 Proteomics in Biomarker Discovery

There is a concept that the human plasma proteome is likely to contain most, if not all, human proteins, and that almost any disease state causes some specific protein expression changes in the blood, (Anderson et al. 2004; Voshol et al. 2005).

Because of theories such as this, the focus of cancer research has expanded from genetic information in the human genome to protein expression analyses.

As this 'proteome' reflects the state of a cell, tissue or organism more accurately, much is expected from proteomics to yield better tumour markers for disease diagnosis and therapy monitoring (Engwegen et al. 2006). Some successes were

claimed in the works by Rai et al. (2002) and Petricoin et al. (2002) who utilised proteomics technologies (SELDI-TOF Mass Spectrometry) to differentiate between malignant and benign ovarian cancers from the proteomic spectra taken from blood serum samples with a 94% and 95% success rate respectively. However a study by Kong et al. (2006) following these two previous reports found that those peaks identified as potential biomarkers were not as accurate as first thought unless combined with the established ovarian cancer biomarker CA125, where they aided in greatly enhancing the tests sensitivity and specificity, thus suggesting that identification of the proteins that caused these peaks could lead to the discovery of significant protein biomarkers for ovarian cancer.

One of the more important techniques in this aspect of modern biomarker discovery: mass spectrometry (and some of its variants) is described below.

1.6.1 Mass Spectrometry

Mass spectrometry is an analytical technique that identifies the chemical composition of a sample on the basis of the mass-to-charge ratio of charged particles (Sparkman 2000). The technique has many uses including:

- Identifying unknown compounds by comparing the mass of the unknown compound (or fragments of it) against databases of known compounds (and their mass spectra profiles).
- Quantifying the amount of a compound in a sample. This is a specialised field of mass spectrometry requiring specific methods and instruments, as

mass spectrometry is not inherently a quantitative method.

- Determining other important physical, chemical, or even biological properties of compounds with a variety of other approaches.

The other important detail to consider when looking at mass spectrometry with regards to biomarker discovery and analysis is that unlike qPCR and microarrays, mass spectrometry does not deal with the detection of genomic information, it is more used for the profiling of larger molecules such as proteins and metabolites.

Concepts of Mass Spectrometry

The basic principle of mass spectrometry is that charged particles (ions) from a sample are passed through electric and magnetic fields, the particles are then separated here as the paths taken through these fields will be dependant on their mass and charge (mass-to-charge ratio, also denoted as ' m/z '), finally the mass-to-charge ratio of each particle exiting the fields are recorded by a detector of some sort, resulting in an output from the experiment.

Mass spectrometers can be divided into three main parts:

- **Ionisation source** – This transforms the molecules of a sample into a variety of ionised (charged) fragments.
- **Analyser** – This section of the mass spectrometer sorts the ions by their masses and charges – this is typically done using electric and magnetic fields.

- **Detector** – This measures a given intensity value for the ions exiting the analyser.

There are many types of mass spectrometry systems available, differentiated by the particular methods used for ionisation and separation of the ions; these different technologies have traits that make them better or worse for different applications.

There are also two different applications of mass spectrometry in the realm of biomarker discovery. Identifying proteins and other molecules in samples that have been marked as being of interest, or directly using mass spectra results as a potential biomarker itself by comparing spectra from individuals with differing conditions to see if there is a distinct pattern to the data indicative of a given condition. It is this second approach that is more directly relevant to biomarker discovery as a field; two of the more common MS implementations in use in this area of modern biomarker discovery are described below.

MALDI-TOF Mass Spectrometry

Franz Hillenkamp, Michael Karas and their colleagues (Karas et al. 1985), developed the technique of Matrix Assisted Laser Desorption/Ionisation, Time of Flight (MALDI-TOF) mass spectrometry in 1985. The first part of the term, MALDI refers to the type of ionisation source that was developed, whereas the second part, TOF refers to use of a time of flight analyser to separate the ions produced from a MALDI surface (but was not created/developed by Hillenkamp and Karas).

MALDI is a 'soft' ionisation technique that allows the analysis of biological compounds that would normally be too fragile for use with traditional laser desorption ionisation (LDI). The method by which MALDI works is that samples are pre-mixed with an absorbing compound known as a matrix, and instead of the laser being used directly on the sample, it is instead used to bombard the matrix compound. The matrix transforms the laser energy into excitation energy for the sample molecules, causing the release of analyte and matrix ions from the surface of the mixture. This allows a form of protection for these more fragile molecules that would otherwise be unsuitable for MS analysis. A basic overview of a MALDI MS system is depicted in Figure 1-4.

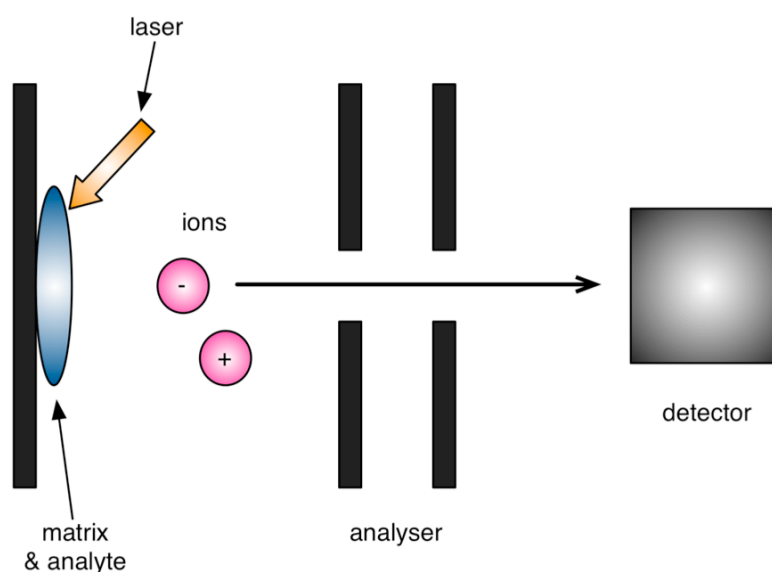


Figure 1-4: A simplified overview of the MALDI-TOF MS technique. The first step in MALDI MS is that the sample to be analysed is absorbed into the matrix. Secondly, a laser is used to bombard the matrix, thus causing an increase in excitation energy and the release of analyte and matrix molecules as charged ions. Finally, these charged ions are accelerated through a time of flight (TOF) analyser – an electric field that causes the ions to accelerate differently based on their charge and mass coupled with a detector to record the time taken for the ions to pass through the electric field – therefore allowing a mass-to-charge ratio (m/z) to be measured for each ion resulting from the sample and matrix.

TOF refers to the Time of Flight analyser that is normally used with MALDI MS systems (and other MS ionisation techniques). In a TOF analyser ions are accelerated through an electric field of known strength towards a detector. Ions of a different mass (as created in the ionization phase) will accelerate differently through the magnetic field based on their m/z ratio – ions with a lower m/z ratio will accelerate faster, whereas ions with a higher m/z ratio would accelerate more slowly. The time that it takes for the particle to reach the detector (at a known distance) is measured, and can therefore be used as a measure of m/z .

These two techniques together (MALDI-TOF MS) produce a complete analytical system that is capable of looking at the makeup of biological samples (made up of molecules such as proteins, peptides and sugars) with high reproducibility and great sensitivity.

SELDI-TOF Mass Spectrometry

Surface Enhanced Laser Desorption/Ionisation, Time of Flight (SELDI-TOF) mass spectrometry is an analytical technique often referred to as a 'Protein Chip' that was originally produced and developed by the company Ciphergen, but is now manufactured and distributed by Bio-Rad Laboratories, Inc (<http://www.bio-rad.com>). SELDI-TOF MS is commonly used in detecting proteins in tissue samples, blood, urine, or other clinical samples. Comparison of protein levels recorded between patients with and without a disease from SELDI-TOF MS can be used for biomarker discovery (Jr et al. 1999; Li et al. 2002; Kong et al. 2006).

SELDI MS is a variation of MALDI MS that is designed specifically for the analysis of protein mixtures. Prior to the addition of the matrix compound to the sample mixtures, the sample is exposed to a selective surface (a protein chip) that binds to specific types of proteins; common surfaces are as follows (descriptions taken from the UCL Institute of Child Health laboratory services webpage - <http://tinyurl.com/6lnzth>):

- **IMAC30** – *immobilized metal affinity capture array*

The IMAC30 ProteinChip Array is activated with transition metals prior to use (Copper, Gallium). Applications include histidine-tagged protein capture, metal-binding proteins, phosphoprotein profiling and protein purification.

- **CM10** – *weak cation exchange array with carboxylate functionality*

The CM10 ProteinChip Arrays incorporate a carboxylate chemistry (negatively charged) that acts as a weak cation exchanger. This array is used for protein profiling and protein purification

- **Q10** – *strong anion exchange array with quaternary amine functionality*

The Q10 ProteinChip Arrays incorporate a quaternary amine chemistry (positively charged) that acts as a strong anion exchanger. This array is used for protein profiling and protein purification.

- **H50** – *bind proteins through reversed phase or hydrophobic interaction chromatography*

H50 ProteinChip Arrays bind proteins through reverse phase or hydrophobic interaction chromatography and have binding characteristics similar to that of a C6 to C12 alkyl chromatographic resin. Typical applications include protein profiling of hydrophobic or membrane-bound proteins.

Following protein selection on these surfaces, non-bound proteins are then washed away, the matrix is added and then a MALDI-TOF MS is used to analyse the samples. This entire process is termed SELDI-TOF MS and is depicted in Figure 1-5.

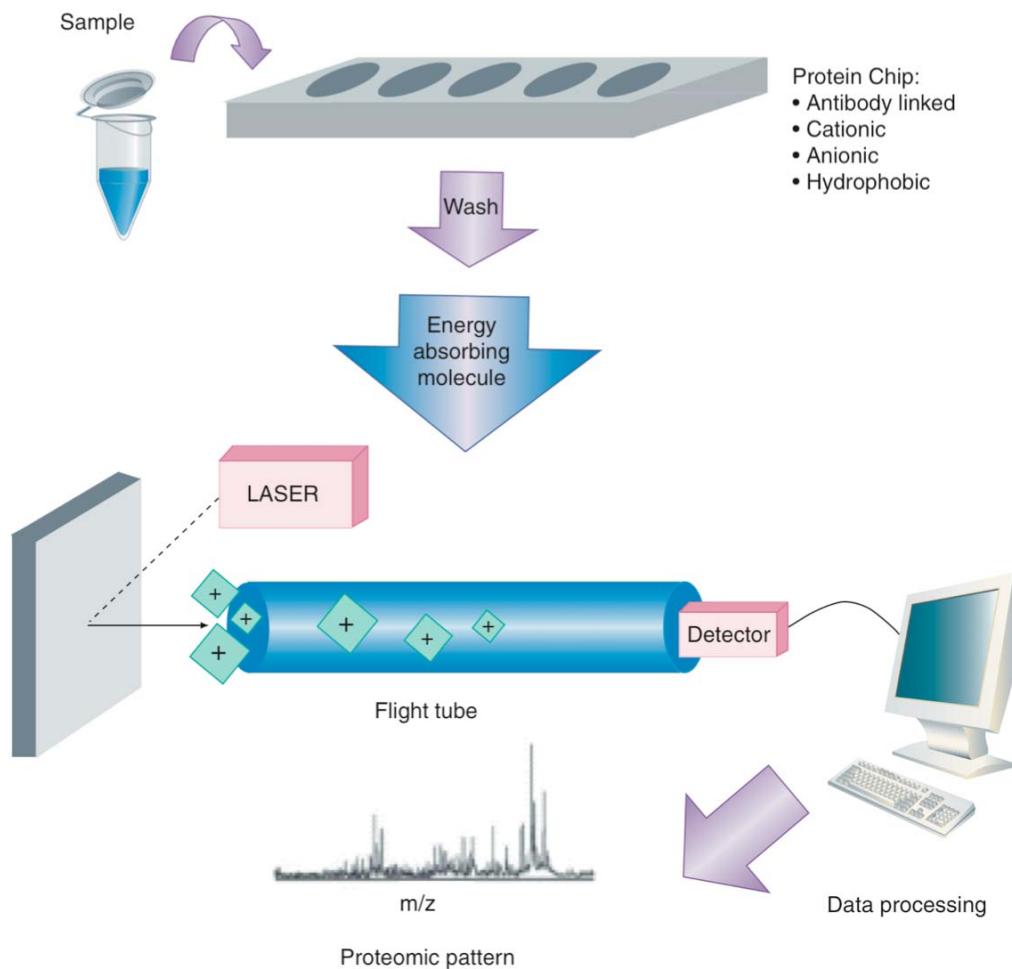


Figure 1-5: SELDI-TOF MS: Surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. A serum sample is loaded onto a chemically selective protein chip. The chip is then washed with buffer to remove unbound proteins and is treated with an energy-absorbing molecule, such as synapic or cinnaminic acid. The bound proteins are then ionized with a laser and accelerated through an electric field where they are separated based upon their mass/charge ratio (m/z). The detector quantifies the proteins, and a spectrum is generated with analytical software. Figure and caption taken from Simpkins et al. (2005).

Data Obtained from Mass Spectrometry

The data produced by mass spectrometry can be quite varied, but the most common form of data resulting from a mass spectrometry run is the 'mass spectrum'. This is a series of intensity measurements taken from the detector within an MS instrument over the course of the run – however, the measurements obtained are not a measure of intensity versus time, they are measurements of intensity versus the mass/charge (m/z) ratio of the ions in the MS instrument. This is due to the analyser within the MS instrument altering the flight patterns of the ions based on their electrostatic/magnetic properties through the instrument. An example of mass spectra can be seen in Figure 1-6, this represents the m/z and intensity measurements obtained during three SELDI MS runs. However, a point that must be noted is that a single MS experiment does not normally produce a single MS spectra; typically a single MS experiment would produce many MS spectra as it is normally used as an analysis technique to identify proteins or molecules following a separation step such as high performance liquid chromatography (HPLC) or polyacrylamide gel electrophoresis (PAGE).

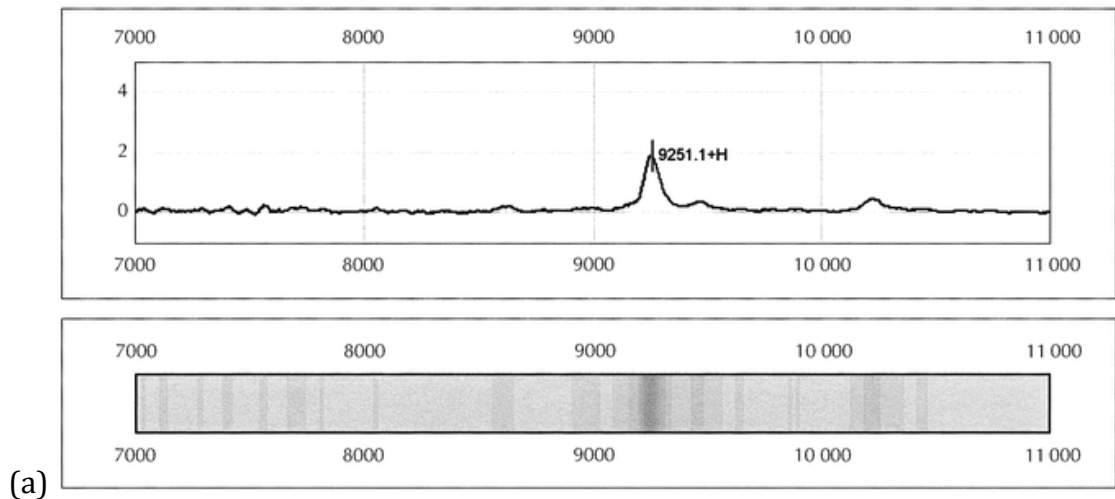
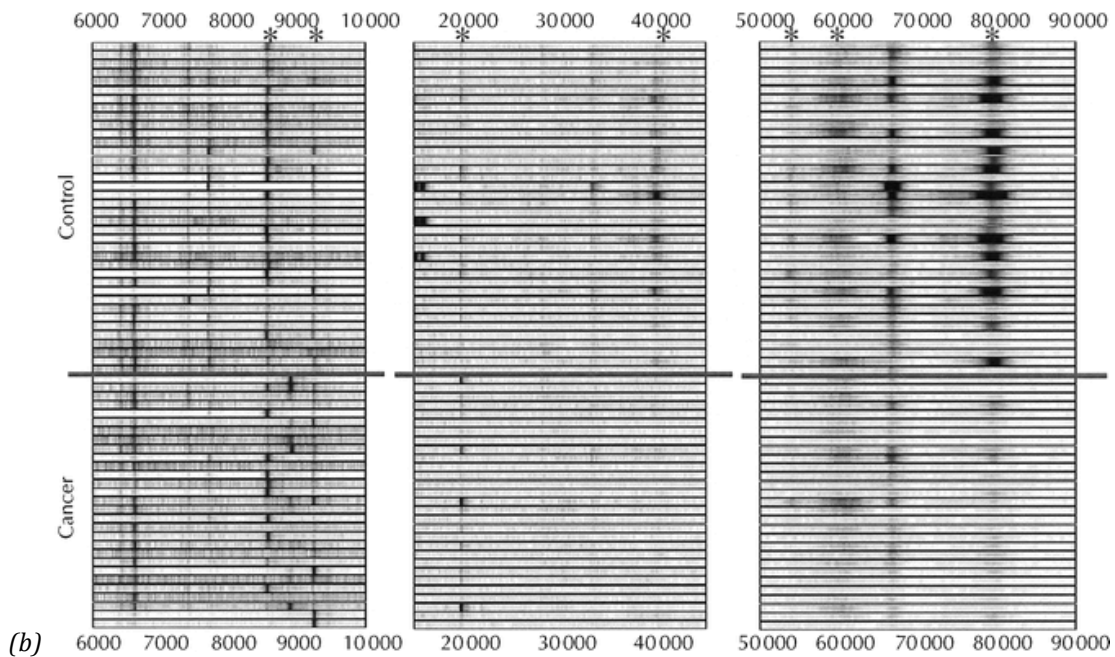


Figure 1-6: Examples of data generated by SELDI MS taken from the study by Rai et al. (2002) where SELDI MS was used to identify and test potential biomarkers for ovarian cancer from blood plasma. (a) A representative spectrum obtained from SELDI analysis (relative intensity vs m/z) generated from a blood plasma sample run on an immobilized nickel (IMAC-Ni) ProteinChip array. The upper panel shows a small portion of the protein profile in a 'regular' spectrum view, while the lower panel shows the same profile portion in a pseudogel view often used with SELDI analysis. (b) Stacked pseudogel views (patient vs m/z) of SELDI data from the analysis of 67 blood plasma samples showing the relative abundance of perspective biomarkers – indicated by the asterisks at the top of the figure. Samples from patients without cancer ($n = 38$) are shown above the solid line; samples from patients with cancer ($n = 29$) are shown below the solid line.

The amount of data that is produced from a single MS run can vary based on the quality of the detector in use, but a single mass spectrum data file is typically about 2Mb in size when in a delimited text format such as CSV (comma-separated variables) files. There are also other data formats in use, depending on the laboratory equipment and software used by individual laboratories that are based on XML (extensible mark-up language), these can help reduce the size of a raw data file to under 1Mb in size.

MALDI-TOF and SELDI-TOF Mass Spectrometry

The raw data that is generated by MALDI and SELDI MS systems is also in the form of intensity versus m/z ratio mass spectra. The major difference to consider with regards to SELDI MS in comparison to 'regular' MS is that the peptides observed within a SELDI MS experiment are proteins that carry specific properties based on the protein chip that was used in the immobilisation step of the experiment.

1.7 Data Management for Modern Laboratory Techniques

As described in the previous sections, the data that is generated from modern laboratory techniques is highly varied; as such this makes the process of storing, handling and managing these different types of data a great challenge. Each individual technique in use by a given laboratory has its own tailored individual data handling solutions or storage methods; often these are supplied/developed by the equipment manufacturer or by the laboratories using the equipment themselves.

This situation has been tolerated in the past as generally when looking into a particular condition, different experimental techniques and their data analysis are kept separate and not considered at the same time, and if they were to be considered together, it would be at a much higher, conceptual level once all of the data processing has been carried out.

However, this approach to research is beginning to change as more focus is being given to looking at different datasets in relation to each other. It has long been understood that the different parts of the biological machinery interact with each other, regulate and are regulated by each other (i.e. genes are regulated by proteins, and the proteins are produced via the genes that they regulate), such research is moving more to towards integrating these different techniques looking at genes, proteins and metabolites together to gain a holistic understanding of what is happening in the biological machinery.

1.7.1 Data Integration

As laboratories and scientific research in general move towards looking at more than just one aspect of research at a time, rather than just looking at transcriptomics, proteomics or metabolomics in isolation, new projects are evolving that study many different aspects of the biological machinery at the same time. These types of studies have come to be labelled “systems biology”.

Systems biology has been described as “the analysis of the relationships among the elements in a system in response to genetic or environmental perturbations, with the goal of understanding the system or the emergent properties of the system”, (Weston & Hood 2004). Although the realm of biomarker discovery does not fall directly into this description of systems biology, there are some common activities between these fields.

One commonality is the need for data integration – this is the primary goal of systems biology, to look at everything as a whole, singular system. This is also an approach that must be followed in the future of biomarker discovery as it is envisaged that future complex diagnostic tests could be made up of measurements of a series of genetic, proteomic and metabolomic markers to give an accurate picture of the progression of a disease, or to predict the likelihood of someone developing a disease – but with the increased complexity of such tests comes an equally so, if not more complex, development cycle. As such, developments in this arena for biomarker discovery could potentially benefit from utilizing ideas and even technologies developed for the field of systems biology.

Two of the more interesting developments with regards to data integration, housing and analysis are the Systems Biology Experiment Analysis Management System (Deutsch et al.) and Bioinformatics Resource Manager (Shah et al. 2007).

- ***Systems Biology Experiment Analysis Management System***

Systems Biology Experiment Analysis Management System (SBEAMS) is a modular database and software framework for collecting, storing, and accessing data produced by different experimental techniques developed by the Institute for Systems Biology (<http://www.systemsbiology.org/>).

Currently there are only basic modules for importing and working with data from microarray experiments (Marzolf et al. 2006), and from proteomics experiments in the form of mass spectra. There are also facilities for importing and organising annotations around the genes and proteins housed within the SBEAMS system from external database sources available on the Internet. A diagram of the concepts of SBEAMS is shown in Figure 1-7, however, not all of the proposed functionality shown in the diagram is available as yet.

- ***Bioinformatics Resource Manager***

The Bioinformatics Resource Manager (BRM) is a software environment that, like SBEAMS, provides the user with data management, retrieval and integration capabilities. BRM also allows the integration/management of both microarray and proteomic data and allows the user to integrate their data with functional annotation and interaction data from public sources. In addition to this BRM provides connectivity to common analytic tools

(such as Cytoscape, Shannon et al. (2003); Christmas et al. (2005)) through the outputting of data readily formatted data for these tools. An overview of the BRM system is shown in Figure 1-8.

Both these systems allow for the integration of microarray and mass spectrometry data, and also provide functionality for the analysis of these datasets.

The concepts shown in both of these systems demonstrate the kind of ideas and concepts that need to be applied to biomarker data storage and analysis solutions. However, the implementation of these two systems is not suitable for applying to biomarker datasets as they are geared firmly for the needs of systems biology through looking at the interactions found in data rather than looking at how the data effects an endpoint in the host environment; and extending or modifying the systems to become more suitable would be a very demanding task as both are highly complicated (and in the case of SBEAM, very incomplete). Therefore, a suitable candidate for housing and analysing different types of biomarker datasets should take ideas and concepts from these platforms for the combining of datasets, but have a sharp focus on analysing data for the needs of biomarker identification and diagnostic test development.

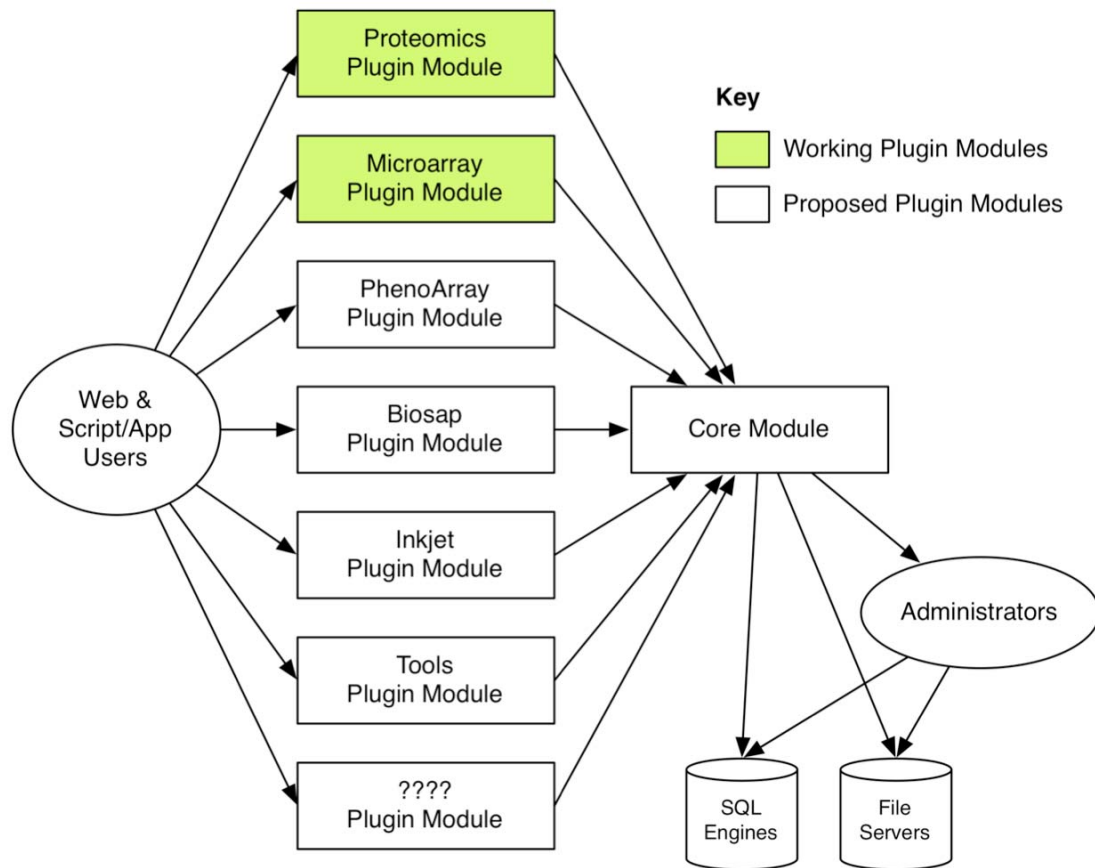


Figure 1-7: An overview of the SBEAMS modular structure concept. This represents the proposed functionality that SBEAMS will eventually provide – modules for many different types of analysis including genomics/transcriptomics, proteomics and many more. However, work on the SBEAMS project is at a very early stage. Currently the only usable components of the system are the proteomics and microarray modules. Image adapted from the SBEAMS website (<http://www.sbeams.org/>)

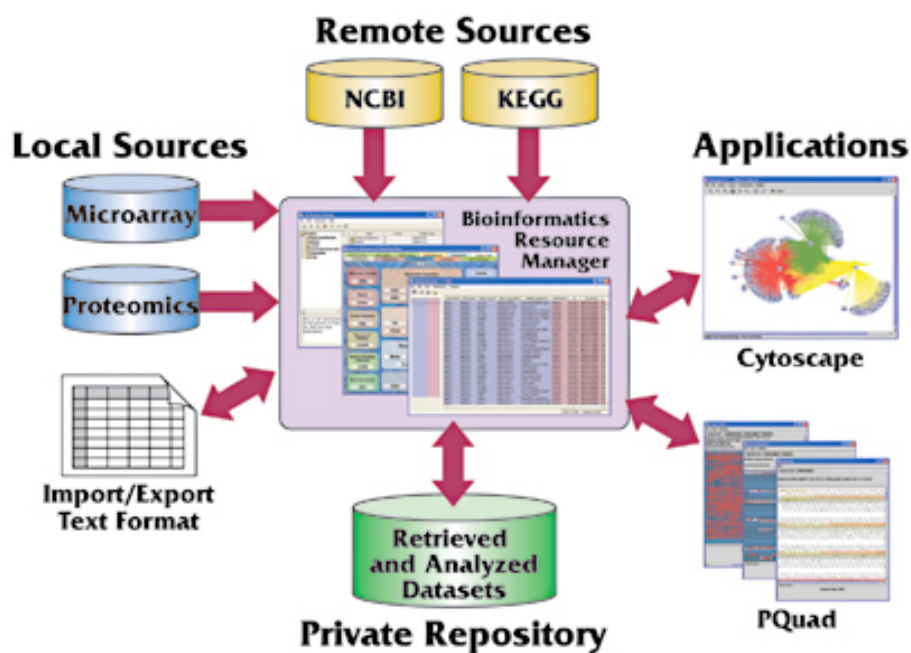


Figure 1-8: An overview of the Bioinformatics Resource Manager and the potential connections that can be made between disparate data with it. BRM allows the combination of microarray and proteomic data from local repositories, with information from remote sources such as the NCBI (The National Center for Biotechnology Information - <http://www.ncbi.nlm.nih.gov/> a repository of gene, protein SNP and other data) and KEGG (The Kyoto Encyclopedia of Genes and Genomes - <http://www.genome.jp/kegg/> which contains detailed information about biological systems and pathways) so that they can be analysed effectively with tools such as Cytoscape (<http://cytoscape.org/>). Image taken from the Pacific Northwest National Laboratory website (<http://www.sysbio.org/>).

1.8 Working With Data From Modern Laboratory Techniques

As has been discussed, all of the newer laboratory techniques involved in modern biomarker discovery generate data for multiple variables in a single experiment, (many thousands in the case of techniques such as microarrays), and as such the traditional univariate approach of simply comparing measured variables and their correlation with a given state or condition is unlikely to be sufficient for all applications.

The data analysis that is performed with these new technologies is known as multivariate data analysis. Multivariate data analysis deals with observations on more than one variable where there is some inherent interdependence between the variables (Mardia et al. 1980) and is therefore well suited to the types of bioanalytical techniques in use today.

Although the term ‘multivariate analysis’ is typically used to describe statistical techniques (multivariate statistics), its definition can also be used to cover a broad range of data analysis techniques useful in this scope of work (biomarker discovery and classification system development) as the true goal is that of pattern recognition through trying to identify any underlying patterns of information within the data that has been generated. Described below are some of the more relevant and used techniques in this field.

1.8.1 Multivariate Statistical Methods

The term multivariate statistics describes a collection of statistical procedures that involve the observation and analysis of more than one statistical variable at a time. There are many different models that are used in this field. However, two of the more significant in the biomarker discovery field are principal components analysis and partial least squares discriminate analysis.

Principle Components Analysis

One of the first techniques usually considered when working with multivariate data is the exploratory data analysis technique of principal components analysis (PCA) as it allows the visualisation of relationships between individual samples based on data from a number of given variables.

The statistical origins of PCA are in biology and psychology; psychometricians have for many years had the need to translate numbers such as answers to questions in tests into relationships between individuals: How can verbal ability, numeracy and the ability to think in three dimensions be predicted from a test? Can different people be grouped by these abilities? And does this grouping reflect the backgrounds of the people taking the test? Are there differences according to educational background, age, sex or even linguistic group? (Brereton 2003). All of these are the types and kinds of questions that PCA allows researchers to ask of their data.

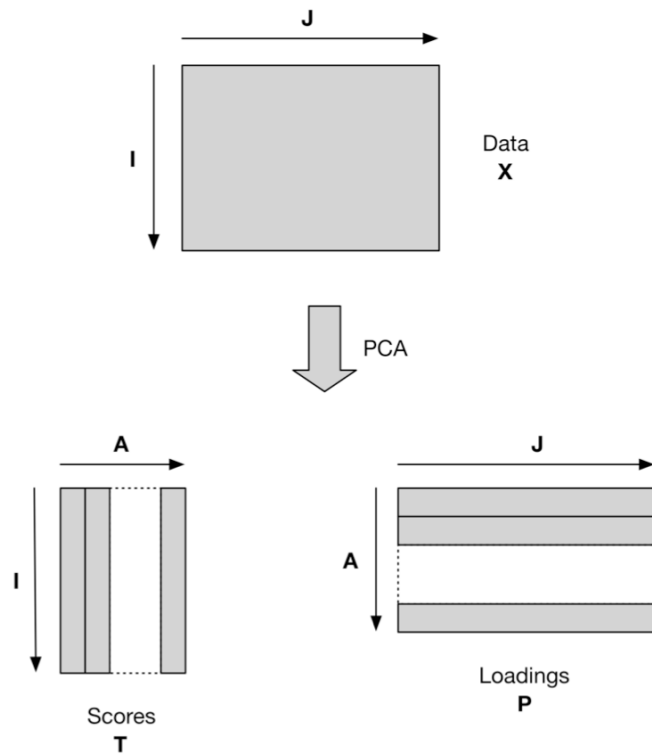


Figure 1-9: Principal components analysis (PCA), image adapted from Brereton (2003). This figure shows the basic principle of PCA – the decomposition of the original data matrix X into the scores and loadings matrices (T and P respectively). The original data matrix represents the recorded values of experimental variables (J) for different samples (I). This is then transformed into the scores matrix (T – samples, I versus scores, A), and the loadings matrix (P – variables, J versus scores A).

The mathematical principle behind PCA is that the multivariate data can be decomposed by linear projections onto a new co-ordinate system. The new axes, known as principal components (PCs), are orientated such that the first PC captures the largest amount of common variance. The next PC is orthogonal (meaning totally uncorrelated) to the first and captures common variance in its direction. The maximum number of PCs is limited to the number of variables in the original multivariate data set but due to the reorientation of the co-ordinate system to maximise common variance. The majority of the data variance can be captured by a much smaller number of principal components because of correlation and redundancy in the original data. In mathematical terms, PCA is the reduction of the data matrix, $\mathbf{X}$, into two smaller matrices, the scores, $\mathbf{T}$, and loadings, $\mathbf{P}$. This can be observed in Figure 1-9, (Brereton 2003).

There are a number of algorithms for calculating $\mathbf{T}$ and $\mathbf{P}$, the most common being nonlinear iterative partial least squares (NIPALS) and singular value decomposition (SVD). In both cases, $\mathbf{T}$ is a matrix of column eigenvectors, with the columns in order of largest variance first, hence PCs are delivered in order of their information content. The scores matrix $\mathbf{T}$ is determined by multiplying $\mathbf{X}$ by the matrix of loadings, $\mathbf{P}$. In simple terms, this means that the scores for a particular sample are weighted sums of the original variables (Brereton 2003).

PCA can also be used as a form of data reduction. PCA allows the reduction of the original (large) multivariate data matrix $\mathbf{X}$, into a matrix with a much smaller number of variables, whilst retaining the important information within the data, the scores matrix, $\mathbf{T}$. Thus, with particularly large datasets, such as mass

spectroscopy results, PCA would allow for a more clear analysis of the data using the resulting matrices.

Partial Least Squares Discriminate Analysis

Partial least squares discriminate analysis (PLS-DA) is a statistical classification technique based upon the regression technique partial least squares (PLS). In order to best describe PLS-DA, it is better to first define the basic principle of PLS.

PLS

PLS was developed in the 1960's by Herman Wold, and is a multivariate regression technique that is similar in appearance to PCA. It is often used in the field of chemistry and chemometrics for calibrating equipment, which is required to return quantitative information from more than one measured variable.

The principles of PLS are fairly straightforward in that two sets of models are created for both the ***X*** (measurements for each compound in a mixture) and ***Y*** (concentrations for each compound in a mixture) variables as follows:

$$X = T P + E$$

$$Y = T Q + F$$

where the product of ***T*** and ***P*** approximates to the measurements and ***T*** and ***Q*** approximates to the concentrations; the common link between the two is the scores matrix ***T***. It is here that we need to note that ***T*** and ***P*** for PLS are different to ***T*** and ***P*** obtained in PCA, as unique sets of scores and loadings are obtained for each compound in the dataset. Hence if there are ten compounds of interest there

will be ten sets of T , P and Q . It is in this way that PLS differs from PCA where there is only one set of T and P , the PCA step taking no account of the Y block (Brereton 2003).

The algorithm for PLS is known as the PLS2 algorithm, which was developed in the 1980's and is a slight improvement on the original PLS algorithm (known as PLS1). There are several different algorithms for use with PLS, all with differing advantages and disadvantages; the more commonly used are the PLS2 algorithm, and a more recent variation on it known as SIMPLS, (De Jong 1993).

PLS-DA

As described above the application of PLS is normally calibration, as it is used to output a continuous value (e.g. concentrations of key analytes over time). This method however has been adapted for the purpose of classification of linearly separable groups; this is the function of PLS-DA, and has been proven to be of use in analysing biological data (Bo Nordén 2005; Karp & Griffin 2005; Musumarra et al. 2005; Marengo et al. 2008).

The way in which the method of PLS had been adapted to perform classification is very simple – the Y matrix is replaced with a matrix indicating class membership for each compound of interest (typically made up of ones and zeros). PLS is then performed on both the X and the new Y matrix, and the regression values obtained for Y can be used to predict the class of unknown related compounds.

Optimising PLS-DA Models and Cross-Validation

When working with PLS-DA (and most other classification tools), prior to producing the best classification model possible, a form of optimisation must be carried out. In PLS-DA, this is the selection of the number of latent variables (LVs) to use in the final PLS-DA model – LVs are analogous to PCs in PCA, in that the number of LVs in a data matrix are equal to the number of variables, and each contributes a calculated percentage of the variance in the dataset.

It is the selection of the LVs that contribute the most variance to the dataset (and excluding less important LVs) for use in the final PLS-DA model that can enhance a model's performance. The only way to choose the most appropriate number of LVs is to test a series of models built using different numbers of LVs. Typically a form of cross validation is used to assist in selecting the optimised parameters for a PLS-DA model as this helps to guard against any errors occurring due to extreme outliers in the data. Two of the more common forms of cross validation are:

- ***n*-Fold Cross-Validation:** In *n*-fold cross-validation the training dataset is split into *n* groups. One sample group is removed from the pool and retained as a validation set for testing the model – the remaining groups are used to train the model. This process is then repeated *n* times, with each of the *n* groups used once as the validation dataset. The results from the *n* number of models can then be averaged to give an estimation of the accuracy of the model.
- **Leave-One-Out Cross-Validation:** The principles of leave-one-out (LOO)

cross-validation are very similar to n -fold cross-validation. The difference is that the whole training dataset is used as a group and only one sample is retained as the validation dataset. The cross-validation is then repeated until each sample in the dataset has been used as the validation set once, then the results of the models can be averaged to give an estimation of the accuracy of the model. This is the same in principle to n -fold cross-validation if n was equal to the number of samples in the training dataset.

1.8.2 Machine Learning

Machine learning is a subfield of artificial intelligence that is focused on the development of algorithms and techniques that allow computers to “learn”. What do we mean by learn? The intention is that should a properly developed system be produced, it could extract relevant and important information, (such as rules and patterns) from a set of data automatically. This is achieved through the use of computational and statistical methods; hence, machine learning is closely related not only to data mining and statistics, but also theoretical computer science.

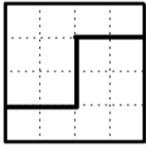
Machine learning has many applications including natural language processing; image recognition; search engines; bioinformatics; medical diagnosis; chemoinformatics; detecting credit card fraud and financial analysis. From the point of view of biomarker analysis and bioinformatics, it's the pattern recognition and classification techniques that are of the greatest interest, two of the more common techniques used are artificial neural networks (ANNs) and support vector machines (SVMs).

Artificial Neural Networks

Artificial Neural Networks (ANNs) are mathematical or computational models based originally on biological neural networks (i.e. the human brain). They are made up of an interconnected group of neurons that in essence combine to become a non-linear data modelling tool that can be used to model complex relationships between sets of input and output data, or find patterns within data that are of importance (MacKay 2002).

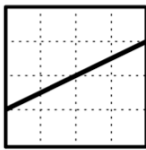
The architecture of an ANN is deceptively simple. At the most basic level there is the neuron, a simple processing element that is linked to many other neurons to form a complex network. The network itself is made up of layers of neurons that are interconnected to each other using weighted values established during the training of the network and influence the overall behaviour of the system, this is depicted in Figure 1-10.

A neuron can also be described as a 'transfer function' as it is in essence a simple mathematical function that receives one or more inputs, and then sums them to produce a single output which then passes onto the following layer of the ANN (The MathWorks 2009). There are several common types of neurons; combinations of these neurons are what give ANNs their analytical power:



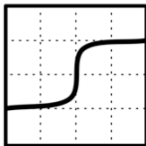
Step Function

The output of this neuron is binary, depending on whether the inputs meet a specified threshold. If this threshold is met, the output is one, if not, the output is zero.



Linear Function

The output of this neuron is simply the weighted sum of its inputs plus a bias term dependent on the linear function.



Sigmoid Function

These neurons represent simple non-linear functions. Like the linear functions the output of this neuron is simply the weighted sum of its inputs plus a bias term dependent on the sigmoid function.

ANNs have many advantages in that they are very powerful when trained correctly and are able to solve both linear and non-linear problems, but they also have disadvantages in the fact that they are essentially a 'black box', once built and implemented it is not easy to determine how they are working (unlike statistical methods), they can also be slow to train, and are prone to over training so great care must be taken when implementing ANNs.

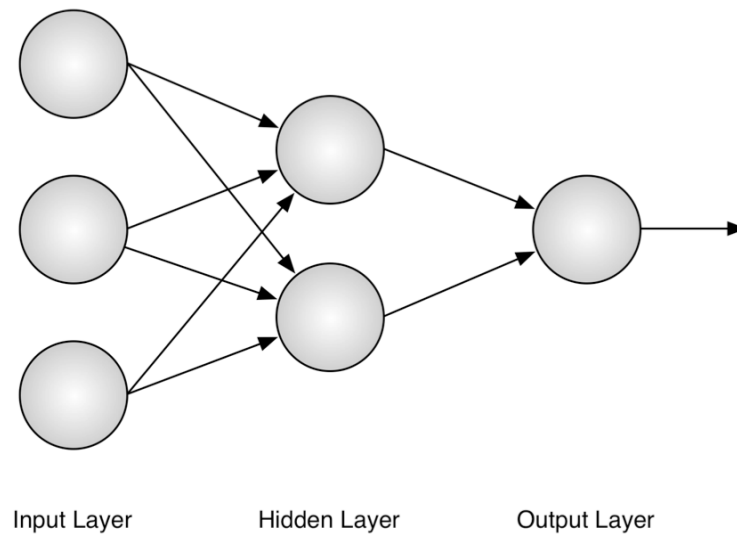


Figure 1-10: A basic representation of a multi-layered artificial neural network. The interconnections between the neurons are weighted during training using error functions in order to map the inputs and outputs of the network correctly.

Support Vector Machines

Support Vector Machines (SVMs) (Vapnik 1998) are supervised learning machines for regression and classification problems that were introduced in the 1980s; their main application in the field of biomarker discovery and biological science in general however are in the role of classifiers (Noble 2006).

SVMs are capable of solving both linear and non-linear problems. The simplest way of describing how an SVM works is an example of linear classification, which is depicted in Figure 1-11. As is shown the central line (the maximum margin hyperplane) splits the two classes, while the two support hyperplanes maximise the space between the central hyperplane and each class. When an unknown sample is presented to the SVM, it is possible to classify it as either -1 or +1 dependent on which side of the support hyperplanes the sample falls on (and therefore which class it belongs to).

As said previously, SVMs are capable of solving both linear and non-linear problems. This is achieved by taking an n -dimensional input space and mapping it to a higher dimension – often referred to as the kernel trick – and once in this space, construct a linear classifier to solve the problem. This function of SVMs is depicted in Figure 1-12.

The key part of the SVM is the kernel – the mathematical function that performs the transition from a linear space to a non-linear space. The majority of SVM architectures implement one of four kernel functions, polynomial, radial basis

(RBF), Gaussian radial basis and sigmoid; in most cases the RBF is a more than suitable starting point as it is capable of handling both linear and non-linear relationships and is the kernel most used in biological applications (Balding et al. 2007).

SVMs have many advantages in that they are able to solve both linear and non-linear problems, can use different kernels more suited to a particular dataset, are much simpler to implement than ANNs and are less prone to over training compared to ANNs. They do however have some disadvantages; like ANNs, once trained they are 'black boxes'; they are mathematically highly complex and are very computationally expensive, so they can take a long time to train and optimise.

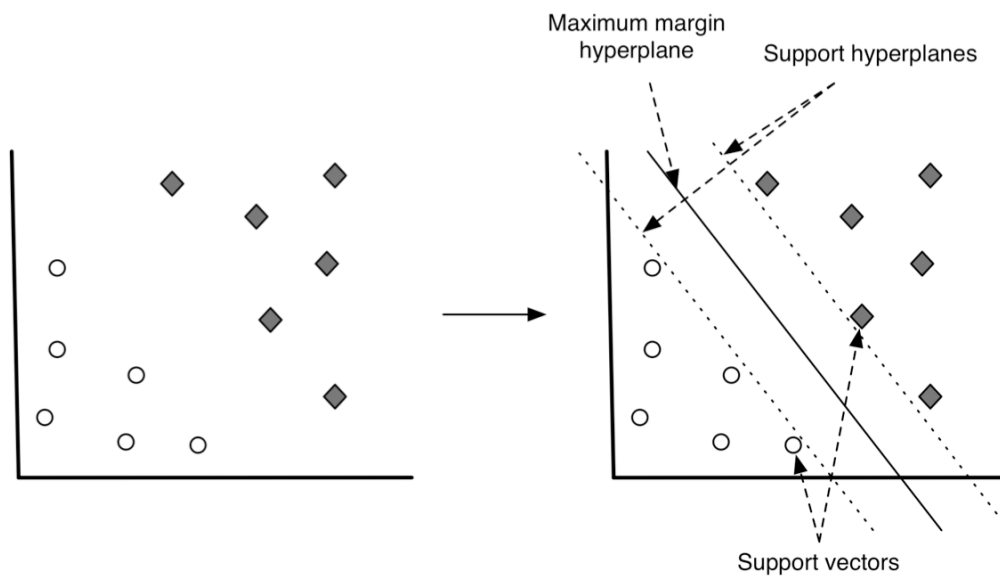


Figure 1-11: An application of an SVM to a linearly separable problem. In the plot on the left there are two groups that are separable in linear space. The plot on the right shows how a SVM defines the relationship between the two groups. The two classes are separated by a central plane, and to either side of the dividing plane are parallel 'support' hyperplanes that define the maximum distance possible between each class and the central hyperplane – the space between the support hyperplanes and the dividing hyperplane contains no data points. The name support vector machine comes from the name given to the data points that lie on the support hyperplanes and define their position, these are known as the support vectors.

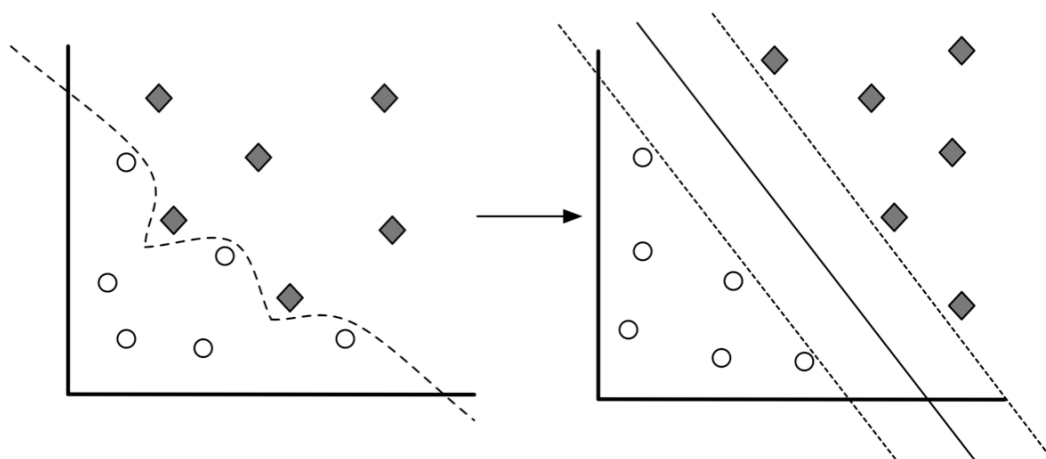


Figure 1-12: An SVM solving a non-linear problem with a linear classifier. The approach the SVM uses to solve this problem is to map the data points to a higher dimensional space where linear separation of the data can be achieved.

Optimising SVM Models

As with statistical classification tools such as PLS-DA described earlier, there is a degree of optimisation that can be carried out when building an SVM model to ensure that the best model possible is produced.

One of the first steps in optimising an SVM is the choice of kernel to use as the different kernels that are available are made up of different mathematical functions, the four most common types of kernels in use in SVMs are:

- Polynomial: $k(x, x') = (x \cdot x')^d$
- Radial Basis Function (RBF): $k(x, x') = \exp(-\gamma \|x - x'\|^2)$, for $\gamma > 0$
- Gaussian Radial Basis Function: $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$
- Sigmoid: $k(x, x') = \tanh(kx \cdot x' + c)$, for some (not every) $x > 0$ and $c < 0$

When presented with different datasets, each of these kernels will perform differently, and some will be more suitable than others based on the nature of the dataset in question. As stated above, a good starting point when creating a new SVM on as yet untested data is the RBF kernel as this kernel is able to separate both linear and non-linearly separable data, there are however parameters (C and γ) that need to be optimised in order for this kernel to perform efficiently – optimal values for these parameters are usually obtained using simple grid searches: testing pairs of values for C and γ until an optimal pairing is found.

1.9 Project Aims

Following on from the topics that have been discussed in this chapter, the aim of this thesis is to produce a combined data handling and analysis platform that allows the integration of data from differing laboratory analyses techniques for biomarker discovery and classification tool development. Four steps to achieving this aim have been identified:

1. Design and construct a database schema that is flexible enough to be able to store and serve data from a variety of laboratory analysis techniques used in biomarker discovery and analysis.
2. Program an API for interacting with the database schema developed in objective 1. This should allow for the loading of data into and retrieving of data from the schema, thus allowing for the flexible construction of interfaces and data analysis routines.
3. Use the developed data handling solutions (database and API) and construct data analysis routines for working with data from two different medical case studies.
4. Evaluate the performance of the systems developed and results obtained in the case studies using real medical patient data.

This series of objectives is represented in the structure of this thesis. Chapter 2 describes the work produced towards objectives one and two, whereas Chapters 3 and 4 describe the case studies carried out for objective three; the final objective is

discussed in Chapter 5. It is envisaged that the completion of these objectives will provide an indication of the feasibility of a combined data housing and analysis platform produced in this thesis within the field of biomarker discovery and analysis.

2 Development of a Heterogeneous Data Management System and Analysis Platform for Biomarker Identification and Analysis

Typically in biomarker studies that utilise more than one laboratory technique there are separate systems and software (for each technique) in place to assist in the cataloguing and storage of raw data; this is normally in the form of commercial/proprietary software packages supplied with the laboratory equipment itself; in the case of techniques such as microarray analysis and certain types of mass spectrometry there are also online data storage repositories where scientists are encouraged to upload their data to in order to allow others to view and utilise it. However, these are all very separate systems and are not linked in any way.

When it comes to the analysis of these datasets, several of the common tools in use were discussed in the previous chapter, numerous packages for this analysis are available such as technique-specific tools like GeneSpring for microarrays or more general statistical analysis packages like R and Matlab, however the analysis employed is generally in isolation from other experiments and datasets; there is no (published) common framework linking experiments and investigations to aid in increasing the speed of such research. Two solutions developed in the field of systems biology were discussed in Section 1.7 , these were SBEAMS and BRM, and

go a long way towards integrating data aimed at interaction analysis, however these systems are not ideally suited for biomarker studies. As such, this chapter presents the tools developed in this thesis to fill this requirement.

2.1 A Different Approach:

XPA – Cross-Platform Analysis

XPA (pronounced ‘cross-platform analysis’) is the name given to the generic biomarker identification and analysis platform presented in this study. The theory behind XPA is relatively straightforward, it is a database system designed to house the most common type of data generated in modern laboratory techniques and ‘omics’ analysis – expression and quantity measurements, coupled with scripts/interfaces for loading and analysing the resulting datasets. An overview of this basic concept is presented in Figure 2-1.

The type of data that would be considered suitable for such an approach is quite broad; microarray expression data, mass spectrometry data, and any other type of analysis where the end result is a value that records the quantity or presence of a substance in a sample, (such as the common techniques described in the first chapter). Then once they are stored in the common XPA schema, subsequent analysis of the datasets can be achieved using standardised analysis routines – this approach also opens up the potential of performing analysis on combined datasets from more than one experimental procedure.

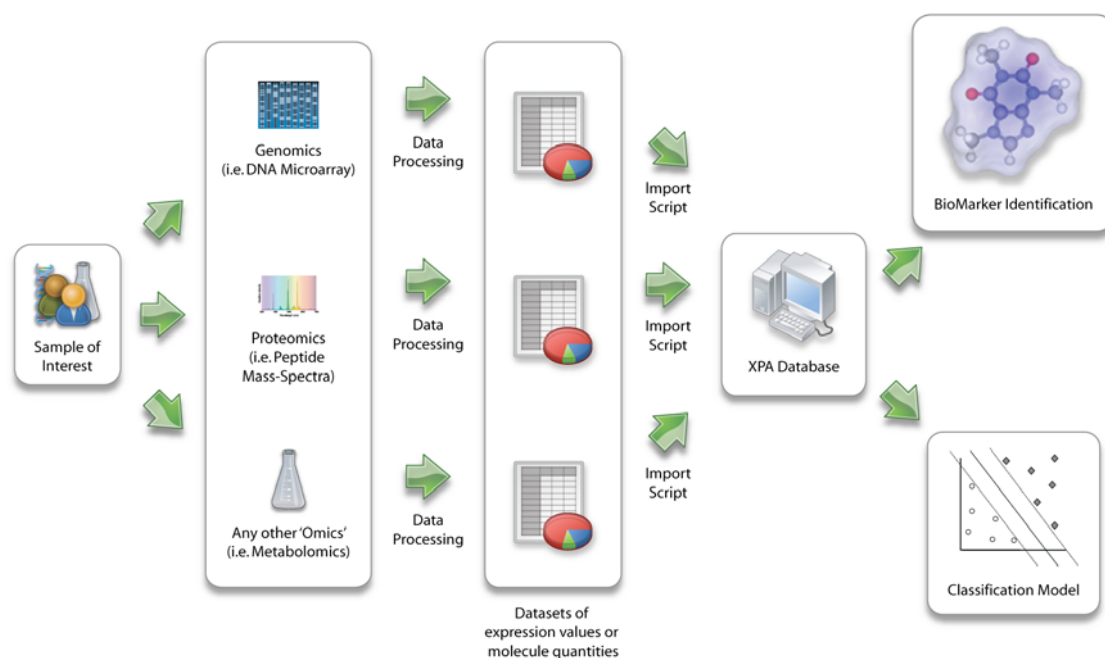


Figure 2-1: An Overview of the XPA vision and process flow. Patient samples and information about the patients, and samples themselves are combined with various types of experimental data from different techniques within the XPA database schema. From here different data analysis approaches and investigational techniques can be used on this data to develop classification models or identify points within the data that could be considered putative biomarkers.

2.2 XPA System Overview

The XPA system is proposed as a central database storage area for housing experimental results and related metadata, accompanied by interfaces for loading data into the system and analysing the datasets. This section of the report describes the details of the XPA database schema, and the layers of software that are intended to interact with the XPA database.

A schematic outlining the layers of the XPA model is shown in Figure 2-2, in which it can be observed that the proposed system is made up of essentially three layers: the database, a software layer (APIs for interacting with the data store) and an interface layer for interaction with end users and for carrying out data analysis tasks on the data held with the XPA database. Each section is now described in more detail.

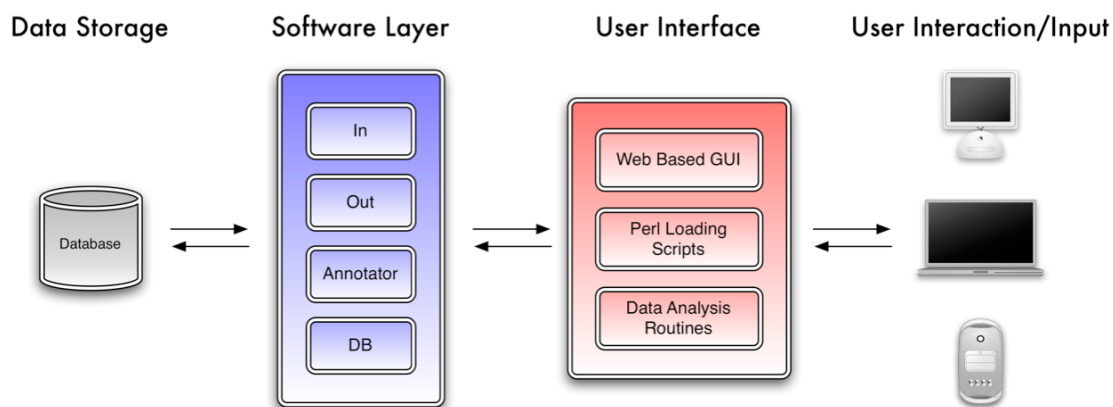


Figure 2-2: A diagrammatic representation of the XPA system (from a software point of view). This shows a relatively common approach to software construction. The most important aspect of the system is the central database that houses all of the patient, sample and experimental data. This is then wrapped and interacted with via a software layer (often called a 'middleware' layer) as this allows interfaces and data analysis routines to be constructed easily without taking into account basic interactions with the database itself – this would be taken care of by the XPA API.

2.2.1 Class and Database Structure

The database design of XPA was built around a group of related classes of information that are used together to make up all of the information needed to efficiently describe an experiment. Inspiration for this approach to modelling the data, and some of the table structures used to represent these classes were taken from the FuGE – ‘Functional Genomics Experiment’ project (A. R. Jones et al. 2007). An overview of these classes and their relationships is presented in Figure 2-3.

The main classes of data in the XPA system are as follows:

- Investigation
- Experiment
- Protocol
- Patient
- Dataset
- Molecule
- Ontology

Each class of data with the XPA schema and its relevance in the overall XPA system is described in more detail in the following sections.

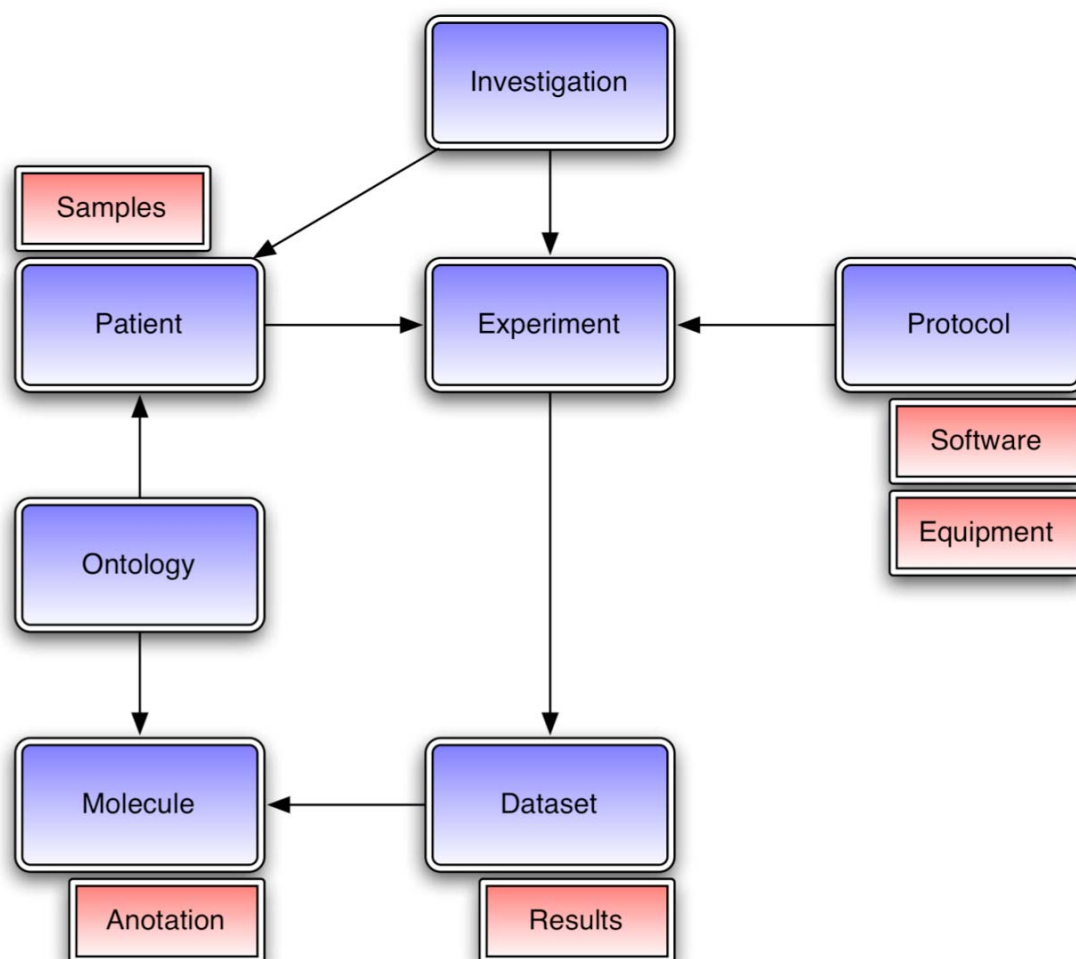


Figure 2-3: A flowchart giving a representation of the class structure used in the design of the XPA database. The Blue boxes represent the main classes of data, with the arrows defining the flow of data between them. The red boxes represent extra detail and data that are related to the classes that they sit next to.

Investigation

‘Investigation’ is the highest level of class in the XPA schema. It was envisaged that XPA would be used to house experimental data from more than one project in an organisation, so the concept of investigations was introduced as a way of grouping together related experiments.

Experiment

‘Experiment’ is the second highest level of class in the XPA schema, and is the central class in XPA – everything is based around the concept of an experiment, and the experiment class (and its related classes) provide a way of encapsulating all of the data for a given laboratory experiment, and multiple experiments can be assigned to individual investigations.

Protocol

‘Protocol’ is the class used to describe the protocol of an experiment, the methods, the equipment used and any software required to record/interpret the data.

Although this class (and type) of data is not essential for the main purpose of XPA (to analyse data), it was deemed useful information to record as it allows a way of recording how each experiment was carried out within the XPA schema. The benefit of this is twofold in that it could aid in diagnosing faults with equipment or software if there were common errors occurring during experiments, and also allows the database to be more effectively mined by experimental procedure.

Patient

‘Patient’ is the class used to record a basic identifier for a given patient taking place within a study. Individual patients can also be associated with more than one experiment or investigation as it is not uncommon practice for researchers to use data from individual patients in more than one investigation. A subclass of patient in the XPA model is ‘sample’ as each patient could potentially have more than one sample or even sample type associated with them – this set-up makes XPA flexible enough to easily allow and record the re-use of patients and samples in multiple experiments and investigations.

More detailed patient data such as age, weight, height, place of birth or any other parameter that would be considered suitable to record for a given study is associated with each individual patient and sample record via the use of the ontology class (described below) via the use of controlled vocabularies for each given data type. It is also envisaged that the sample grouping (i.e. test vs control) would be recorded within such a manner using the ontology classes.

Dataset

‘Dataset’ is the class used to record the raw and processed results from an experiment. In the XPA schema, a dataset (also known as a ‘results_set’ within the database table structure) is made up of individual ‘results’ where a result is the recorded measurement for an individual experimental variable (in XPA, each variable relates to a ‘molecule’, the class for which shall be discussed shortly).

However, it is not the intention of the XPA schema to be a housing area for raw data as producing an efficient data housing solution for each possible type of raw data that XPA is envisaged to handle would be an extremely complicated and would not add any extra benefit to the overall function of XPA. Efforts have been made towards this in the SBEAMS system described in Section 1.7 but this comprises of many very distinct (and separate) database schemas used for housing each data type used, and a central database schema used to house linking and management information to this data.

Instead of reinventing storage solutions for data types that already have appropriate storage mechanisms in place, the approach that has been developed for XPA is for the raw data to be stored in text files (outputted straight from a laboratory instrument), or in a specialised data storage solution, specific for the type of experimental data, and for a link to this data (i.e. the filename/storage location, or a database ID) stored in the XPA schema, and instead only store processed data (i.e. the significant data that is intended to be analysed) in the 'Results' subclass, thus reducing the amount of storage space taken up by XPA, and making the storage of laboratory data as streamlined as possible.

Molecule

'Molecule' is the class within XPA used to define an experimental variable, and can be used to represent anything from a gene (i.e. in genomics/transcriptomics studies), to a protein (in proteomics studies), to metabolites (in metabolomics studies) – whatever the molecule of interest in the given experiment is. As stated

before, molecules are linked to experimental results via the 'Results' class (a subclass of Dataset), and further information about each molecule can be recorded in the associated 'Annotation' subclass, and any classification data can be recorded in the 'Ontology' class.

Ontology

'Ontology' is the final class within XPA; this is used to store data from controlled vocabularies commonly known as ontologies. Capturing ontological information in XPA (related to both patient/sample data and 'molecules'), allows the data within XPA to be classified in an organised fashion – i.e. all samples relating to cancers would have an ontological entry from the same family, or all molecules involved in a metabolic pathway that have been studied in some way in an organisation would be easily identifiable via the ontology class.

The XPA Database

The XPA database schema is shown in Figure 2-4, it shows the entire design and layout of the XPA database, and the information that it is intended to house. Each class of data (as described above) is represented within the schema by a collection of tables, and where possible, the classes are identified in the figure by grouping the tables together in coloured boxes.

The final design of the schema was formed following the production of numerous lists containing the data that should be stored for each class of data, these lists were then normalised into database tables (in the third normal form) and linking tables created to connect classes that needed to be connected to each other.

2.3 Data Handling

The flow of data in and around the XPA system is all handled by a series of Perl modules. These can be classified into three main categories: data input, data output and annotation; everything else is classified as ‘miscellaneous’. The data input and output modules control all of the interactions with the XPA database; the annotation modules govern the process of annotation within XPA (gathering extra information about the molecules housed within the system); and the miscellaneous modules are general helper routines used by the other three groups.

Interaction with the database itself (the input and output modules) is handled through the use of the Object Relational Mapping (ORM) facilities that are part of the “DBIx::Class” Perl modules*. The benefit of using this approach for interacting with the XPA database is that it reduces complexity for the programmer when interacting with a large database, by mapping each database table to a programmatic object and therefore each row within that table as an instance of that object type. It greatly reduces the need for low-level interaction with the XPA database (writing SQL statements and querying the database directly through DBI), and therefore reduces the amount of Perl code that is needed to be written, makes the existing code more easy to maintain, and makes extending the system easier.

An overview of the Perl modules created for interacting with the XPA database and the functions contained within is presented in Figure 2-5. This shows that there

* <http://search.cpan.org/dist/DBIx-Class/lib/DBIx/Class.pm>

are three main groups of functions:

- **XPA** – This is the main group of the Perl modules (and functions) containing three perl module files:
 - o **XPA::In** – Functions responsible for entering data into the XPA database.
 - o **XPA::Out** – Functions used for retrieving and exporting data from the XPA database.
 - o **XPA::Annotator** – Functions to run the annotation pipeline for gathering data about molecules in the XPA database.
- **XPA::Annotator::???** – Helper modules and functions for use in the annotation pipeline.
- **XPA::DB::???** – DBIx::Class model files representing each table in the XPA database schema. These models are used throughout the XPA code base for interacting with the schema.

It must be noted at this point that these Perl modules (with the exclusion of the DB modules) are not currently object orientated in their programming and design.

This was mainly due to the inexperience of the programmer in the creation of these modules, but if further work were carried out it would be extremely beneficial to change the design of these modules to be object orientated as this would encourage better coding and enable more code reuse.

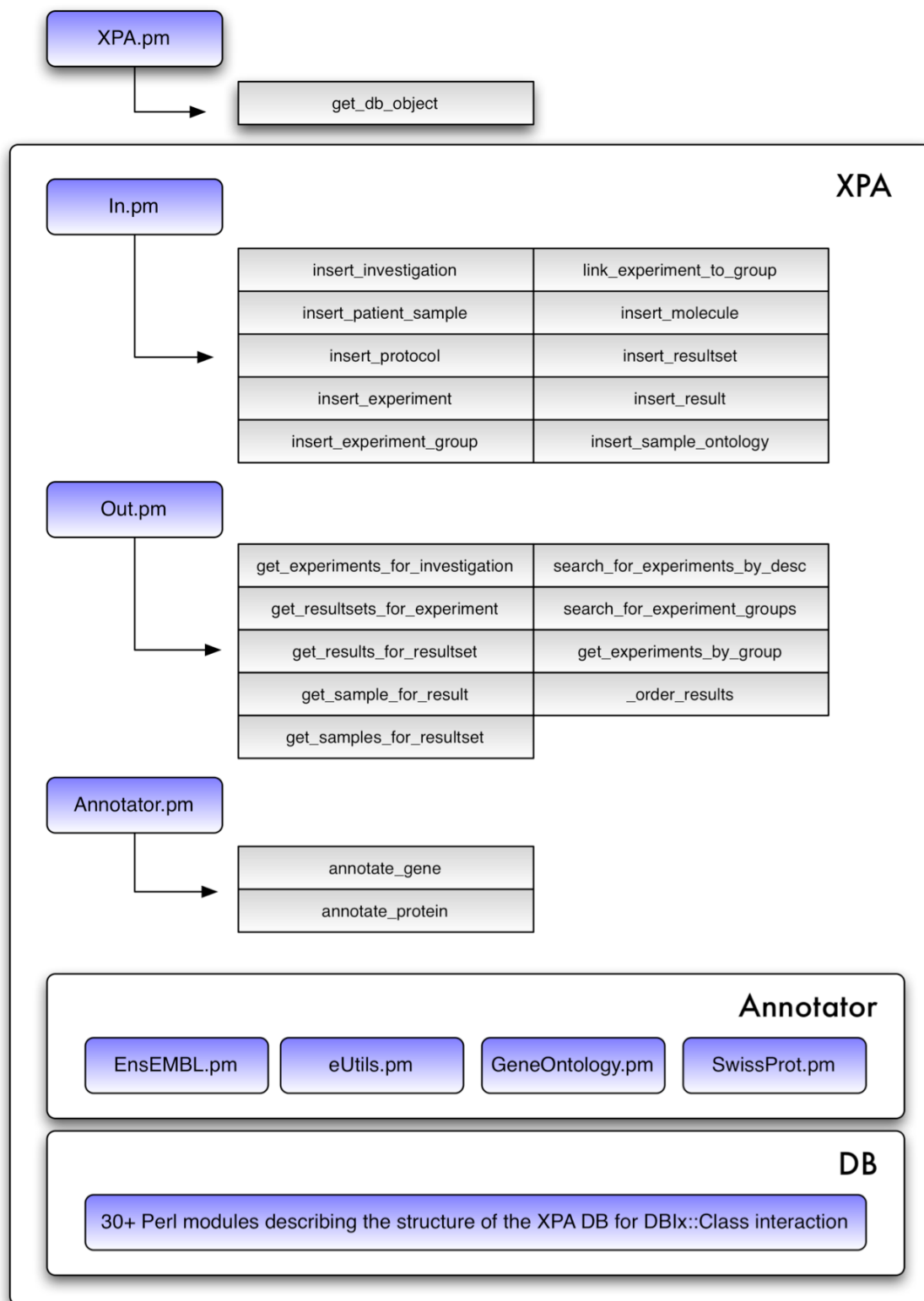


Figure 2-5: The structure of the XPA Perl module libraries. This diagram represents the layout and contents of the four main Perl modules that make up the XPA system at present (XPA, In, Out, and Annotator). The grey boxes represent the methods/subroutines that are currently with the given modules (for interacting with the XPA database).

2.4 Data Annotation

Annotation is the process of noting extra information on a particular point of interest, adding detail to aid in the understanding of an object. In terms of this thesis, and in the fields of bioinformatics and systems biology, annotation generally refers to the annotation of a given gene, protein or other molecule of interest – finding out as much information as possible about the given subject, from a variety of data sources. This can aid the understanding or interpretation of experimental results.

One of the early aims of this PhD study was to investigate annotation techniques, with the aim of automating these processes, so that complex and meaningful annotation of molecules of interest can be generated as quickly and as easily as possible, and automatically applied to previously acquired data. This work was first approached from the viewpoint of microarray probe annotation, but the work that was developed (described below) is a generic tool that can be used in a wider field than microarray probe annotation alone.

2.4.1 Annotator

Annotator is the name given to the annotation software that was developed in the first year of the PhD programme. Its intention was first to be an annotation tool for microarray probes (or the genes that the microarray probes are intended to represent), but also a general annotation package that would have a use in different omics fields, and was able to provide a reliable source of automated

annotation data from a large array of on-line data sources.

A flowchart giving specific details of the search process used by Annotator, and the information obtained during the course of a search for information about genes can be seen in Figure 2-6. The input format for these Annotator searches is a very simple comma-separated variables file, with two columns on each row, (one row per microarray spot), column 1 is an internal ID for a given microarray spot, column 2 is the GenBank Accession number for that spot. The annotation results retrieved are then output into a text file, in one of three (user-selected) formats – comma-separated variables, tab-delimited text, or XML. However, as shown later, the input and output processes of annotator were changed to take its input from, and import annotation data directly into the XPA schema.

The approach that has been adopted is beneficial in the fact that it makes direct use of central repositories such as GenBank, PubMed, Ensembl, Swiss-Prot and the Gene Ontology instead of relying on a third-party database typically made up of information extracted from these same sources. The latter approach can produce annotation that are outdated, whereas Annotator ensures that the annotations retrieved are always up-to-date.

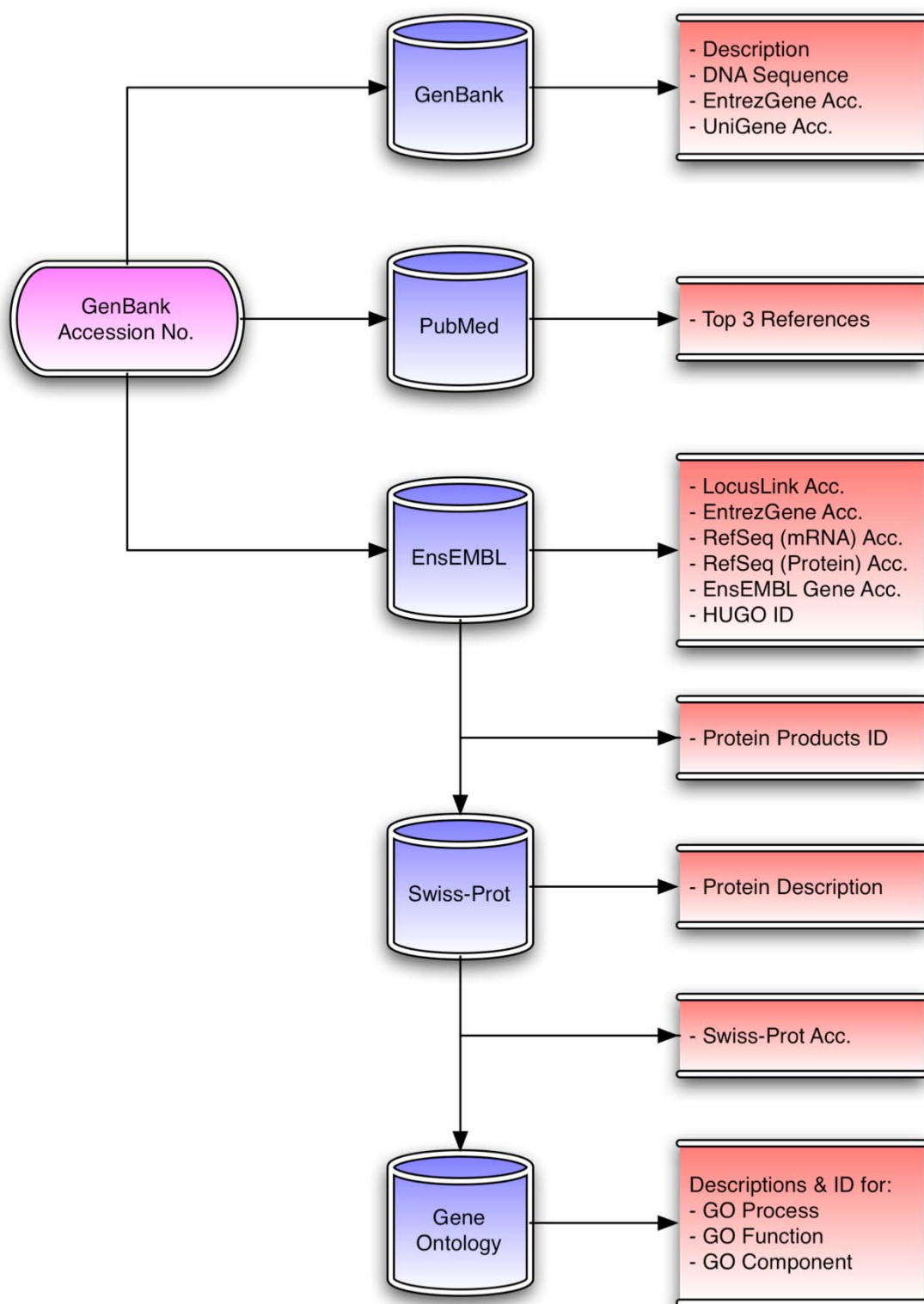


Figure 2-6: The Annotator process flow. This diagram represents the way in which Annotator visits various different databases and the information that is retrieved. This represents the steps taken to annotate genomic information within XPA.

Programming Specifics

Annotator was also written in the Perl programming language using common bioinformatics Perl tools including the BioPerl libraries (Stajich et al. 2002) and the Ensembl Perl API. These libraries/APIs were used as they provide a stable programming platform for accessing data in all of the major bioinformatics data repositories. However, the use of these APIs does not guarantee that compatibility will not be broken (and, by implication, the Annotator program will be broken) with each update of the on-line databases, but it does make keeping the program up-to-date and compatible with all databases used a far more simple process than direct access to the databases would allow. Generally, with the use of supplied APIs in programs, if a database schema is altered (as is fairly common with many bioinformatics databases), the only work needed to update a program to use the newer database build is to use a newer version of the supplied API - a fairly quick and easy process. If APIs were not used, and direct database (SQL) access was the preferred method – a change in a database schema could potentially involve major re-writes to applications, a difficult and often very time-consuming process.

2.4.2 Annotator within XPA

Data annotation with the XPA system was a highly desirable feature, as this would allow users to better investigate the datasets and experiments in question as it would be possible to have background information about all of the molecules studied easily accessible within the database, permitting more useful data mining and analysis. As such, the annotator program was updated and integrated with the

XPA system.

As the XPA system was envisaged to house more than just gene expression data, Annotator was modified to be a more flexible system than the original model and is now adaptable to retrieve annotation data for more than one molecule type (it was originally intended only to annotate genomic data). The current implementation of Annotator is capable of annotating both genes and proteins*, and it is anticipated that adding functionality for more molecule types would be a simple operation once suitable annotations have been sourced and an annotation plan formulated.

As before (with annotator), gene entries require a Genbank accession number to be associated with them so that Annotator has a starting point from where it can obtain the other annotation data. Proteins on the other hand require either a Uniprot or Entrez Protein accession number associated with them to allow annotator to start gaining related annotation data. The current method for interfacing with XPA and Annotator is through the use of command-line driven Perl programs, it is envisaged however that a GUI would be developed for the setting up and running of Annotator (and other XPA processes).

2.5 Data Analysis within XPA

As described previously, XPA is not just a data storage platform; it is intended to be a data storage and analysis platform. So far this chapter has been dedicated to

* Annotation data for proteins is collected from the online databases: NCBI Entrez (Entrez Protein and PubMed); EnSEMBL; Swiss Prot and Gene Ontology (via the QuickGO query engine provided by the European Bioinformatics Institute).

describing the data storage and handling aspects of XPA, this section briefly describes the data analysis possibilities integrated into XPA before moving onto the two case studies that are presented in this thesis.

In traditional biomarker analysis many different data analysis techniques have been used, and a great deal of them were discussed in the previous chapter. With XPA the intention is not to change the techniques used, but to streamline the way in which they are implemented. The basic premise is that once any dataset is properly loaded within the XPA database, the subsequent data analysis (no matter what the experiment type) would follow a mostly similar pattern, therefore it would be plausible to use the XPA system to automate the analysis of different datasets to some degree. In fact, once a data analysis method has been implemented in XPA, it can be applied to any of the types of data supported by XPA (either individually or in combination) as once the data is housed within the XPA schema, it is in a common format that would be suitable for the analysis routines.

This obviously is not of much merit when the analysis would only be calling for univariate analysis on a pre-determined biomarker - the overhead of loading data into the database and running analysis scripts from there would only add time and complications to the analysis. However, if the experiment is using newer technologies such as those introduced in Chapter 1 producing large amounts of data on many variables this is where it has been envisaged that XPA will be of great benefit, in both decreasing the time spent doing basic analysis of the data, and reducing the complexities of analysing different datasets, or even making it possible to treat data from different experimental techniques as a single dataset.

2.5.1 Exploratory Analysis

Consider the situation: “A researcher has a series of microarray gene expression profiles taken from patients with a form of cancer *versus* control subjects and you would like to see if there are any variables within the dataset that could be used to distinguish between the two populations (cancer vs. control).”

This aspect of data analysis can be achieved using data exploration techniques such as principal component analysis (PCA – as discussed in the previous chapter), and is something that is conceptually straightforward to automate within the XPA system by presenting the user with various different data pre-treatment options and controls, prior to automatically carrying out the analysis and identifying the variables that contribute the largest amount of variance within the dataset.

This approach would enable the users of the system to avoid learning the complexities of a statistical programming environment such as R or Matlab that are normally required to perform this type of analysis and get on with the task at hand – extracting the important information from the data. There would still need to be some human input in defining the best pre-treatment procedures (i.e. looking at the resulting PC scores plots to determine the quality of the data separation), but the ensuing model investigation (identifying the most significant variables) is entirely automatable and something that is very possible within the XPA framework.

2.5.2 Developing Classification Systems

Now consider this situation: “A researcher has already done their exploratory analysis, and has a set of biomarkers that they believe will be good for identifying or diagnosing a particular condition. They would like to quickly develop and test numerous different classification methods on their data to see which would be the most suitable to take forward with either more research or toward a product.”

This aspect of data analysis can be achieved using a variety of multivariate classification tools such as PLS-DA, ANNs and SVMs as described previously in Section 1.8 . Many of these tools (like PCA) can also be semi-automated after an appropriate configuration from an end user and thus would fit well into the XPA workflow. This aspect of the XPA system is covered in more depth as part of the case studies in the following chapters.

2.6 Case Studies

The following two chapters of this thesis present two case studies where the data from experimental investigations has been imported into the XPA system and analysed semi-automatically through the use of the XPA API discussed previously in Section 2.3 (for data import/export) and automated scripts for data investigation and classification using a combination of multivariate statistical techniques and machine learning, in order to define which would be the more appropriate classification tool for the data in hand.

3 Early Prediction of Prostate Cancer from qPCR Data

Prostate cancer is a disease in which cancer develops in the prostate, a gland in the male reproductive system. It occurs when cells of the prostate mutate and begin to multiply uncontrollably.

Prostate cancer is the most common cancer in men in the UK representing a quarter of all new cases of cancer diagnosed in men (Cancer Research UK), and is the sixth most common cancer in the world (in the number of new cases observed), the third most common cancer globally in men, and the most common cancer in men in Europe, North America, and some parts of Africa (Grönberg 2003).

Prostate cancer develops most frequently in men over fifty. Major known risk factors include age, family history, and country/ethnicity. Type-2 diabetes appears to reduce risk, while high birth weight and adult height are linked to increased risk of aggressive prostate cancer (Wigle et al. 2008). However, many men who develop prostate cancer never have symptoms, undergo no therapy, and eventually die of other causes.

3.1 Traditional Methods of Diagnosing Prostate Cancer

Prostate cancer is most often discovered by physical examination, such as a prostate biopsy or by screening blood tests, such as the PSA (prostate specific antigen) test. PSA is a 30-kDa glycoprotein produced primarily by the prostatic epithelial cells and secreted into the seminal plasma where it is present at relatively high concentrations ($\sim 0.5\text{--}3\text{ mg ml}^{-1}$) in individuals suffering from prostate cancer. Much lower concentrations ($\sim 10^6$ times lower) are found in the serum of normal males (Diamandis 1998).

PSA testing has come into widespread use as a cancer marker, both for initial diagnosis and for the monitoring of response to treatment. Use of PSA has also aided the prediction of prostate cancer risk and of treatment outcome. However, its routine use as a screening tool has been questioned because of a lack of specificity when levels are only moderately elevated (4 to 10 ng ml⁻¹). Twenty-five percent of men with PSA levels in this range do have biopsy-proven prostate cancer, but 75% have negative biopsy results (Catalona et al. 1991) indicating that the PSA test has only 25% sensitivity in such cases. Guidelines from Cancer Research UK however indicate that overall the PSA test has a sensitivity of 33% as it states “2 out of 3 men with a raised PSA do not have prostate cancer” on its guidelines for GPs talking about the PSA test (Cancer Research UK 2008) based on

the work by Selley et al. (1997).*

There is also continuing disagreement over the threshold level of PSA that should indicate biopsy as it has been shown that PSA levels vary naturally within individuals over given time periods as small as two-weeks (Nixon et al. 1997), and can even be effected by the current season (Salama et al. 2007).

PSA testing cannot be expected to resolve all ambiguity with respect to prostate cancer; instead, it may be best considered as an indicator of risk to be weighed in combination with other factors (Lilja et al. 2008). One such suggestion has been the examination of genes that are differentially expressed in prostate cancer (Eder et al. 2004) as a means of generating improved biomarkers and tests for diagnosis.

3.2 Developing Diagnostic Tools for Prostate Cancer with qPCR and XPA

The steps taken within this study to look at the condition of prostate cancer and develop improved methods for diagnosing the condition in men due to their genetic make-up are discussed below. All of the following clinical and laboratory work was carried out by collaborators at St George's Hospital Medical School, University of London (SGUL). The notes on the methodology described below are taken from the published article in the journal *Biomarkers* from which the data for this study was taken: "*Measurement of blood E2F3 mRNA in prostate cancer by quantitative qPCR: a preliminary study*" (Pipinikas et al. 2007).

* No published specificity figures for the PSA test were found at the time of this research.

3.2.1 Patient Recruitment and Groups

One hundred and forty-nine patients attending the uro-oncology clinic at St. George's Hospital (London, UK) were recruited on the basis of diagnosis (of prostate cancer) by prostate biopsies. Blood samples were obtained following fully informed consent, following the correct procedures (according to the World Medical Association), and with full ethics approval.

Patients were classified into two groups based on clinical diagnosis and histopathological information as well as radiological information - magnetic resonance imaging (MRI) scans. The 'benign' group included the individuals whose biopsy results were negative for cancer and patients diagnosed with benign prostatic hyperplasia (BPH). The 'malignant' group included the patients whose biopsy results indicated that they had cancer. This meant that the benign group contained 39 individuals, whereas the malignant group contained the remaining 110 subjects.

3.2.2 Laboratory Methodology

Whole blood samples of 10ml were collected from patients, and total RNA was extracted within 30 min of blood sampling. Gene-specific primers for the genes of interest were designed ready for PCR analysis. The genes of interest were as follows:

- E2F3 (accession number NM_001949)

- RECK (accession number NM_021111)
- MMP9 (accession number NM_004994)
- E2F4 (accession number NM_001950)
- HIF1A (accession number NM_001530)
- EF1a (accession number NM_001402)
- PIM1 (accession number NM_002648)
- MTSP-1 (accession number NM_021978)

In addition to the genes of interest, primers were designed for the house keeping gene GAPDH (accession number NM_002046) – this gene was used in the qPCR calculations to produce the relative gene expression values.

Following cDNA synthesis, qPCR (quantitative polymerase chain reaction) was carried out for each gene of interest and the relative gene expression levels calculated. From the acquired data for individual genes, expression ratios of several genes were calculated as these have been shown to improve performance of any further analysis, in terms of diagnosis (Fenske 2007), these were as follows:

- MMP9 / RECK
- HIF1A / RECK
- E2F3 / RECK

- E2F3 / E2F4

This data was then supplied to Cranfield University in the form of an Excel spreadsheet including the patient classifications and the serum PSA (prostate specific antigen) levels recorded at the time of blood sampling. This meant that the initial raw data matrix of samples (patients) versus variables (genes) was 149 samples x 8 genes, which was then increased up to 149 samples x 12 variables when the gene ratios were calculated, the data matrix was further increased to 149 samples x 13 variables when the PSA measurements were included.

3.2.3 Data Handling and Analysis

The processed gene expression and PSA data was imported into the XPA database using Perl programs taking advantage of the XPA API described previously in Chapter 2. The ‘molecules’ in this case were the genes of interest (and PSA), and the ‘results’ were the gene expression measurements (and PSA levels). This allowed the data to be moved into a standard data format so that all subsequent steps taken from XPA could be covered by any general analysis routine that had been defined.

The processed experimental data was then exported from the XPA schema into file formats suitable for further data analysis, so no further data manipulation was required prior to processing and analysis. Automated analysis and classification of the samples and datasets was then carried out using PLS-DA and SVM classification models as described below.

3.2.4 Classification Models using qPCR Data in XPA

The aim of this study was to determine if it was possible to accurately predict the incidence of prostate cancer from a combination of gene expression values, gene expression ratios and PSA measurements. This combination of variables was chosen as although sometimes controversial, PSA testing has been proven to be a successful predictor for prostate cancer – the aim of combining its measurements with those of the genes selected is to try to improve the sensitivity of the PSA test rather than replace it, and it is also a good (albeit rather small) example of how XPA can be used to combine both gene and protein data for building classification models. Two approaches were chosen for this analysis task: statistical methods and machine learning tools.

Statistical Methods

Several statistical methods suitable for classification problems were described previously in Section 1.8.1 – the methods that were used to analyse this dataset were PLS-DA and PCA. PLS-DA was used as the classification tool, and PCA was used for exploratory data analysis as a means to see if there was likely to be a simple way to differentiate between the two disease states according to the qPCR and PSA data.

Both sets of analysis were carried out using Matlab®, version 2006b (The MathWorks, Natick, MA, USA), with the PLS Toolbox, version 3.5 (Eigenvector Research, Inc. Wenatchee, WA, USA), on a Linux (Ubuntu Linux 7.04) workstation

using the exported (pre-prepared for each analysis technique) data files described in Section 3.2.3

For PCA analysis the following steps were carried out:

- All data values were logged (using $\log_{10}$) as this helps to emphasise the variation and differences between data points and should therefore aid the task of classification.
- The PCA analysis was carried out using the 'pca' function, which is part of the PLS Toolbox, on the logged qPCR expression data and PSA values, retaining values for the first 5 principal components. The results of this analysis are shown later in the chapter. This analysis was repeated twice more using the following pre-treatment methods to attempt to reveal further information about the data:

- o **Normalisation and Mean Centring:** for a 1 by N vector x , the norm

n_x is given by $n_x = \left(\sum_{j=1}^N |x_j|^p \right)^{1/p}$. The normalised 1 by N vector n_x is

given by x/n_x , this procedure was carried out using the Matlab

function 'normaliz'. This data was then mean centred given

$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$ using the Matlab function 'mncn'.

- o **Range Scaling:** The data matrix was range scaled using the function

$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{\max}} - x_{i_{\min}})}$ this was performed using the Matlab function

‘mapminmax’.

The PLS-DA model was prepared using the following steps:

- The dataset was split into two sets. 70% (104 samples, made up from 77 malignant and 27 benign samples) for cross-validation, optimisation and training of the model (described below), and 30% (45 samples, consisting of 33 malignant and 12 benign samples) of the dataset was held back completely for testing the final optimised model. The two datasets retained the balance of malignant versus benign sample groups to ensure that there was no training bias.
- Both the training dataset, and the test dataset were normalised and mean centred as with the PCA models described above.
- The PLS-DA models were created and trained using the (70%) training dataset. Five-fold cross-validation was used during the training to select the optimal number of latent variables (LVs) to use in the final PLS-DA model; n -fold cross validation was described previously in Section 1.8.1 in this case the data was split into five groups ($n=5$).
- Once the training and optimising of the models was completed, the models were then tested using the testing datasets (the remaining 30% of samples). The results of which are shown later in the chapter.

Machine-Learning

The machine-learning method that was chosen for use on this data set was Support Vector Machines (SVMs) due to their proven use and good performance in classifying biological data (Furey et al. 2000; Yeang et al. 2001; Pirooznia et al. 2008).

This analysis was carried out using the *libsvm* implementation of SVMs, a set of executable programs that are used to build, train and test SVMs (Chang & Lin). Version 2.84 of *libsvm* was used on the same Linux (Ubuntu Linux 7.04) workstation using the exported data files described in Section 3.2.3 . All training and optimisation of the SVM models was done through the use of the 'easy.py' python script that is supplied with *libsvm*, the steps carried out during the running of the 'easy.py' are described below.

The SVM model was constructed using the following steps:

- As with PLS-DA, each dataset was split into two sets: 70% for training (104 samples, made up from 77 malignant and 27 benign samples), and 30% (45 samples, consisting of 33 malignant and 12 benign samples) for testing the final model, with both sets retaining the balance of malignant versus benign sample groups.
- These datasets were then analysed and SVM models were trained, optimised and tested using the 'easy.py' script using the following steps:
 - o Data was range-scaled between -1 and +1 using the function

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{\max}} - x_{i_{\min}})}. \text{ This scaling method was used as this was highly}$$

recommended by the *libsvm* documentation, (Chang & Lin).

- o A grid search was used to optimise the values of C and γ for the SVM kernel (an RBF kernel was used). Five-fold cross-validation was used to train and test the SVM models during this search and was repeated upon each parameter change for C and γ until suitable values were selected. More details on the grid search procedure are described shortly.
- o Training was performed once more to generate the final SVM model (again using five-fold cross-validation, and the optimised values of C and γ).
- o The optimal SVM model was then used on the test (30%) dataset to assess the quality of the model, the results of which are shown later in this chapter.

Performance Metrics

When working with classification models, there needs to be a defined way in which the models are built and tested in order for their results to be reliably compared and contrasted. As such this short section is dedicated to describing the steps that were taken to build, test and compare the classification models used in this case study. For consistency, the metrics were essentially independent of classification

method (PLS-DA or SVM).

Optimising, Training and Testing

- As described in the previous section, the models (both PLS-DA and SVMs) were trained using 70% of the samples available in this study (the training dataset) – 104 samples, made up from 77 malignant and 27 benign samples.
- The parameters that needed optimising (LVs in PLS-DA and C and γ in SVMs), were optimised using five-fold cross validation whilst testing various different values for the parameters until an optimal model was found.
- The optimal classification models were tested on the remaining 30% of the samples available in this study (the test dataset) – 45 samples, consisting of 33 malignant and 12 benign samples.

Performance Metrics Recorded/Calculated

- *Accuracy*

This measure was calculated based on how many classifications the models made correctly.

- *Sensitivity*

This measure was one of two calculations used as a way of assessing the quality of the predictions made by the classification models. The sensitivity measures the proportion of actual positives that are correctly identified (i.e. the percentage of samples in the 'malignant' group correctly classified as

such). The sensitivity of a model was calculated as follows:

$$\text{Sensitivity} = \frac{\text{number of True Positives}}{\text{number of True Positives} + \text{number of False Negatives}}$$

- *Specificity*

This was the second calculation used to assess the quality of the predictions made by the classification models. The specificity measures the proportion of negatives that are correctly identified (i.e. the percentage of samples in the 'benign' group correctly identified as such). The specificity of a model was calculated as follows:

$$\text{Specificity} = \frac{\text{number of True Negatives}}{\text{number of True Negatives} + \text{number of False Positives}}$$

3.3 Results

The results achieved when using classification models on the qPCR dataset are presented below.

3.3.1 Data Investigation with PCA

Prior to building any statistical or machine learning classification models for this dataset preliminary analysis was carried out with PCA to see if there was likely to be a simple way to differentiate between the two disease states according to the qPCR and PSA data, as this would give more confidence in any subsequent analysis.

PCA was performed on logged values of the qPCR and PSA data as described

previously in the chapter*, and the first three principal components (PCs) were plotted against each other, these results can be seen in Figure 3-1 (raw data), Figure 3-2 (normalised and mean centred data) and Figure 3-3 (range scaled data). No significant separation between the two classification groups was observed in any of the three plots. Some sample grouping was observed, but the origin or reason for this grouping of samples could not be determined from the metadata provided and appeared to be unrelated to prostate cancer diagnosis.

* PSA data was not analysed in isolation, as PCA requires more than one sample variable to be of use.

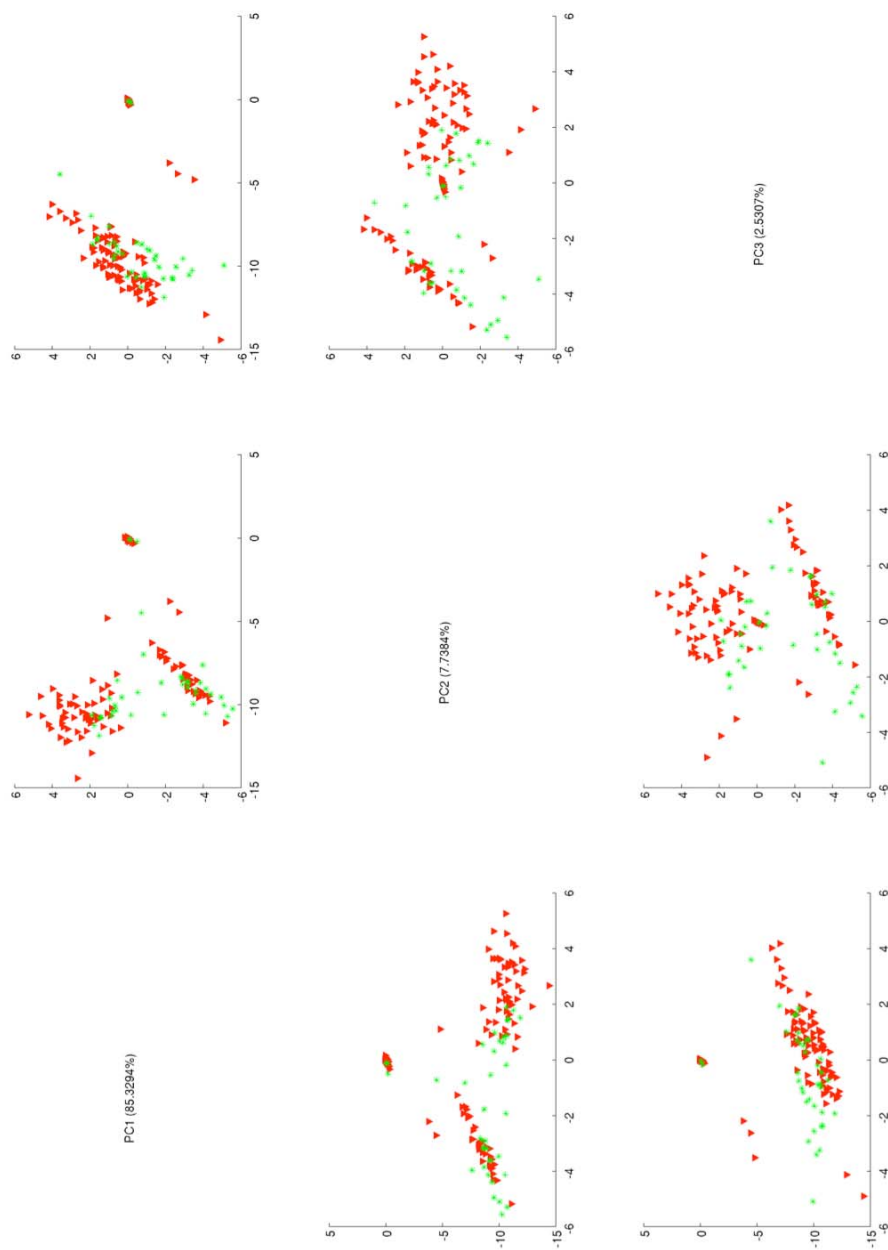


Figure 3-1: Results from PCA performed on the raw qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.

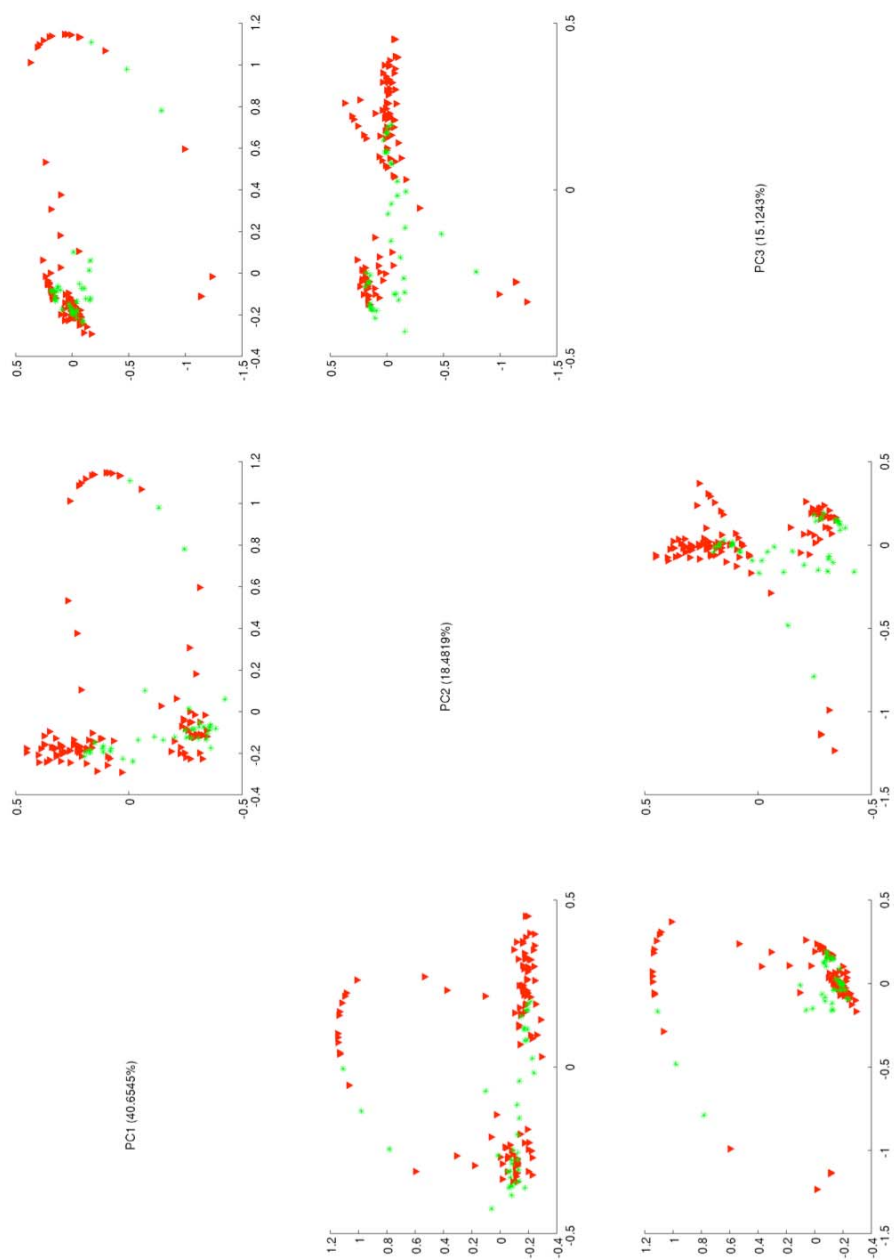


Figure 3-2: Results from PCA performed on the normalised and mean centred qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.

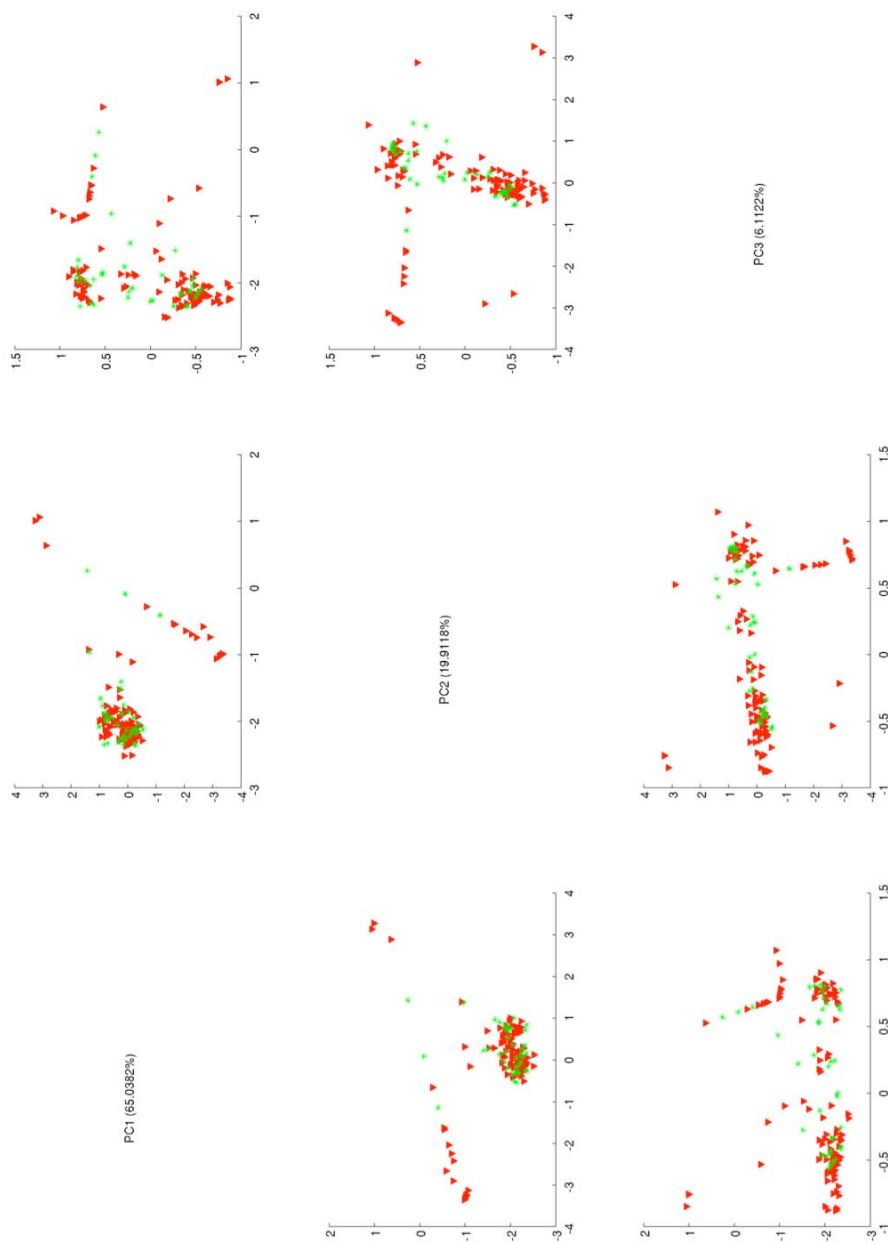


Figure 3-3: Results from PCA performed on the range scaled qPCR data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the group of patients in the malignant group, and green indicated samples in the benign group.

3.3.2 Classification – PSA Alone

The training dataset was used to train and produce a PLS-DA model using five-fold cross validation with a single latent variable (as only one variable was used – $\log_{10}$ PSA). The results presented in Table 3-1 are the classification results achieved using the test dataset with the trained PLS-DA model.

A radial bias function (RBF) kernel was chosen for use in the SVM model, as these are a very good kernel for classification models capable of working with both linearly and non-linearly separable data, (Balding et al. 2007). As such, an extra stage of model optimisation was needed in order to select the appropriate values of C and γ for use with the RBF kernel and the particular type of data that was used. As more than one parameter needed to be optimised (both C and γ), a grid search approach was used to find the best values to use for these parameters.

The grid search that was carried out by the *libsvm* 'easy.py' script consisted of first starting with a very coarse set of exponentially growing sequences; for example $C = 2^{-5}, 2^{-3}, \dots, 2^{15}$ and $\gamma = 2^{-15}, 2^{-13}, \dots, 2^3$. As with the PLS-DA models, cross-validation was used to test and train each SVM with the different pairs of C and γ . Once the initial, very coarse, search was completed, secondary searches were carried out using values for C and γ from a more focused range of values, based on the better models from the first step. The final values of C and γ selected for use in the SVM model were 2048.0 for C and 8.0 for γ as these helped to produce the best classification model. Results from the grid searches are represented in Figure 3-4.

Table 3-1: Results of the PLS-DA model trained and tested on the PSA data only.

Accuracy	57.1%
True Positives	27 out of 33 (81.8%)
False Positives	6 out of 33 (18.2%)
True Negatives	1 out of 16 (6.3%)
False Negatives	15 out of 16 (93.8%)
Sensitivity	64.3%
Specificity	14.3%

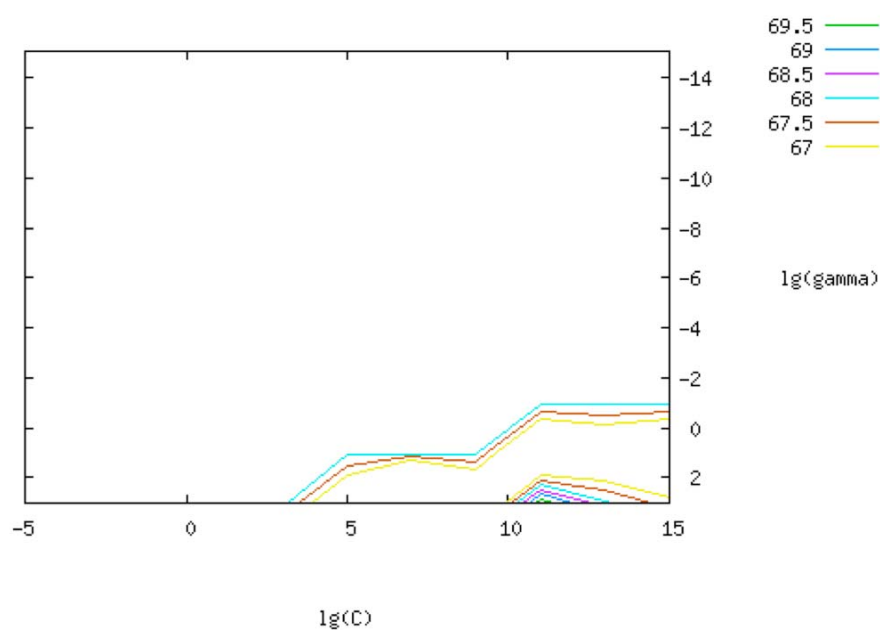


Figure 3-4: A plot representing the data acquired during the grid search used for training and optimising of the SVM model using only the PSA measurements. The different coloured bars represent the accuracy of the SVM models produced with the values of C and γ represented by the axes.

Table 3-2: Results of the SVM model trained and tested on PSA data only.

Accuracy	69.4%
True Positives	29 out of 33 (87.9%)
False Positives	4 out of 33 (12.1%)
True Negatives	5 out of 16 (31.3%)
False Negatives	11 out of 16 (68.8%)
Sensitivity	72.5%
Specificity	55.6%

3.3.3 Classification – PSA with Gene Expression Data

Again, the training dataset was used to train and produce PLS-DA models using five-fold cross validation, however this time they were produced using one to fifteen latent variables, the results of the cross validation models are shown in Figure 3-5. It was decided to use four latent variables for the final PLS-DA model as this number of LVs gave the lowest amount of error with the training dataset.

The results presented in Table 3-3 are the classification results achieved using the test dataset with the trained and optimised PLS-DA model using four latent variables.

As with the PSA only data, an RBF kernel was used in the SVM models, and as such, the optimisation of C and γ was carried out using a grid search as described previously. The final values of C and γ selected for use in the SVM model were 8.0 for C and 2.0 for γ as these helped to produce the best classification model. Results from the grid searches are shown in Figure 3-6.

The results presented in Table 3-4 are the classification results achieved using the test dataset with the trained SVM model.

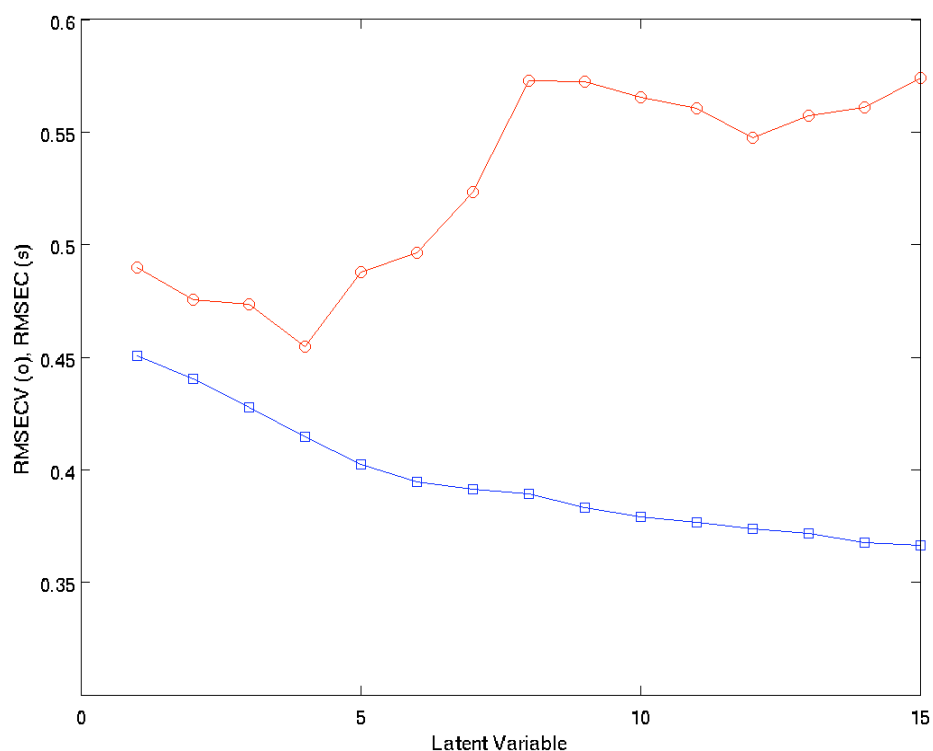


Figure 3-5: A plot representing the data generated from the five-fold cross validation of the PLS-DA models using different numbers of LVs in order to optimise the final model. The blue line represents the root mean square error of calibration – the error when the classifier was tested against the training datasets, whereas the red line represents the root mean square error of cross validation.

Table 3-3: Results of the classifications made by the PLS-DA model on unseen test data containing both PSA measurements and gene expression values/ratios.

Accuracy	73.5%
True Positives	25 out of 33 (75.8%)
False Positives	8 out of 33 (24.2%)
True Negatives	11 out of 16 (68.8%)
False Negatives	5 out of 16 (31.3%)
Sensitivity	83.3%
Specificity	57.9%

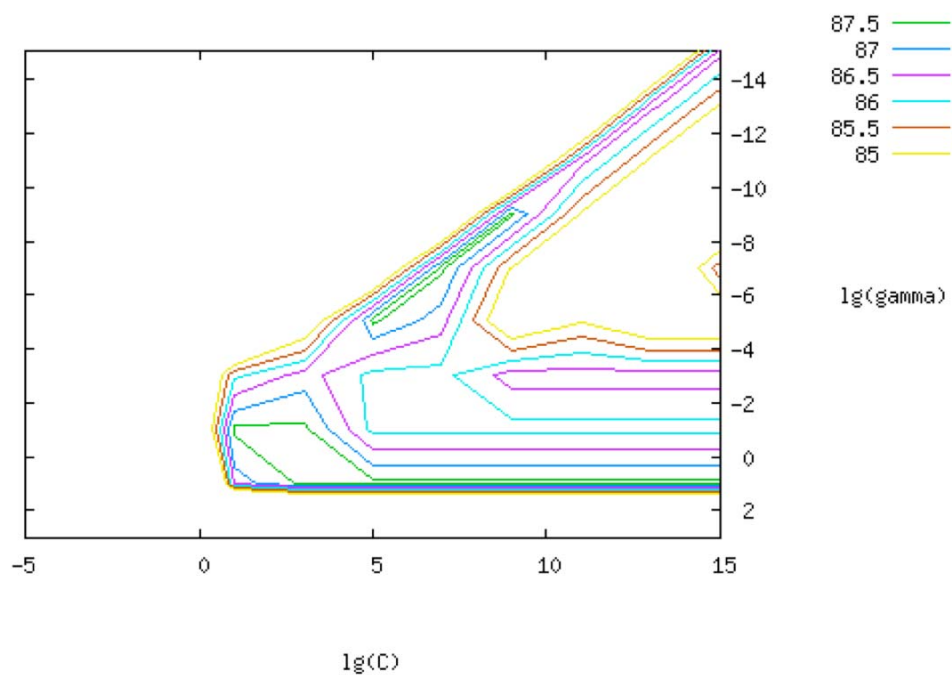


Figure 3-6: A plot representing the data acquired during the training and optimising of the SVM model. The different coloured bars represent the accuracy of the SVM models produced with the values of C and γ represented by the axes.

Table 3-4: Results of the classifications made by the SVM model on unseen test data containing both PSA measurements and gene expression values/ratios.

Accuracy	83.7%
True Positives	30 out of 33 (90.9%)
False Positives	3 out of 33 (9.1%)
True Negatives	11 out of 16 (68.8%)
False Negatives	5 out of 16 (31.3%)
Sensitivity	85.7%
Specificity	78.6%

3.4 Discussion and Conclusions

Having applied two distinct classification approaches (statistical: PLS-DA and machine learning: SVMs) using both PSA measurements alone and PSA in combination with gene expression values/ratios, it is now possible to compare results achieved with each other and with the published sensitivity figures for the current PSA test.

3.4.1 Expected PSA Test Sensitivity vs. SVM and PLS-DA models

The expected sensitivity for the PSA test is approximately 33% (Selley et al. 1997), but the sensitivity results obtained from this dataset using PSA data alone were 72.5 and 64.3% for the SVM and PLS-DA models respectively – far higher than the current methodology; when gene expression data was added to the PSA data, the sensitivity of the models generated increased to 85.7 and 83.3% for SVMs and PLS-DA respectively, this data is shown in Figure 3-7.

These results show that the addition of the gene expression data to the already established PSA test greatly enhances the predictive power of the test, thus presenting the potential for a far more accurate test for prostate cancer to be carried forward and developed. This approach of combining an established biomarker test with data points from modern analytical techniques is novel for the study of prostate cancer, but the theory however is not new. Another such study by Kong et al. (2006), looking at ovarian cancer diagnosis, discussed previously in Section 1.6 demonstrated that the combination of intensity values for 3 peaks from

SELDI MS from blood plasma samples along with measured CA125 levels enabled the production of prediction models significantly more accurate than the predictive power of CA125 alone.

In addition to this, the results achieved indicate that the mathematical methodology used by the SVM and PLS-DA models were able to make a much better prediction of the patients outcome based on the data presented to them, even when using PSA levels alone, rather than using a single cut-off value as is part of the traditional PSA test (Cancer Research UK 2009) before referring the patient for further testing.

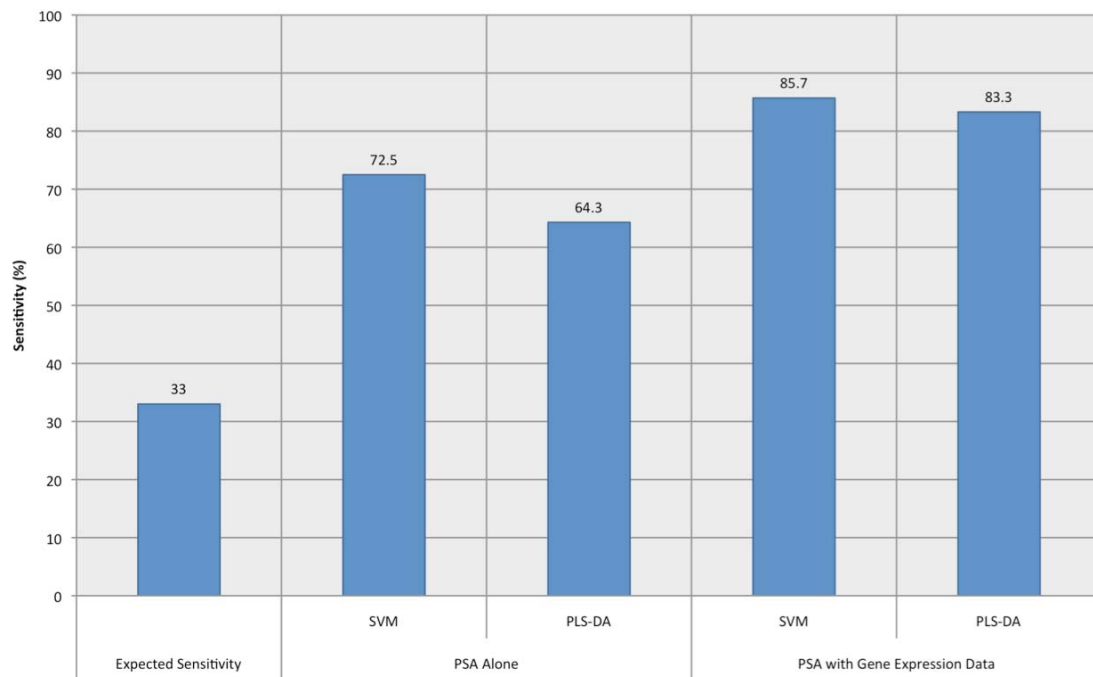


Figure 3-7: A comparison of the sensitivity results achieved using SVM and PLS-DA models against the published expected sensitivity figures for the PSA test when predicting the condition of prostate cancer. The results displayed in this figure indicate that the models produced using SVMs and PLS-DA are over twice as sensitive as the expected performance of the traditional PSA test indicating that they would be very suitable for use in a clinical situation (given further testing) as these models would correctly identify patients with prostate cancer more accurately than the expected performance of the traditional PSA test.

3.4.2 Comparison of the Two Classification Methods

One of the aims in the onset of this investigation was to compare the prediction results achieved by the two classification approaches used to ascertain if there was a clear advantage to using one over the other when applied to the PSA and gene expression data.

SVM and PLS-DA models have been compared before in the work by Sorich et al. (2003) where they were used to classify substrates and non-substrates for different isoforms of human UDP-glucuronosyltransferase (UGT), an enzyme family involved in the metabolism of drugs. In this study it was found that SVMs out performed PLS-DA in all areas. However no published work could be found comparing such predictive models with datasets or techniques related to biomarker studies to compare the following findings with.

In this study the SVM and PLS-DA models produced excellent classifications of the test datasets, it was however clear that the SVM models produced the better classifiers as they not only had a better overall accuracy, but a much higher sensitivity score, meaning that they would be far more likely to produce a positive result if an individual truly has prostate cancer – this is a very desirable trait in disease diagnosis. The results of these classifiers can be seen in Figure 3-8. Therefore, if one of the goals of this work were to recommend the best classification tool, an SVM based model would be the optimal choice.

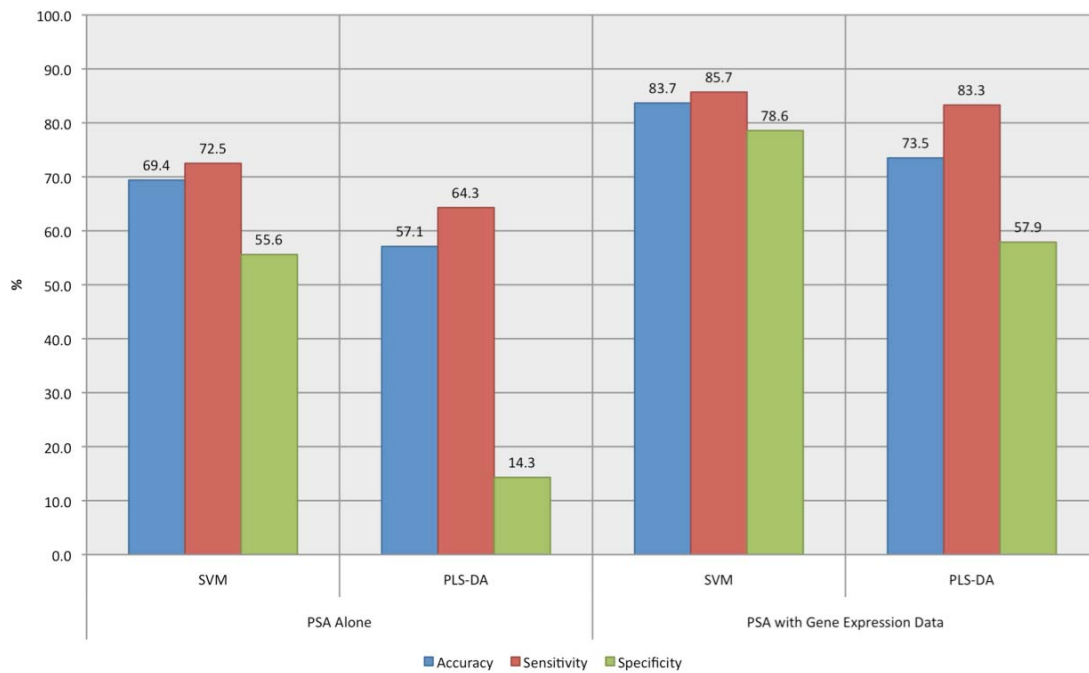


Figure 3-8: A comparison of the results achieved from the two different classification approaches used in this study when working with both PSA alone and PSA combined with gene expression data. The results represented in this figure clearly demonstrate that SVMs were the better classifiers on these datasets as they produced not only more accurate predictions of patient outcome, but also had a higher sensitivity and specificity giving greater confidence in the predictions made by the SVM models.

3.4.3 Using XPA to Study Prostate Cancer

What has been shown in this short case study is the use of XPA in the rapid development of classification models to assess whether the results gained from an experimental study could potentially be used as a diagnostic test.

In this instance it has been demonstrated that classification models based on both SVM and PLS-DA have potential as diagnostic classifiers, and that the results achieved thus far (especially with the SVM model) would give confidence in the classifications made. However, the dataset used here was quite small and may not be large enough to give confidence in a true clinical scenario, and therefore would warrant further study with more substantial amounts of clinical data (to better assess the classifications made by the SVM and PLS-DA models).

One potential function of XPA was not however used in this case study – that is the potential to use the system to fine tune or develop a more robust diagnostic technique. In this case study, that work had been done previously as the small set of genes tested were already selected due to their potential in diagnosing prostate cancer, therefore no further analysis was required, just the development of a potential classification system to be taken further towards a final diagnostic test. The more ambitious aspects of the XPA system (the ability to help identify diagnostic molecules, and deal with larger data sets) were tested in a further case study, which is presented in the next chapter.

4 Early Prediction of Pre-Eclampsia from SELDI MS Data

Pre-eclampsia is a disease of pregnancy characterized by hypertension (high blood pressure) and proteinuria (an excess of protein in the urine) developing after 20 weeks of gestation – it has been estimated that 5 to 7% of pregnancies worldwide are complicated by this condition (Walker 2000; Sibai et al. 2003; Mutter & Karumanchi 2008). Potential foetal complications include low birth weight, prematurity and death. Maternal complications include renal failure, HELLP syndrome (haemolysis, elevated liver enzymes, and thrombocytopenia), liver failure, cerebral oedema with seizures and rarely death (Mutter & Karumanchi 2008).

Traditionally, pre-eclampsia is diagnosed when the mothers' blood pressure becomes too high and an excessive amount of protein is recorded in the urine. However, this simple diagnostic approach is thwarted by the fact that many pregnant women develop mild high blood pressure but most do not have pre-eclampsia – this can often complicate or delay diagnosis. As such, a method for facilitating the early diagnosis of expectant mothers with pre-eclampsia is both highly desirable and needed.

Recent studies have indicated the influence of several proteins to the development of the condition which have the potential to act as biomarkers for pre-eclampsia. The earliest of these studies by Kunjara et al. (2000) found the protein inositol

phosphoglycans P-type (P-IPG), a putative insulin second messenger (in the past studied in relation to diabetes) was present in the placental tissue of women suffering from pre-eclampsia at two to three times higher levels than that of control subjects. This connection was later corroborated in a study by Williams et al. (2007) where it was found that concentrations of P-IPG in the urine were up to 30 times higher in pre-eclamptic patients compared to controls.

Another marker of interest more recently discussed is a high level of circulating soluble fms-like tyrosine kinase-1 (sFlt1, also referred to as sVEGFR-1) of placental origin (Ahmad & Ahmed 2004; Chaiworapongsa et al. 2004; Maynard et al. 2003; Zhou et al. 2002). sFlt1 is an endogenous protein produced by the placenta that is able to bind to the angiogenic growth factors vascular endothelial growth factor (VEGF) and placental growth factor (PlGF), thereby neutralizing their functions (Tjoe et al. 2007). High concentrations of circulating sFlt1 along with decreased free VEGF and free PlGF are seen not only during pre-eclampsia but before the onset of clinical symptoms; it has also been shown that over expression of sFlt1 in rats leads to hypertension, proteinuria and glomerular endotheliosis, the classical symptoms of pre-eclampsia (Venkatesha et al. 2006).

More recently, serum levels of soluble endoglin (sEng) – a protein that inhibits endothelial capillary tube formation and promotes vascular permeability – were also shown to be significantly elevated in pre-eclamptic women and levels of sEng correlated strongly with disease severity (Mutter & Karumanchi 2008). In tests performed in rats, over-expression of sEng caused proteinuria and hypertension, but this was far milder than seen with sFlt1 over-expression. However, when sEng

was co-expressed with sFlt1, the rats developed severe conditions similar to human HELLP syndrome (Venkatesha et al. 2006). These findings suggest that sEng and sFlt-1 both cause endothelial dysfunction by different mechanisms but may cooperate to produce the clinical syndrome of pre-eclampsia (Mutter & Karumanchi 2008). Therefore, measurement of sFlt1 and sEng in the maternal circulation could have the potential to be a useful diagnostic and screening tool for preeclampsia (Tjota et al. 2007).

4.1 Assisting Early Diagnosis of Pre-Eclampsia with SELDI MS and XPA

Discussed below are the steps taken within this study towards developing new methods for predicting the onset of pre-eclampsia in pregnant women. The steps that were carried out are as follows:

- Early antenatal blood samples were collected from a cross section of pregnant women (details on the laboratory preparation and processes can be found in Section 4.1.2), with 30 samples from sufferers of pre-eclampsia and 60 control subjects.
- These samples were analysed using SELDI-TOF mass spectrometry, employing a range of different SELDI chips, and saved in a proprietary Ciphergen XML data format.
- The SELDI data files were converted to raw spectra (CSV) files, normalised and analysed (to identify all observed peaks, and the most commonly

observed peaks across all spectra) using the R PROcess package.

- The resulting peaks and common peaks were then imported into the XPA database, (using Perl scripts taking advantage of the XPA API described in Chapter 2) ready for analysis.
- The processed experimental data was then exported from the XPA schema into file formats suitable for data analysis.
- Automated analysis and classification of the samples and datasets was then carried out using PLS-DA and SVMs.
- Finally, the PLS-DA models were investigated to see if it was apparent which protein peaks were more important in their contribution to determining the classification of the samples. Such analysis could reveal putative biomarkers.

An overview of these processes is shown in Figure 4-1.

4.1.1 Study Goals

As stated previously the development of a diagnostic test for pre-eclampsia is highly desirable and needed to identify expectant mothers who are likely to develop the condition. This study is trying to help with this goal as follows:

- Assess the different types of SELDI protein chips in use in this study and identify which would be the most suitable for use in a diagnostic test based on SELDI MS technologies. In tandem with this, compare different

classification tools and assess which would be the most successful for predicting the onset of pre-eclampsia from SELDI MS data.

- Using the SELDI MS data, identify the more important peaks in the dataset that, if the peptides represented by these peaks were identified, could lead to potential protein biomarkers for pre-eclampsia. The identification of such proteins could lead to a relatively simple test for the condition in the future.

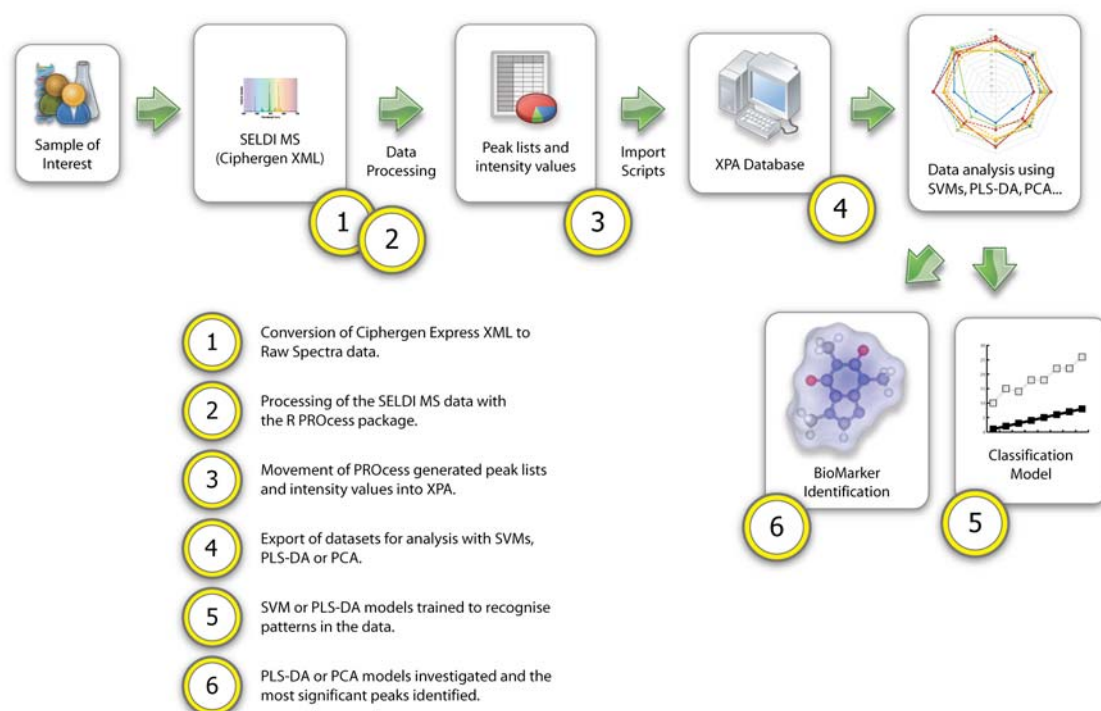


Figure 4-1: An overview of the processes used to look at the SELDI MS data from pre-eclampsia sufferers. This details each individual step that was carried out within this study in order to use the SELDI MS data with the XPA system.

4.1.2 Sample Preparation and Laboratory Methodology

Early antenatal blood samples obtained from a large selection of pregnant women were centrifuged, and the resulting serum was stored frozen until after the pregnancy was completed. Following completion of the pregnancies, thirty specimens from women who subsequently developed severe pre-eclampsia (cases), and sixty specimens from women who never developed pre-eclampsia (controls) were selected for inclusion in this study. Each serum specimen was fractionated six times at different pH values as described in Table 4-1.

Following fractionation, and before any further procedures were carried out, fractions 1 and 2, and fractions 5 and 6 were combined in order to reduce the expense of the following procedures.

The four groups of fractions (fractions 1&2, fraction 3, fraction 4, and fractions 5&6) were then tested on CM10, H50 and IMAC CIPHERGEN SELDI chips at two laser intensities (high and low). Each specimen was analysed in duplicate on the chips and the data used for all subsequent analysis was a mean average of results over the two runs.

All laboratory procedures were carried out by St. George's Medical School, University of London (SGUL) and were performed with the correct ethical approval prior to any involvement from anyone from Cranfield University.

Table 4-1: pH level / solvents used to fractionate the antenatal blood serum samples.

Fraction	pH level / solvent
1	9
2	7
3	5
4	4
5	3
6	organic solvent

4.1.3 Processing SELDI MS Data for Use with XPA

SELDI MS data was supplied in the form of a collection of Ciphergen Express XML files. These files contain a complete set of information about a given experiment, including the subject numbers and the configuration of the SELDI MS instrument – it also contains the raw data from a SELDI MS experiment. As useful as this data format is for containing the surrounding and resulting data from a SELDI experiment, unfortunately any non-commercial tools for subsequent analysis do not use this format – it is specifically designed for the Ciphergen Express software now sold by Bio-Rad Laboratories.

Therefore, the first step involved in looking at this data was to convert it into something more useful for the tools that were available – raw spectra in a comma-separated variables format. This was achieved using a Perl program to read in the XML files and then generate the raw spectra data files. Details of the algorithm used in this process are shown in Figure 4-2.

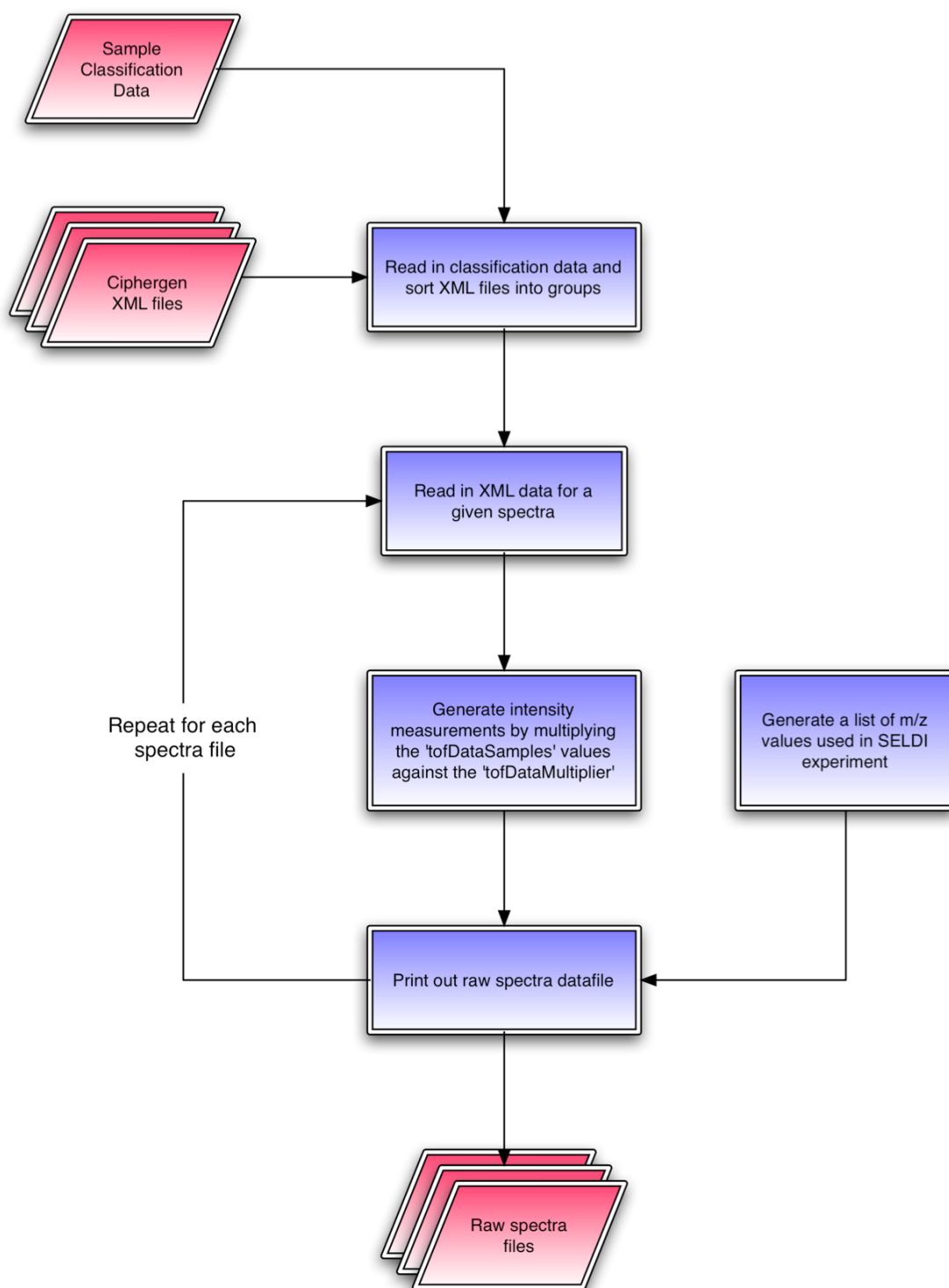


Figure 4-2: A flowchart representing the processes put in place to convert the CIPHERGEN XML data files into raw spectra files that could be used with the R PROcess package. All of these steps were implemented as part of a simple program written using the Perl programming language.

Before the SELDI MS data could be used within the XPA database, several steps of processing and analysis were performed on the raw spectra as the size of the raw spectra data (approximately 2Mb per spectra file, all spectra for this dataset combined were 4.1Gb in size), would not be suitable for housing within the XPA schema – the reasons for not wishing to store raw data within the XPA schema were outlined in Section 2.2.1 This processing was carried out using the R statistical programming environment (R Development Core Team) and the PROcess package for analysing SELDI MS data (Li), part of the Bioconductor library for R (Gentleman et al. 2004). The steps of processing and analysis that were carried out shall now be described.

Firstly, the raw spectra files were read into memory, an example of a raw spectrum can be observed in Figure 4-3, part (a). Noise from the each of the spectra was then removed by applying the process of baseline subtraction; a plot demonstrating this process on a single spectrum can be seen in Figure 4-3, part (b). Baseline subtraction was performed on all spectra using the R function 'rmBaseline', which is a batch function for 'bslnoff', which itself is a wrapper around the 'loess' R function for local polynomial regression fitting, the algorithms used by this function are described below (adapted from Gentleman et al. 2005):

1. For each spectrum find local minima by using moving windows of 0.3% around each m/z value, and take the minimum intensity value for each window.
2. Using polynomial regression, fit a continual line on the minimum values

obtained for the windows calculated above – this forms the new estimated baseline for the spectra.

3. Subtract the estimated baseline from each spectrum.

The next step in the processing was peak detection as the peaks in the spectra represent specific, abundant proteins in the sample. These, and also the differences in the intensities of these peaks between samples are ultimately the most interesting points in an MS dataset. Peaks were identified using the R function 'getPeaks', which is a batch function for 'isPeak'. The algorithm used by the function is outlined below (adapted from Gentleman et al. 2005), a plot demonstrating some of the results of this processing are shown in Figure 4-3, part (c):

1. The spectrum is smoothed using moving averages of the five nearest neighbours (in both directions), around each point in the spectrum in order to remove spurious peaks.
2. Local variability is computed as the median of the absolute deviations (MAD) of 40 of the nearest neighbours (in both directions), to each point in the spectra.
3. Once the above are done, peaks are identified (or *called*). Three thresholds are used in peak calling, they are as follows:
 - (a) The signal to noise ratio, which is calculated as the value the smoothed spectrum (as calculated in step 1) at a given point divided by the local

estimate of variation. This is used to select peak candidates – i.e. we retain peaks with signal to noise ratio greater than two.

- (b) The detection threshold for the whole spectrum, below which the intensities are considered as zero. This is used to keep out low peaks in relatively flat regions, where the local variation is nearly zero and the signal to noise ratio alone might lead to the selection of smaller peaks.
- (c) The shape ratio, which is computed as the area under the curve within a small distance (0.3%) of a peak candidate, as identified by the first two criteria, divided by the maximum of all such peak areas of a spectrum. This third threshold helps discard single spike peaks. Given the limited precision of the SELDI MS technology, a single spike is more likely to be a technical abnormality than a true signal.

Finally, common peaks across the spectra were identified using peak alignment – that is, identifying those peaks, across spectra, which are likely to represent the same protein/peptide. For this it must first be noted that the observed variation in peak location is not simply a constant shift but is proportional to the value of m/z . Currently, the accuracy in the m/z position is believed to be within 0.3% of the observed m/z value; as such, peak alignment was performed using the R function 'pk2bmk', using the methodology described below:

- For each peak, an interval of 0.3% was calculated around its m/z value to account for inaccuracy in the measurement of m/z . This meant that the interval representing each peak was $[0.997x, 1.003x]$.

- The peak intervals were compared across all spectra, and those peaks that are present in 10% or more of the acquired spectra are considered to be common peaks. The center m/z value across the group of aligned peaks was used to represent the common peak, and the maximum intensity value was used as the height of the common peak.

From here, the identified peaks, common peaks and their intensity values were saved into files, and then imported into the XPA database (via the use of Perl import scripts).

Storing SELDI MS Data in the XPA Database

In the previous study, the way in which qPCR data was stored within the XPA schema was very straight forward – the data generated was gene expression measurements, therefore the molecules that were stored were genes, and the data entered into the results classes were the associated gene expression measurements.

With SELDI MS data however, this was not quite as straightforward as although this technique measures the abundance of proteins – the proteins that are being measured are unknown unless it is known which m/z values refer to specific proteins. Therefore the approach that was given to loading SELDI MS data into the XPA database was to treat each m/z value as a molecule (analogous to a gene or protein), and record the intensity of the measurements in the results classes.

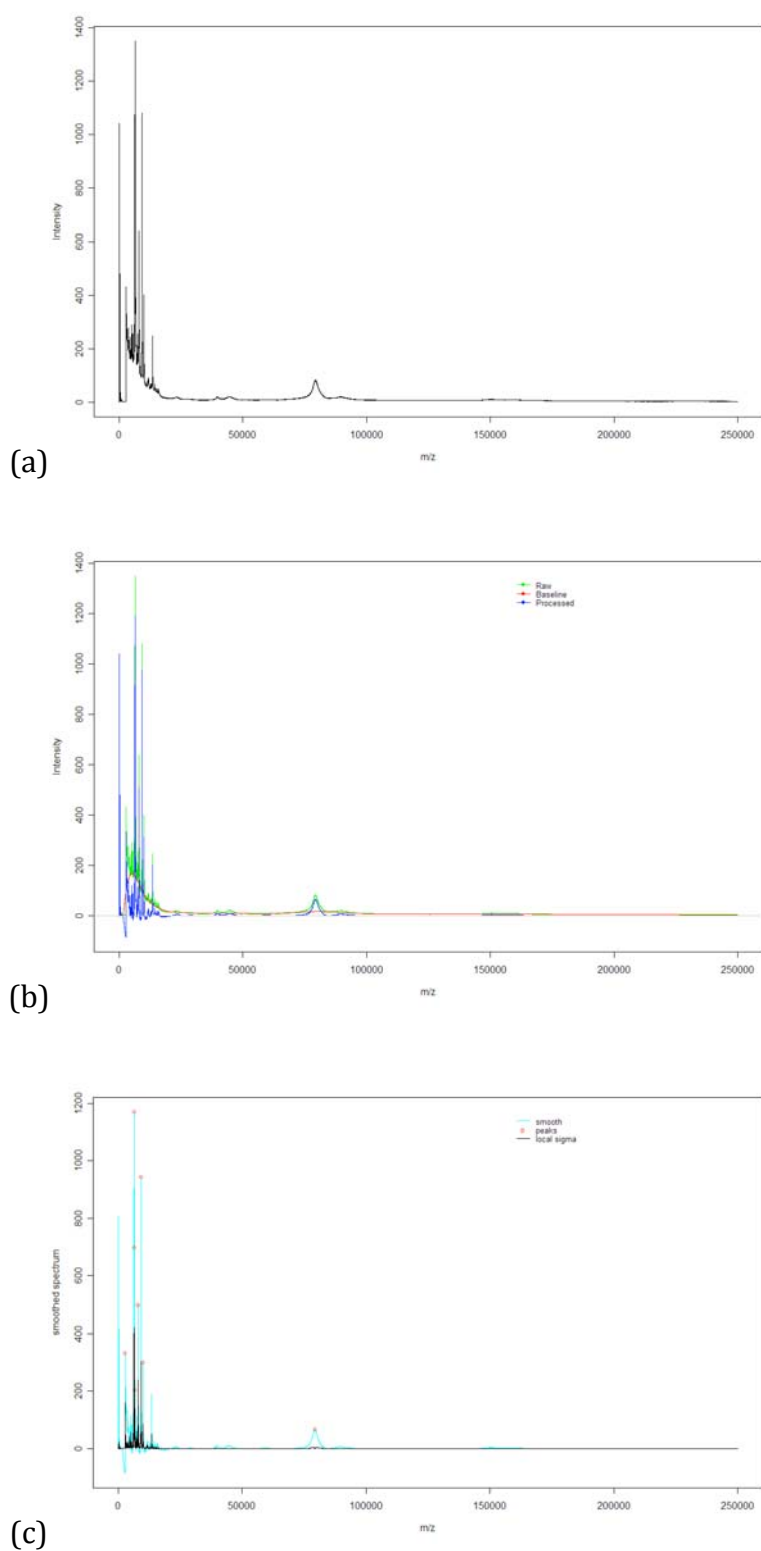


Figure 4-3: Spectrum plots demonstrating the various stages of processing with R. (a) A raw spectrum. (b) Spectrum baseline subtraction – the green line shows the raw spectrum; the red shows the estimated baseline of the raw spectrum; and the blue indicates the raw spectrum minus the estimated baseline. (c) A spectrum with the detected peaks marked by red circles.

4.1.4 Classification Models Using SELDI MS Data in XPA

The first type of analysis that was performed on this dataset was classification to see if it was possible to use the SELDI MS data as an accurate prediction technique for pre-eclampsia. As in the previous study, two approaches were chosen for this analysis task: statistical methods (PCA, PLS-DA) and a machine learning approach (SVMs), these were the same techniques as used in the previous chapter and demonstrates how XPA makes it simple to apply existing analysis techniques to other types of data.

The dataset was split into six groups for analysis and classification; these were based on the type of SELDI chip being used (CM10, H50 and IMAC) and the energy of the laser being used to activate the samples on the chip (high and low). This approach should allow the consideration of whether the use of a single SELDI chip was a better classifier over others. In addition, the common peaks and 'all' peak datasets produced by the R PROcess package were also analysed separately to see if this extra level of data processing (identifying the more common peaks) had any significant improvement on the classification power of the models.

Statistical Methods

Several statistical methods suitable for classification problems were described previously in Section 1.8.1. The methods that were used to analyse this dataset were that of PLS-DA and PCA. PLS-DA was used as the classification tool, and PCA was used for exploratory data analysis as a means to see if there was likely to be a

simple way to differentiate between the two disease states according to the acquired SELDI data.

Both sets of analysis were carried out using Matlab, version 2006b, (The MathWorks, Natick, MA. USA), with the PLS Toolbox, version 3.5 (Eigenvector Research, Inc. Wenatchee, WA, USA), on a Linux (Ubuntu Linux 7.04) workstation.

The PCA analysis was carried out using the 'pca' function (following the methodology defined earlier in Section 1.8.1 which is part of the PLS Toolbox, on the SELDI MS peak data intensity values, retaining values for the first five principal components. The results of this analysis are shown later in the chapter. The analysis was repeated twice more using two different pre-treatment methods to attempt to reveal further information about the data, the two methods used were normalisation and mean centring, and range scaling – both of these techniques were described previously in Section 3.2.4

The PLS-DA model was prepared using the following steps:

- The dataset was split into two sets. 70% for cross-validation, optimisation and training of the model (described below) consisting of 63 samples (21 cases, 42 controls), and 30% of the dataset (27 samples consisting of 9 cases, 18 controls), was held back completely for testing the final optimised model. The two datasets retained the balance of malignant versus benign sample groups to ensure that there was no training bias.
- Both the training dataset, and the test dataset were normalised and mean

centred as with the PCA models described above.

- The PLS-DA models were created and trained using the (70%) training datasets. Five-fold cross-validation was used during the training to select the optimal number of latent variables (LVs) to use in the final PLS-DA model; n -fold cross validation was described previously in Section 1.8.1 in this case the data was split into five groups ($n=5$).
- Once training and optimising of the models was completed, the models were then tested using the testing datasets (the remaining 30% of samples). The results of which are shown later in the chapter.

Machine Learning

As with the statistical classification methods, two different machine-learning techniques were discussed previously in Section 1.8.2. The method that was chosen for use on this data set was Support Vector Machines (SVMs) due to their proven use and good performance in classifying biological data (Furey et al. 2000; Yeang et al. 2001; Pirooznia et al. 2008).

This analysis was carried out using the *libsvm* implementation of SVMs, a set of executable programs that are used to build, train and test SVMs (Chang & Lin).

Version 2.84 of *libsvm* was used on a Linux (Ubuntu Linux 7.04) workstation for all analysis. All training and optimisation of the SVM models was done through the use of the 'easy.py' python script that is supplied with *libsvm*, the steps carried out during the running of 'easy.py' are described below.

The SVM model was constructed using the following steps:

- As with PLS-DA, each dataset was split into two sets: 70% for training (63 samples – 21 cases, 42 controls), and 30% for testing the final model (27 samples – 9 cases, 18 controls), with both sets retaining the balance of malignant versus benign sample groups.
- These datasets were then analysed and SVM models were trained, optimised and tested using the ‘easy.py’ script using the following steps:
 - o Data was range-scaled between -1 and +1 using the function
$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{\max}} - x_{i_{\min}})}$$
This scaling method was used as this was highly recommended by the *libsvm* documentation, (Chang & Lin).
 - o A grid search was used to optimise the values of C and γ for the SVM kernel (an RBF kernel was used). Five-fold cross-validation was used to train and test the SVM models during this search and was repeated upon each parameter change for C and γ until suitable values were selected. More details on the grid search procedure are described shortly.
 - o Training was performed once more to generate the final SVM model (again using five-fold cross-validation, and the optimised values of C and γ).
 - o The optimal SVM model was then used on the test (30%) dataset to

assess the quality of the model, the results of which are shown later in this chapter.

Performance Metrics

As discussed in the previous chapter, when working with classification models, there needs to be a defined way in which the models are built and tested in order for their results to be reliably compared and contrasted. The metrics used in this study were exactly the same as those used in the previous case study (described in detail in Section 3.2.4).

Optimising, Training and Testing

- The models (both PLS-DA and SVMs) were trained using 70% of the total samples (63 samples – 21 cases, 42 controls).
- Optimisation of the models was performed using five-fold cross validation.
- The optimal classification models were tested on the remaining 30% of the samples (27 samples – 9 cases, 18 controls).

Performance Metrics Recorded/Calculated

- Accuracy
- Sensitivity
- Specificity

4.1.5 Using XPA to Look for Biomarkers

The other type of analysis that was carried out on this dataset was that of biomarker detection. This was achieved by using the PCA and PLS-DA models from the previous analysis. Using these two approaches it is possible to extract the more significant data points found via each method, although this is dependant on the results of each analysis – if the data could not be separated/classified according to diagnosis using either technique, investigation of the model was not desirable, as it would not yield any useful markers of disease.

If either the PCA or PLS-DA models were deemed accurate in separating cases from controls, the models were investigated further to ascertain the more important data points (peak m/z values) observed in the SELDI MS spectra.

4.2 Results

The following section of the report outlines the results achieved when using classification models on the SELDI MS datasets. Described first are the results achieved using statistical methods, then machine learning, and then finally the comparison of the results achieved by the two different approaches.

4.2.1 Data Investigation with PCA

Prior to building any statistical or machine learning classification models for the datasets, preliminary analysis was carried out with PCA to see if the data is clearly/easily separable, as this would give more confidence in the results

anticipated before proceeding into the next stages of analysis.

PCA was performed on all samples, and the first three principal components (PCs) were plotted against each other, these results can be seen in Figure 4-4. As can be observed in the figure, there was no clear separation between the two sample groups (pre-eclampsia sufferers and control); this required further investigation.

The data from the first three PCs was re-plotted in Figure 4-5, but this time the data points were coloured based upon the type of SELDI chip that was used for each analysis, CM10 as red dots, H50 as green and IMAC as blue. This series of plots demonstrated a clear clustering of the data points based on the type of SELDI chip being used, not the disease state of the sample.

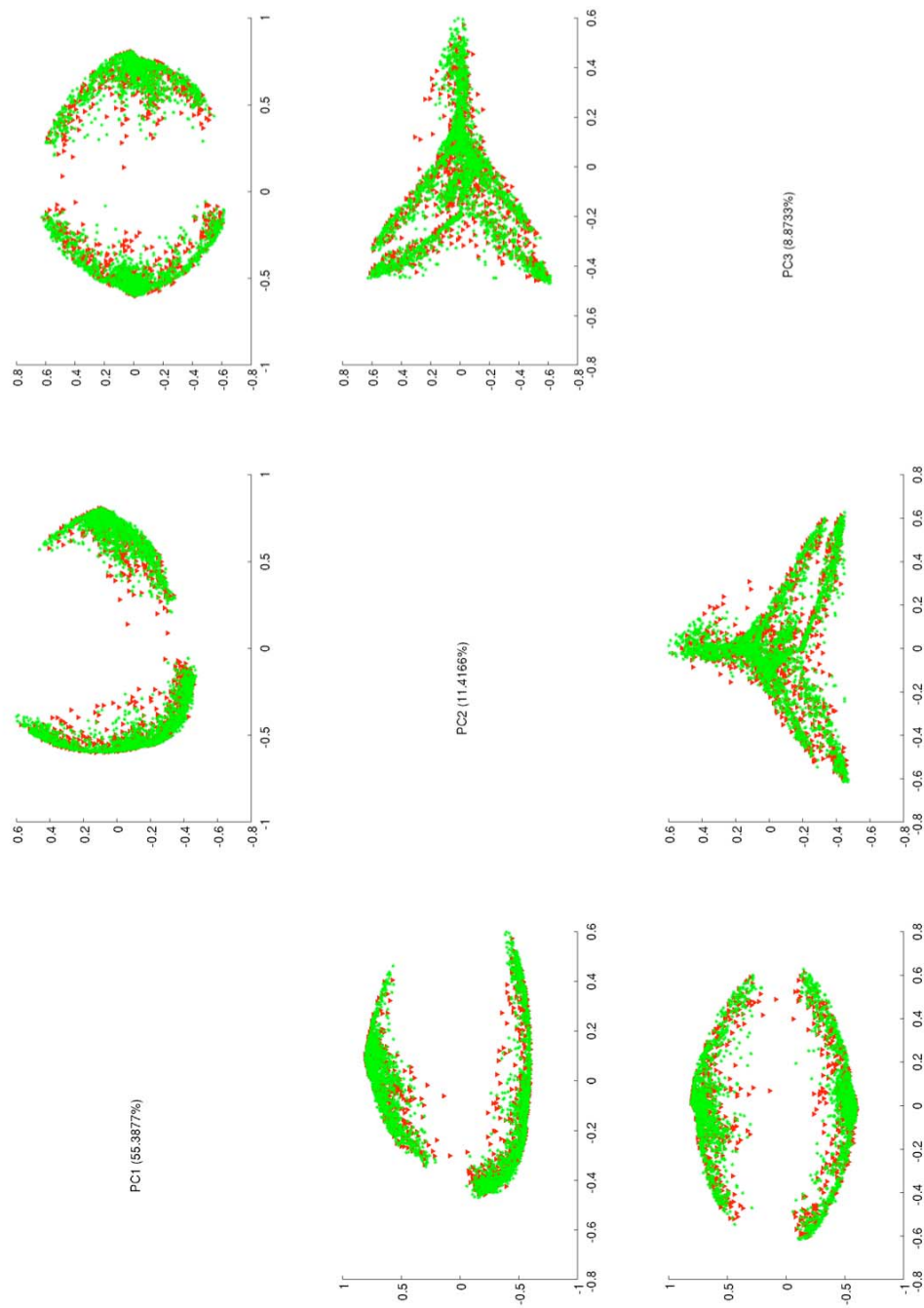


Figure 4-4: Results from PCA performed on all of the SELDI data. These plots represent the first three principal components (PCs), being compared against each other. Markers in red indicate samples in the PE group, and green represents the control group. No clear disease-related separations could be seen in this data to lead to any more findings.

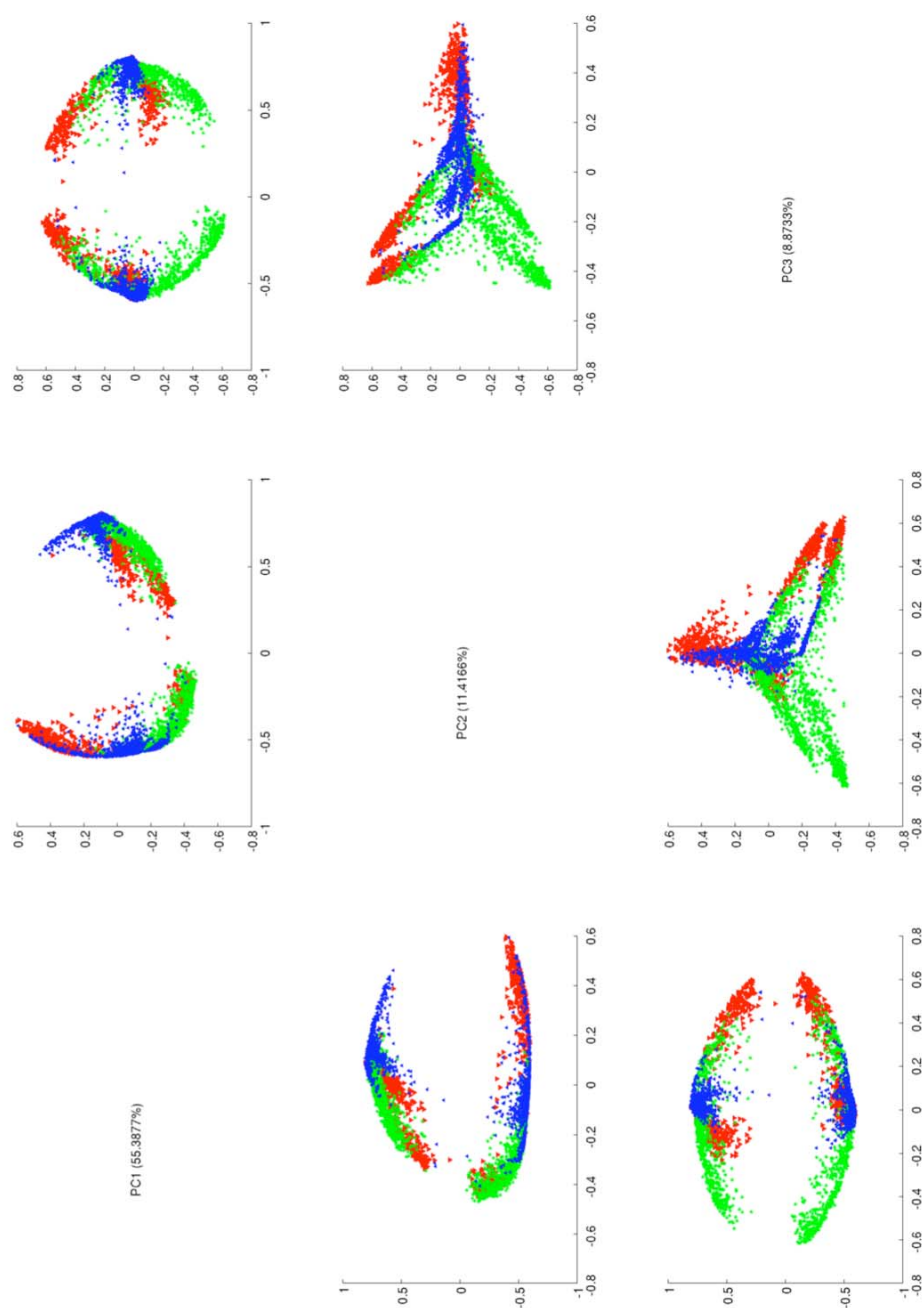
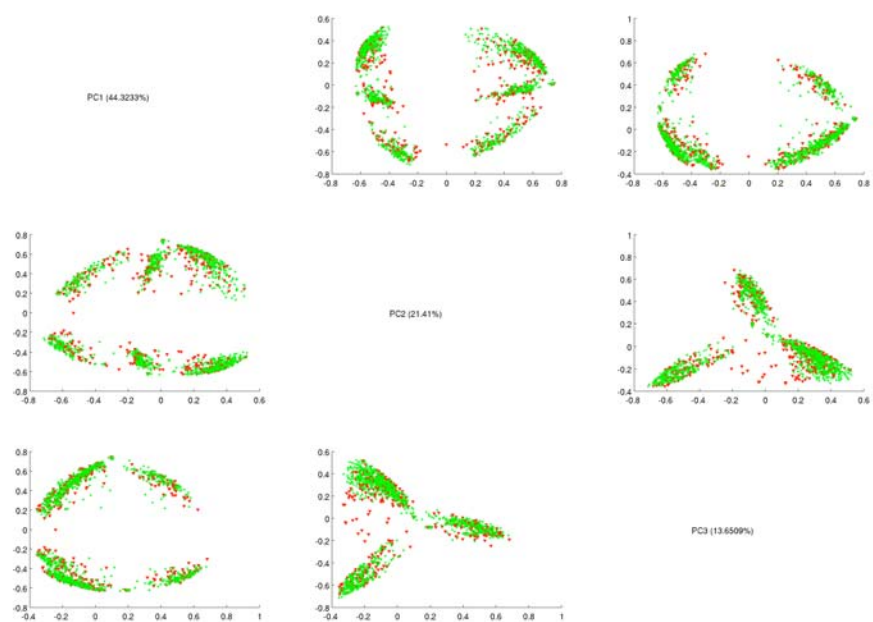
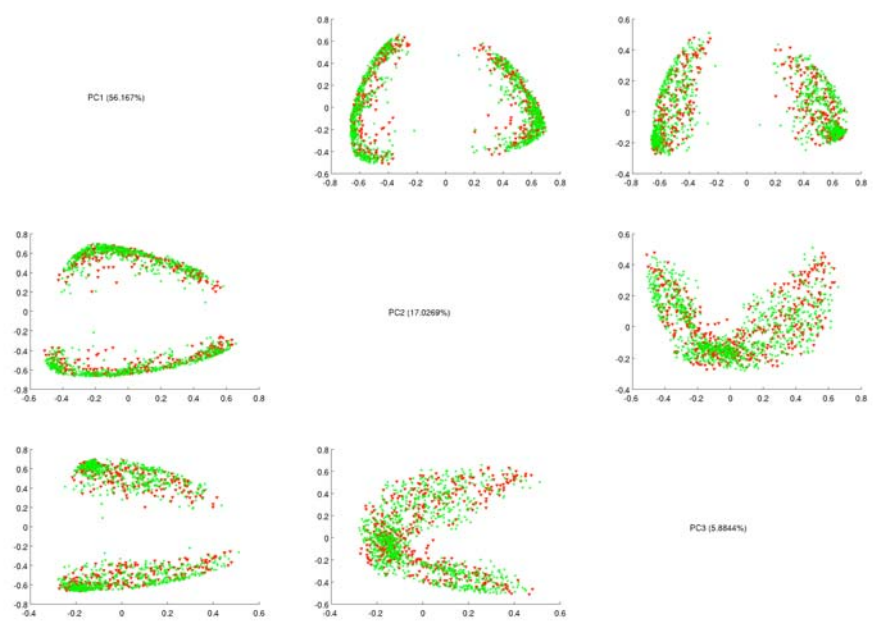


Figure 4-5: Further results obtained by performing PCA on all of the SELDI MS data to see if we were observing differences in the SELDI chip types being used in Figure 4-4 instead of variances in the data due to whether the subject suffered from PE or not. These plots represent the first three PCs compared to each other to see if there was any noticeable separation between the chip classes. Red markers represent the CM10 chips, green the H50 and blue represents the IMAC chip type. As can be seen from the plots, some clear clustering of data points can be observed, suggesting that by performing PCA on all of the SELDI datasets together, most of the variance captured by the technique is due to the SELDI chip type used.

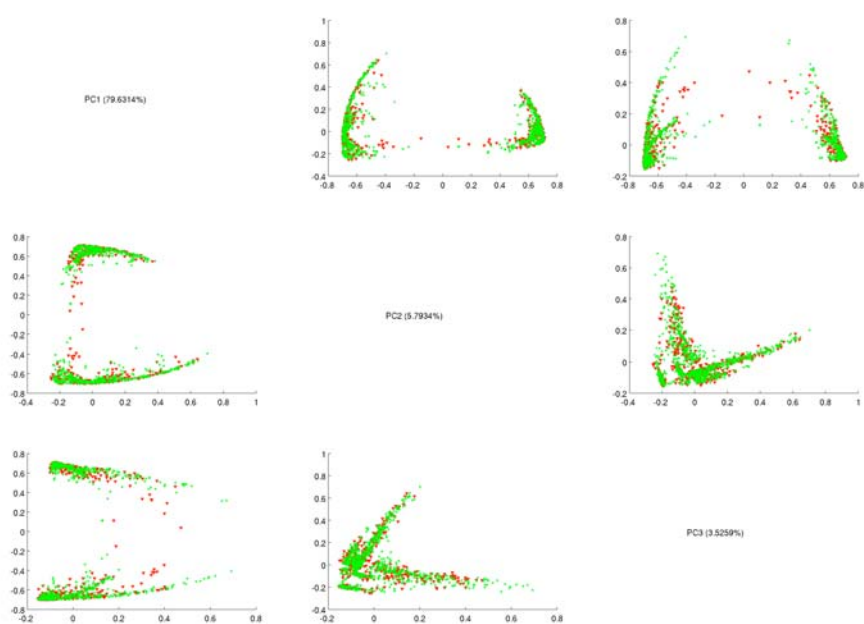
In order to remove the effects obviously being imposed on the data due to the type of SELDI protein chip being used, PCA was then performed on the data from the three SELDI chip types separately, the resulting plots of the first three PCs for each chip can be seen in Figure 4-6. As can be seen from the figure there was still no clear separation observed between the groups of pre-eclampsia sufferers and the controls, indicating that there was no simple means by which to separate the sample groups by disease state. If indeed it is going to be possible to predict the onset of pre-eclampsia, an alternative to PCA is going to need to be used – this is the reason two different classification approaches trialled in the next section.



(a)



(b)



(c)

Figure 4-6: Results from the PCA carried out on each chip type dataset separately. (a) Represents the CM10 SELDI chip, (b) the H50 chip, and (c) the IMAC chip type. As with Figure 4-4, the red markers indicate the PE group, and the green markers indicate the control group. As can be seen in the figures, no clear separation was observed between these groups (on any chip type) that could lead to further investigation with PCA.

4.2.2 Statistical Classifiers

As stated above, PLS-DA was the statistical classification technique of choice for this study. The dataset containing all of the identified peaks was the first to be analysed, the results of which can be seen in Table 4-2 and Table 4-3, followed by the dataset containing only the most common peaks between spectra; these results are presented in Table 4-4 and Table 4-5. All results presented used the test dataset containing 30% of all samples (27 samples – 9 cases, 18 controls).

In the first round of analysis using data from all of the identified peaks from the SELDI MS data; the PLS-DA classification models generated were similar in performance for the CM10 and IMAC protein chip types at around 50-70% accuracy. The H50 dataset produced on average better classification models and generated the best model with this dataset – the combined fractions 1 and 2 on the H50 chips provided the most accurate classifiers at 77.8 and 88.9% accurate when using high and low laser intensity respectively. The same was true of the quality of the classifications, in that models for the CM10 and IMAC datasets achieved averages of 45-49% sensitivity and 76-80% specificity, but the models based on the H50 dataset achieved averages of 50-59% sensitivity and 81-86% specificity. Looking more closely at the differences between the classifications made with the different sample fractions it was also observed that the data from the combined fractions 1 and 2 on average produced the better classification models.

This pattern was continued with the datasets made up of the most common peaks, (Table 4-4 and Table 4-5), the data from the H50 dataset created the best classification model using combined sample fractions 1 and 2 and low laser intensity as this not only produced a model that classified more of the samples correctly, but had the best combination of sensitivity and specificity scores of all of the common peaks datasets tested. As with the datasets containing all the peaks, looking more closely at the differences between the classifications made with the different sample fractions it was also observed that again the data from the combined fractions 1 and 2 on average produced the better classification models.

Table 4-2: Results for the PLS-DA models trained and tested against the peak lists generated by the R PROcess package. Reported values represent the percentage accuracy of the PLS-DA models created for each dataset when used on the test (30% of total samples) dataset.

Fraction	CM10		H50		IMAC		Mean
	High	Low	High	Low	High	Low	
1&2	74.1	51.9	77.8	88.9	59.3	59.3	68.5
3	66.7	70.4	51.9	66.7	66.7	70.4	65.4
4	51.9	70.4	70.4	66.7	63.0	77.8	66.7
5&6	51.9	70.4	70.4	74.1	63.0	37.0	61.1
Mean	61.1	65.7	67.6	74.1	63.0	61.1	

Table 4-3: A more in-depth look at the results obtained by the PLS-DA models created from the peak lists. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.

Fraction	CM10				H50				IMAC			
	High		Low		High		Low		High		Low	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1&2	64.3	100.0	35.7	69.2	61.5	92.9	80.0	94.1	40.0	70.6	41.7	73.3
3	50.0	76.5	53.3	91.7	30.0	64.7	50.0	84.6	50.0	84.6	53.8	85.7
4	35.7	69.2	54.5	81.3	53.8	85.7	50.0	80.0	46.2	78.6	61.5	92.9
5&6	37.5	72.7	54.5	81.3	54.5	81.3	58.3	86.7	46.7	83.3	25.0	54.5
Mean	46.9	79.6	49.5	80.8	50.0	81.1	59.6	86.3	45.7	79.3	45.5	76.6

Table 4-4: Results for the PLS-DA models trained and tested against the most common peaks identified by the R PROcess package. Reported values represent the percentage accuracy of the PLS-DA models for each dataset when used on the test (30% of total samples) dataset.

Fraction	CM10		H50		IMAC		Mean
	High	Low	High	Low	High	Low	
1&2	74.1	70.4	77.8	81.5	51.9	63.0	69.8
3	63.0	48.1	70.4	51.9	66.7	74.1	62.3
4	48.1	59.3	66.7	55.6	63.0	59.3	58.6
5&6	66.7	63.0	70.4	63.0	40.7	55.6	59.9
Mean	63.0	60.2	71.3	63.0	55.6	63.0	

Table 4-5: A more in-depth look at the results obtained by the PLS-DA models created from the most common peaks. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.

Fraction	CM10				H50				IMAC			
	High		Low		High		Low		High		Low	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1&2	57.1	92.3	53.8	85.7	63.6	87.5	70.0	88.2	38.9	77.8	45.5	75.0
3	47.4	100.0	35.3	70.0	55.6	77.8	33.3	66.7	50.0	80.0	57.1	92.3
4	30.8	64.3	37.5	68.4	50.0	80.0	38.5	71.4	45.5	75.0	43.8	81.8
5&6	50.0	80.0	44.4	72.2	53.8	85.7	46.2	78.6	26.7	58.3	40.0	75.0
Mean	46.3	84.1	42.8	74.1	55.8	82.7	47.0	76.2	40.3	72.8	46.6	81.0

4.2.3 Machine Learning Classification

As described in the methods above, SVMs were chosen as the machine-learning tool to be used to develop classification models for these datasets. This was partly due to favourable reviews in the literature, but also because of the good results obtained in the previous chapter. SVMs were constructed and trained using the 70% (of the whole dataset) training dataset (63 samples – 21 cases, 42 controls) as described previously using the *libsvm* tools, and tested using the 30% test dataset (27 samples – 9 cases, 18 controls).

In the first dataset to be analysed (all peaks identified), all three chip types produced good classification models with accuracies between 63 and 92% (Table 4-6 and Table 4-7), however the clearly better dataset with regards to these classification results was the H50 chip dataset as this produced the two most accurate classification models. This was also the case when the quality of the classifications was investigated as the classifiers built and tested on the H50 chip type were more accurate than the other two chip types. Using these datasets (all peaks) it was also observed that the best classification results were achieved using the combined sample fractions 1 and 2 and a high laser intensity.

The second dataset analysed with SVM models (the common peaks) achieved similar classification results (shown in Table 4-8 and Table 4-9), with all classification models achieving between 63 and 92% accuracy. Again it was the H50 datasets that produced the better classification models both in terms of the accuracy, and the sensitivity and specificity of the classifications made. Once again the combined sample fractions 1 and 2 produced the best models using high laser intensity.

Table 4-6: Results for the SVMs trained and tested against the peak lists generated by the R PROcess package. Reported values represent the percentage accuracy of the SVM created for each dataset when used on the test (30% of total samples) dataset.

Fraction	CM10		H50		IMAC		Mean
	High	Low	High	Low	High	Low	
1&2	66.7	70.4	92.6	88.9	59.3	66.7	74.1
3	85.2	74.1	77.8	77.8	70.4	70.4	75.9
4	74.1	77.8	77.8	74.1	70.4	66.7	73.5
5&6	63.0	77.8	66.7	81.5	74.1	66.7	71.6
Mean	72.2	75.0	78.7	80.6	68.5	67.6	

Table 4-7: A more in-depth look at the results obtained by the SVMs created from the peak lists. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.

Fraction	CM10				H50				IMAC			
	High		Low		High		Low		High		Low	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1&2	50.0	71.4	57.1	75.0	81.8	100.0	87.5	89.5	37.5	68.4	50.0	68.0
3	72.7	93.8	100.0	72.0	71.4	80.0	100.0	75.0	57.1	75.0	57.1	75.0
4	62.5	78.9	63.6	87.5	71.4	80.0	62.5	78.9	60.0	72.7	50.0	71.4
5&6	0.0	65.4	100.0	75.0	50.0	76.5	83.3	81.0	75.0	73.9	0.0	66.7
Mean	46.3	77.4	80.2	77.4	68.7	84.1	83.3	81.1	57.4	72.5	39.3	70.3

Table 4-8: Results for the SVMs trained and tested against the most common peaks identified by the R PROcess package. Reported values represent the percentage accuracy of the SVM created for each dataset when used on the test (30% of total samples) dataset.

Fraction	CM10		H50		IMAC		Mean
	High	Low	High	Low	High	Low	
1&2	74.1	77.8	92.6	85.2	66.7	70.4	77.8
3	63.0	70.4	74.1	66.7	77.8	59.3	68.6
4	66.7	74.1	66.7	77.8	66.7	63.0	69.2
5&6	88.9	85.2	59.3	70.4	70.4	66.7	73.5
Mean	73.1	76.9	73.1	75.0	70.4	64.8	

Table 4-9: A more in-depth look at the results obtained by the SVMs created from the most common peaks. This table gives calculated measures for the sensitivity (Sens.) and specificity (Spec.) of each model.

Fraction	CM10				H50				IMAC			
	High		Low		High		Low		High		Low	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1&2	62.5	78.9	100.0	75.0	88.9	94.4	55.6	81.8	50.0	73.7	66.7	70.8
3	40.0	68.2	54.5	81.3	75.0	73.9	50.0	71.4	66.7	83.3	42.9	76.9
4	50.0	69.6	66.7	76.2	50.0	71.4	66.7	83.3	0.0	66.7	45.5	75.0
5&6	87.5	89.5	72.7	93.8	33.3	66.7	53.8	85.7	53.8	85.7	0.0	66.7
Mean	60.0	76.5	73.5	81.5	61.8	76.6	56.5	80.6	42.6	77.3	38.7	72.4

4.2.4 Comparison of the Two Classification Approaches

Having applied two distinct classification approaches (statistical: PLS-DA and machine learning: SVMs) it was possible to compare the two methods to see if there was any clear advantage of using one classification tool over the other.

Figure 4-7 presents the differences between the PLS-DA and SVM models when used with the all peaks datasets to train and test the classifiers. This clearly shows that the SVM models produced generally better classifiers from this dataset, and that the best results were found using the H50 SELDI chip type, and the combination of sample fractions 1 and 2.

When the classification results generated from the common peaks datasets were compared (Figure 4-8), it was once again observed that the SVMs generally produced better classification results. With this set of samples again it was the H50 SELDI chip that produced the best classifier when in combination with the data obtained from fractions 1 and 2 (combined fractions).

Mean averages were also calculated for the sensitivity and specificity values recorded to identify which would be the most appropriate protein chip type for future classification models and tests. These values were collated with the mean averages of the accuracy of each classification model, and are presented in Figure 4-9. This plot shows that the H50 chip type gave the most useful information for the prediction of pre-eclampsia, as it produced the most effective classification models.

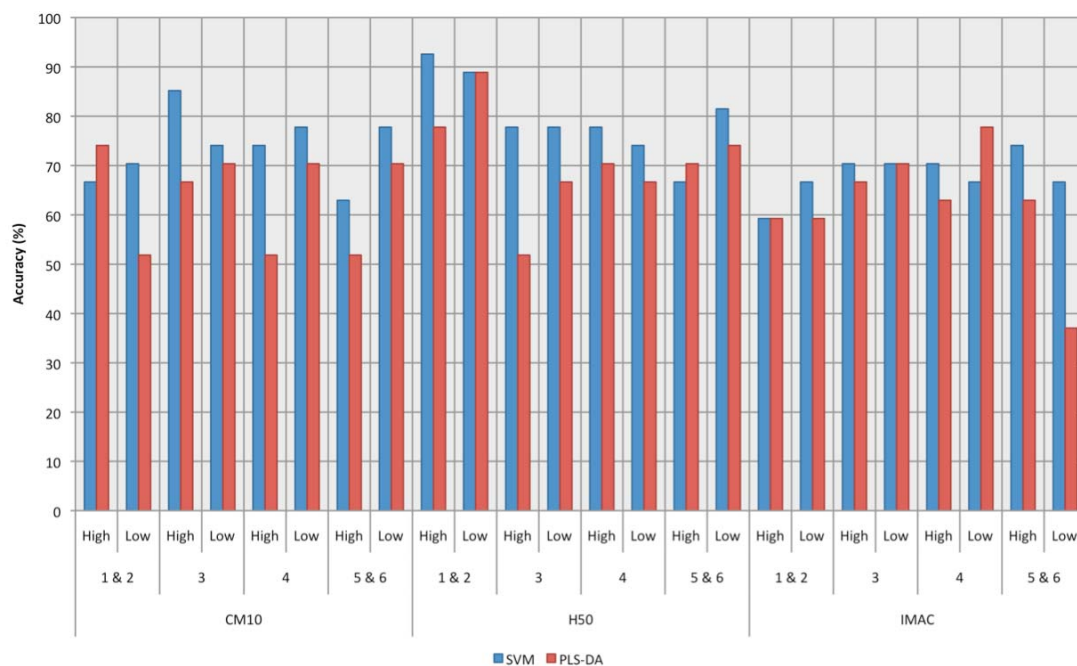


Figure 4-7: Comparison of the results obtained from all of the spectrum peaks identified by the R PROcess package (the 'peak lists') when they were used to train and test classifiers based on SVMs (blue bars) and PLS-DA (red bars). It can be clearly observed that the SVM models were the better classification tools with this dataset.

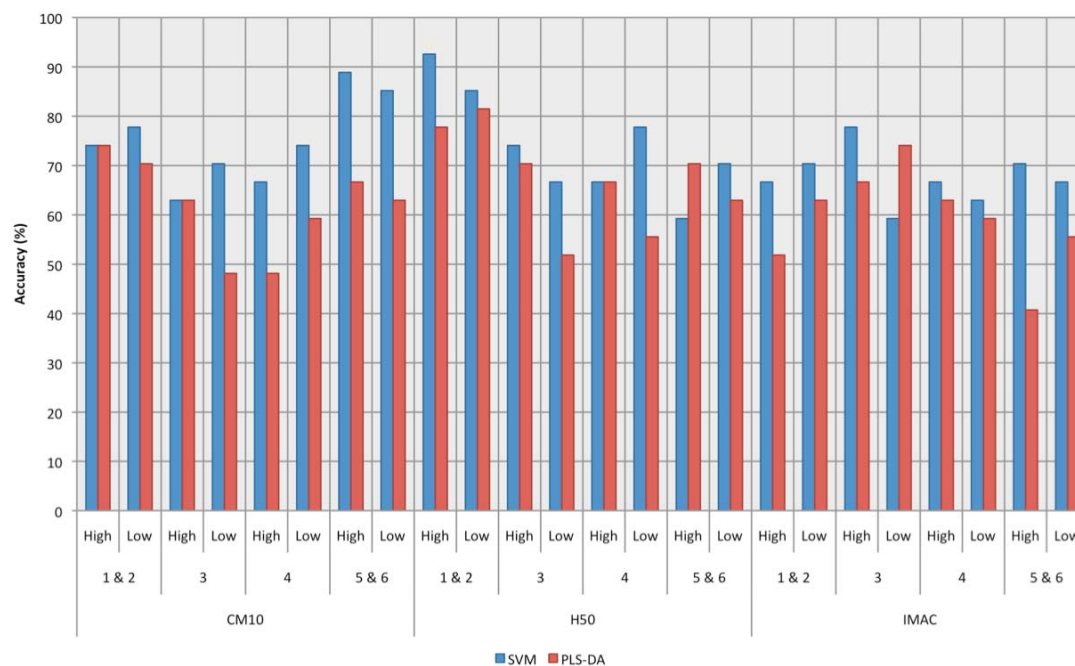


Figure 4-8: Comparison of the results obtained from the common peaks (identified by the R PROcess package) when they were used to train and test classifiers based on SVMs (blue bars) and PLS-DA (red bars). Once again it can be clearly observed that the SVM models were the better classification tools when used with this dataset.

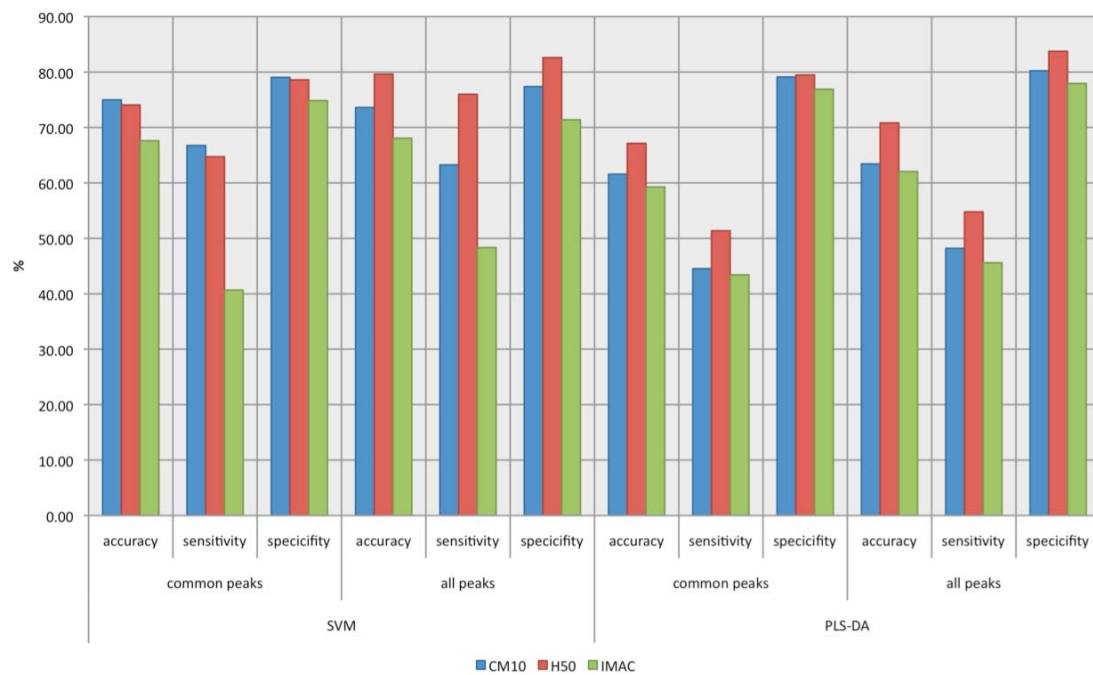


Figure 4-9: A comparison of the mean average classification accuracy, sensitivity and specificity achieved from the different SELDI MS protein chip types. Representing the data this way allows the clear observation that the most effective chip type to use for the classification of the pre-eclampsia data was the H50 chip (red bars) as it produced consistently better classification models.

4.3 Biomarker Identification

The second goal of this case study was to show the applicability of XPA and the associated data analysis methods to identify the most important features in the data. Of particular importance was to identify the more important peaks in the SELDI MS spectra from the point of view of classification – that is the peaks that contributed the most information as to whether the models decided a sample was from a pre-eclampsia sufferer or not.

In order for the data from the classification models to be used to identify the more informative points (peaks) in the dataset they must have first produced an effective classification of the samples. The analysis using PCA was not suitable for further investigation in this way as it was not possible to effectively separate the sample sets using this technique. It was also not possible to investigate the models formed by the SVMs no matter how effective they were as classifiers, as the models produced by SVMs are essentially ‘black boxes’, (described earlier in Section 1.8.2) in which the inner workings of the models are near impossible to discover how they arrive at their classifications.

However, the PLS-DA models were also effective in their classifications of the samples and were more suitable candidates for further investigation as it is possible to look at the loads attributed to each sample variable by the PLS-DA models, an approach that is common practice in the field of chemometrics (Malecha et al. 2003; Brereton 2003; Otto 1999).

The loadings matrix for the first principal component of each PLS-DA classification model was extracted, and standard deviations were calculated for each of these matrices. Data points (the m/z values representing peaks in the mass spectra) that had loads (absolute values) greater than two standard deviations from the mean were identified as the data points that contributed the most information towards the PLS-DA classification model.

It is to be noted that this analysis was only carried out on the common peaks datasets. The reason for this is that the process of biomarker identification is a form of data and noise reduction; the end goal is to extract only the most important points in the original complex datasets that have the potential to identify differing sample groups. The identification of the most common peaks in the SELDI spectra was itself a form of noise reduction, therefore it would not be appropriate to investigate the models from the all peaks datasets as this would re-introduce noise into this final step of analysis.

The results of the analysis (the data points identified as carrying more information to the PLS-DA models) are presented in Table 4-10. These peaks (indicated by their m/z value) represent the most important points observed in the SELDI MS datasets by the PLS-DA models.

Table 4-10: The m/z values (MS peaks) identified in the common peaks datasets as contributing the most information content to the PLS-DA models for the classification of patient samples.

Chip	Fraction	Laser	m/z
CM10	1&2	High	6428.47, 6627.46, 8108.57, 13552.4
		Low	3156.98, 6632.14, 8311.55, 79431.7
	3	High	28043.7
		Low	6814.4
	4	High	13817.1, 66873.7, 133991
		Low	9444.27
	5&6	High	4142.96, 4155.91, 4156.83, 4157.76, 14538.6
		Low	1498.41, 3436.89, 3480.81, 4171.2, 14549.9
H50	1&2	High	3150.93, 3151.74, 3152.55, 3219.83, 3438.56, 3484.61, 6621.04
		Low	3315.89, 3441.51, 4514.22
	3	High	3151.74, 3152.55, 3156.18, 8794.31, 8909.14, 15980.6, 15982.4, 16001.4, 17110, 17228.6, 17280.4
		Low	8801.04, 8909.14, 16002.3
	4	High	8791.62, 14536.9, 17227.6, 17825.5, 17864.9, 18030.3
		Low	1279.14, 1430, 8708.31, 14405.6, 14490.2
	5&6	High	14537.8, 14728, 17592.3
		Low	4153.6, 14552.5
IMAC	1&2	High	2847.05, 2847.82, 2850.12, 2850.89, 2852.42, 2853.95, 2855.49, 2937.43, 6630.38, 7429.59, 7437.03, 7712.02, 9496.66, 79395.3, 79413.4, 79429.6, 79437.7, 79468.1, 79476.2, 79488.3, 79494.4, 79662.5
		Low	1676.28, 6629.21
	3	High	4416.8, 5983.9, 6174.26
		Low	5974.48, 6178.21
	4	High	14024.7, 14360.9, 28063, 67070.7, 134219
		Low	67000.1
	5&6	High	4161.47, 4162.39, 4163.32, 4164.24, 4167.02, 4168.42, 4170.27, 4180.48, 4183.26, 4339.83
		Low	4174.92, 4457.95

4.3.1 Testing the Important MS Peaks

Following the identification of the more important peaks in the SELDI MS datasets, it was necessary to see if these data points were still able to produce effective classification models, or whether too much information content has been removed from the dataset.

As such, the data from these peaks was then extracted from the XPA database and formatted once again for classification with PLS-DA, using the same approaches as earlier within this study, classification models were trained using five-fold cross validation upon 70% of the dataset (63 samples – 21 cases, 42 controls), and then tested using the remaining 30% (27 samples – 9 cases, 18 controls). The results of these models are presented in Table 4-11 and Table 4-12, with a comparison of these results vs. the ‘all’ and ‘common’ peaks dataset represented in Figure 4-10.

As can be observed, the extraction of these individual peaks from the datasets has generally had a negative impact on the overall ability of the PLS-DA models to produce efficient classifiers, with the majority of the important peaks datasets producing worse classification models than the all/common peaks – indicating that too much information has been removed from the model, and that identifying individual peaks of importance within these datasets is not an optimal approach. There was however a few notable exceptions, with the most apparent being the H50 chip type and combined fractions 1&2, (already the most effective dataset from the previous analysis), producing efficient classification models using data from only the few peaks that have been extracted. This suggested that the

information contained within these data points were enough to distinguish between the two sample groups. This was also confirmed by testing the data from these peaks with an SVM model (again using the same procedures and metrics as earlier in the study), the results of which are presented in Table 4-13. These results show that the peaks effectively match the classifications made with the common peaks datasets.

These results indicate that the m/z values identified as the more important peaks for the H50 1&2 dataset, and as such the proteins that they represent (if they were to be identified) could be considered as potential protein biomarkers for the condition of pre-eclampsia.

Table 4-11: Results of the PLS-DA models trained and tested using only data for peaks identified as contributing the most information to the previous PLS-DA models. The cells with no results represent chip/fraction/laser combinations that did not contain enough important peaks to generate a new PLS-DA classification model.

Fraction	CM10		H50		IMAC		Mean
	High	Low	High	Low	High	Low	
1&2	77.8	59.3	77.8	81.5	66.7	44.4	67.9
3	-	-	48.1	44.4	51.9	70.4	53.7
4	55.6	-	55.6	37.0	66.7	-	53.7
5&6	59.3	44.4	66.7	59.3	48.1	51.9	54.9
Mean	64.2	51.9	62.0	55.6	58.3	55.6	

Table 4-12: Performance metrics for the PLS-DA models trained and tested using only data for peaks identified as contributing the most information to the previous PLS-DA models. The cells with no results represent chip/fraction/laser combinations that did not contain enough important peaks to generate a new PLS-DA classification model.

Fraction	CM10				H50				IMAC			
	High		Low		High		Low		High		Low	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
1&2	66.7	83.3	41.7	73.3	66.7	83.3	83.3	81.0	50.0	73.7	31.3	63.6
3	-	-	-	-	33.3	66.7	25.0	60.0	33.3	66.7	60.0	72.7
4	33.3	66.7	-	-	38.5	71.4	10.0	52.9	50.0	76.5	-	-
5&6	41.7	73.3	28.6	61.5	50.0	84.6	43.8	81.8	35.3	70.0	35.7	69.2
Mean	47.2	74.4	35.1	67.4	47.1	76.5	40.5	68.9	42.2	71.7	42.3	68.5

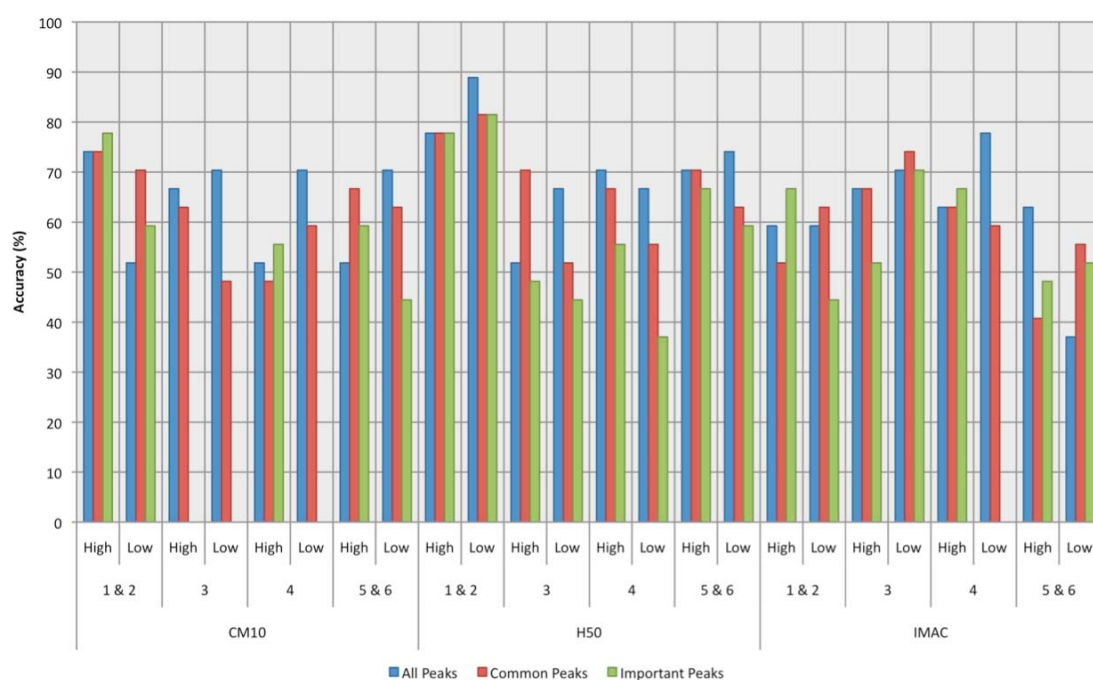


Figure 4-10: A comparison of the accuracies achieved by the PLS-DA models using the all, common and important peaks datasets. Looking at the datasets this way shows a trend in many of the chip/fraction/laser combinations that as the dataset becomes reduced (i.e. moving from all peaks to common peaks to important peaks) so too does the accuracy of the model produced. However, there are some exceptions to this rule – one particular example is the H50, fractions 1&2 datasets that consistently prove to build the best classification models.

Table 4-13: Results and performance metrics for the SVM model trained and tested using only data for the peaks identified as contributing the most information to the PLS-DA models for the H50 (fractions 1&2) datasets compared against the results achieved using the common peaks dataset.

Important Peaks

Chip	Fraction	Laser	Accuracy	Sensitivity	Specificity
H50	1&2	High	92.6% (25/27)	88.9%	94.4%
		Low	85.1% (23/27)	55.6%	100%

Common Peaks

Chip	Fraction	Laser	Accuracy	Sensitivity	Specificity
H50	1&2	High	92.6% (25/27)	88.9%	94.4%
		Low	85.2% (23/27)	55.6%	81.8%

4.4 Discussion and Conclusions

Two aims for this case study were set out at the onset:

- Demonstrate the use of XPA in the fast development of classification models to assess which of the three SELDI MS protein chip types in use in this study would be the most successful in use as part of a diagnostic test.
- Also demonstrate that due to the data analysis techniques used, the results produced from this study could be used to further explore the data and identify the more important points in the data, and therefore reveal potential biomarkers for future study.

4.4.1 Diagnostic Platform Development

With regards to the first aim of this study, XPA has been used successfully to produce a number of classification models to assess whether the results gained from any of the experiments carried out had the potential to be developed into a diagnostic test.

The H50 protein chip was identified as the most suitable chip for this purpose as the data generated from it helped to produce the most accurate classification models of all the chips tested.

It has also been demonstrated that both SVM and PLS-DA based classification models would be a suitable diagnostic classifier for the condition of pre-eclampsia,

and that the results achieved thus far would give high confidence in the classifications made.

At present, the only other published example of work towards the development of a diagnostic tool for pre-eclampsia (to draw a comparison of the quality of this work) is the study by Williams et al. (2007), describing the development of an ELISA-based immunoassay for measuring the quantity of P-IPG in urine and amniotic fluid, the accuracy and sensitivity/specificity of this test however was not discussed within the publication.

This clearly demonstrates the potential of XPA to the investigation of conditions such as pre-eclampsia, to go from raw data to solid classification metrics in a short timescale – one of the overall goals of this work.

4.4.2 Biomarker Identification

The second aim of this study was to identify the more important data points within the datasets in an effort to leverage the XPA system to aid in the identification of putative biomarkers.

Progress towards this goal was largely dependant on the success of the data separation with PCA or classification models generated with PLS-DA. The data generated by PCA was not of a high enough quality to be used for this process, but the PLS-DA classifications were sufficiently successful in their classifications of the sample groups to be suitable for this type of investigation.

Following a standard chemometric approach (Brereton 2003; Otto 1999) the

loadings of the variables in each dataset were investigated, and the more important points in the data were identified, as shown in Table 4-10. The data from these variables were then used to build and test further PLS-DA classification models. This extended analysis revealed that the peaks present in spectra from the H50 chip type, using combined fractions 1&2 and high laser intensity provided enough information content for the PLS-DA models to be as efficient as the models using all, and the most common of the MS peaks, with only a small sub-set of this data.

Clearly the peaks (and the proteins that they represent) identified at m/z values 3150.93, 3151.74, 3152.55, 3219.83, 3438.56, 3484.61, 6621.04 convey enough information to efficiently predict the condition of pre-eclampsia, and should be considered potential protein biomarkers for the syndrome.

If these proteins were to be investigated and their identities found, it could indicate new protein biomarkers for pre-eclampsia, or indeed if they were identified as any of the proteins currently linked to pre-eclampsia – P-IPG (Williams et al. 2007; Kunjara et al. 2000), sFlt1 (Ahmad & Ahmed 2004; Chaiworapongsa et al. 2004; Maynard et al. 2003; Zhou et al. 2002) and sEng (Mutter & Karumanchi 2008; Tjoa et al. 2007; Venkatesha et al. 2006) it would further confirm their relationship with the condition.

Identification of these proteins would also allow the development of diagnostic tests for the condition (via these protein biomarkers) to be investigated on platforms other than SELDI MS as the equipment needed for SELDI MS is very

expensive, it is not an ideal candidate for the development of a widespread diagnostic test as it would exclude establishments that could not afford the high expense of the equipment (as well as the expertise needed to run such equipment).

A more suitable platform for a regular diagnostic test would need to be identified, such as an immunoassay as was developed in the study by Williams et al. (2007) which would be far more cost effective and a much simpler test to implement, not requiring large investment or expertise from any laboratory wishing to carry out the test. Such technologies have already been used with good success in many conditions, but most notably in the detection of HIV infection (MacDonald et al. 1989).

5 Discussion and Conclusions

The aim of the work described in this thesis was to develop a generic system for the housing and analysis of data for biomarker discovery, and apply it to some real world clinical applications. The objectives set out to achieve this goal were as follows:

1. Design and construct a database schema that is flexible enough to be able to store and serve data from a variety of laboratory analysis techniques used in biomarker discovery and analysis.
2. Program an API for interacting with the database schema developed in objective 1. This should allow for the loading of data into and retrieving of data from the schema, thus allowing for the flexible construction of interfaces and data analysis routines.
3. Investigate the use of the developed data handling solutions (database and API) and construct data analysis routines for working with data from two different medical studies.
4. Evaluate the systems developed and results obtained in the case studies using real medical patient data.

All of these aims were achieved, and the analysis of data from medical studies was completed, with varying degrees of success. The outcomes are explored in the discussion section that follows.

5.1 Discussion

5.1.1 Generic Data Storage

One of the main challenges in the onset of this study was the development of a data storage system that would be suitable for housing the vast array of data types are produced from modern laboratory equipment in use in biomarker studies.

The solution that was developed within this study was a compromise in the fact that it does not actually house the raw data from the various laboratory instruments in use; instead it was chosen to only store the processed data relating to the molecules of interest in studies and any related meta-data that was of interest to XPA. The reason for this pragmatic compromise was that many other data storage solutions for the data formats in question are already in existence; and are no doubt in use already in the laboratories that use these techniques, therefore yet another alternative data storage solution is not entirely required.

Instead, the solution that was adopted was an approach similar to that used by the systems biology tool, Bioinformatics Resource Manager (BRM), previously discussed in Section 1.7.1 a central store for data directly related to XPA (the processed datasets and related meta-data), and links (in the form of references to file locations or other databases in which the data is stored) out to the raw data itself.

This arrangement means that XPA can carry all of the information about a given experiment and its subsequent data analysis but not have the burden of managing

the raw datasets itself, giving two key advantages to the system, in that it is both simpler in its implementation (than it would have been had data housing been an issue), and that development on the more interesting aspects of the system (data manipulation and analysis) could be completed much more quickly and efficiently. One disadvantage of this arrangement though would be when it comes to outputting data into standard data sharing formats such as the MIAME standard for microarray data (Brazma et al. 2001); as the raw data storage is not directly under the control of the XPA schema, it is not possible to confidently program export mechanisms for such a task – therefore the system would be reliant on the raw data storage solution or other pieces of software in use within a laboratory for producing such an output.

However, this is just a single disadvantage to an otherwise very flexible system that has the potential to house and interact with data from any laboratory analysis technique, and which has been shown to be fit for purpose during the case studies in this thesis.

5.1.2 Generic Data Analysis

The second major challenge in this study was the production of a generic workflow for analysing data obtained from biomarker studies. This was in the form of two related analysis paths:

- Looking at data from laboratory experiments with an aim to develop a classification model or diagnostic test.

- Looking at data from laboratory experiments with an aim to identify the most significant points in the dataset – i.e. the possible ‘biomarkers’.

Both these analysis paths were demonstrated in the two case studies within this thesis. In chapter 3, classification models were developed for the condition of prostate cancer from gene expression data (in combination with PSA measurements), and in chapter 4 classification models were also developed for the condition of pre-eclampsia from peptide expression data; but in this second instance the models were also investigated to reveal the most significant variables in the data; which given further study could lead to novel protein biomarkers for the condition of pre-eclampsia.

Both of these case studies shared the same basic methodology:

- Data pre-treatment (i.e. scaling and normalisation).
- Classification model optimisation.
- Classification model training (with optimal parameters).
- Classification model testing.

The vast majority of the code and procedures that was used to analyse the two datasets were also very much the same, with differences only being made for the different sample group names and the sizes of the datasets. This meant that data analysis was straightforward and the possibility of simple errors was reduced.

The advantage of the approach developed is simplicity – a standard set of tools are

used to do a distinct job, however this approach may not be so suitable for all types of data, where more flexibility is required. In particular, the current solution would not be suitable for multi-group classifications (with more than two sample groups) as the SVM implementation in use *libsvm* cannot cope with more than two classification groups, (whereas PLS-DA based models can), so the *libsvm* implementation would need to be substituted for an SVM implementation capable of multi-class classification such as the work described by Statnikov et al. (2005) and Ding & Dubchak (2001).

These issues however are merely a matter of developing and implementing the correct tools for the task that is being carried out by a given laboratory; after the initial development of a workflow suitable for a given task, the system would then be more than capable of carrying out the needed analysis task.

5.1.3 Conclusions

What has been shown thus far is that the overall aims of this study have been met.

- XPA has proven to be a suitable system for housing and processing data generated from modern laboratory techniques.
- The system has been used successfully with two case studies generating data on different conditions using very different laboratory methodologies.
 - o In the prostate cancer study (Chapter 3), XPA was used to generate classification models that were far more accurate and sensitive than the established PSA test.

- o In the pre-eclampsia study (Chapter 4), the XPA system was used for two threads of analysis.
 - To generate effective classification models that were able to distinguish between case and control subjects using the data generated directly from SELDI MS.
 - It was also used to identify important points within the data (peaks) that possessed a higher discriminatory power compared to the other observed data points. These were then used to generate more focused classification models, and provided targets for future identification.
- The use of multiple data analysis techniques (SVMs, PLS-DA and PCA) demonstrated the ability of XPA to be used with a multitude of analysis workflows and gives confidence that it could be adapted for use with any given analysis technique.
- The XPA database schema has proved suitable in handling data from two very different laboratory analysis techniques (qPCR and SELDI-TOF MS), indicating that it has the potential to be flexible enough for almost any type of data formats that would be required in future studies.

5.2 Future Work

This final section of the thesis details potential future work that would be

beneficial for the progression of the XPA system towards a complete solution for biomarker discovery and diagnostic test development. It also discusses other areas of work where the concepts and indeed the tools generated in this thesis may be applicable and what work would have to be carried out in order for this to be a possibility.

5.2.1 Further Data Integration

During the development of the XPA schema and systems it was envisaged that an end goal of the system would be the potential of analysing data from multiple laboratory techniques looking at targets such as genes, proteins and metabolites as a single dataset. Unfortunately such data was not available during the course of the thesis study, so this aspect of the XPA system was not tested. If any further work was to be carried out with XPA, it would be interesting to see if this type of analysis would indeed be possible.

5.2.2 Integrated Automatic Data Analysis

In the two case study sections of this thesis there were descriptions of the how the work was carried out to achieve the aim of each study. In each case there were two similar steps:

- Data was imported to XPA to allow a uniform way of accessing the data.
- Data was then exported from the XPA schema into a format suitable for the following rounds of data analysis.

Ultimately, this is one of the aspects of the XPA system as it currently stands that is flawed, and is therefore one of the next steps of work that needs to be addressed, should any work be carried on with XPA. Fully automated analysis of biomarker and diagnostic test data would greatly increase the throughput of modern biomarker discovery analysis as it would allow for (once set up), an automatic workflow of data being loaded into the database, followed immediately by data processing and analysis giving the ultimate end point of whether the data the laboratory has would be suitable for identifying the condition that they are interested in or not.

This goal could be achieved through the following areas of work:

- Develop methodology and tools for automatically selecting the optimal number of LVs for use in PLS-DA based analysis.
- Develop tools for the automated interrogation of PCA and PLS-DA models in order to extract and identify the most significant variables within the datasets.

5.2.3 Interfacing With XPA

Another notable flaw with the current XPA system is the way in which it is used; this is presently through a series of Perl programs for importing and exporting data – this is not a suitable interface for such a system. As such, another future goal for the further development of XPA would be the refactoring of the XPA API into a full object-orientated codebase as this would make further development far

more flexible and easier. This should also be accompanied by the production of a graphical interface so that users can tell what data is in the system, interact with it, load more data in, and even carry out data analysis within the confines of an XPA interface.

The development of an interface to XPA would not be a huge task as a lot of work has already been done in the development of the API for importing, exporting and interacting with data from the system, it is just the user facing side of the software that needs to be developed.

5.2.4 Systems Biology

The focus of this thesis study, and indeed the development of the XPA system was solely aimed at the realm of biomarker discovery and diagnostic test development. However, many of the ideas and concepts of data integration are shared with tools in another field – systems biology. Therefore it is envisaged that it would not be entirely impossible for the analysis routines in use within XPA to be expanded to look at situations such as interaction data. If given some time and thought this would most definitely be a worthwhile future direction for the development of the XPA system.

Bibliography

- Adam, B. et al., 2002. Serum Protein Fingerprinting Coupled with a Pattern-matching Algorithm Distinguishes Prostate Cancer from Benign Prostate Hyperplasia and Healthy Men. *Cancer Res*, 62(13), 3609-3614.
- Adams, M.D. et al., 2000. The genome sequence of *Drosophila melanogaster*. *Science (New York, N.Y.)*, 287(5461), 2185-95.
- Affymetrix, 2009. Affymetrix - Arrays. Available at: http://www.affymetrix.com/products_services/arrays/index.affx [Accessed February 5, 2009].
- Ahmad, S. & Ahmed, A., 2004. Elevated placental soluble vascular endothelial growth factor receptor-1 inhibits angiogenesis in preeclampsia. *Circulation Research*, 95(9), 884-91.
- American Diabetes Association, 2007. Standards of Medical Care in Diabetes--2007. *Diabetes Care*, 30(suppl_1), S4-41.
- Anderson, N.L. et al., 2004. The Human Plasma Proteome: A Nonredundant List Developed by Combination of Four Separate Sources. *Mol Cell Proteomics*, 3(4), 311-326.
- Atkinson, A.J. et al., 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*. *Clin Pharmacol Ther*, 69(3), 89-95.
- Balding, D.J., Bishop, M. & Cannings, C., 2007. *Handbook of Statistical Genetics* 3rd ed., WileyBlackwell.
- Bassett, D.E., Eisen, M.B. & Boguski, M.S., 1999. Gene expression informatics--it's all in your mine. *Nature Genetics*, 21(1 Suppl), 51-5.
- Blattner, F.R. et al., 1997. The complete genome sequence of *Escherichia coli* K-12. *Science (New York, N.Y.)*, 277(5331), 1453-74.
- Bo Nordén, P.B., 2005. Analysis and Understanding of High-Dimensionality Data by Means of Multivariate Data Analysis. *Chemistry & Biodiversity*, 2(11), 1487-1494.
- Brazma, A. et al., 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4), 365-71.
- Brereton, R.G., 2003. *Chemometrics: data analysis for the laboratory and chemical*

plant, John Wiley and Sons.

Brown, P.O. & Botstein, D., 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(1 Suppl), 33-7.

Cancer Research UK, 2009. Cancer Research UK : Prostate Cancer screening.
Available at:
<http://info.cancerresearchuk.org/cancerstats/types/prostate/screening/>
[Accessed February 7, 2009].

Cancer Research UK, Cancer Research UK : UK Prostate Cancer mortality statistics.
Available at:
<http://info.cancerresearchuk.org/cancerstats/types/prostate/mortality/>
[Accessed May 17, 2008].

Cancer Research UK, 2008. CancerHelp UK: Screening for prostate cancer.
Available at: <http://www.cancerhelp.org.uk/help/default.asp?page=2719>
[Accessed December 28, 2008].

Cappuzzo, F. et al., 2005. Epidermal Growth Factor Receptor Gene and Protein and Gefitinib Sensitivity in Non-Small-Cell Lung Cancer. *J. Natl. Cancer Inst.*, 97(9), 643-655.

Carninci, P., 2007. Constructing the landscape of the mammalian transcriptome. *J Exp Biol*, 210(9), 1497-1506.

Carrasquillo, M.M. et al., 2009. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet*, 41(2), 192-198.

Catalona, W. et al., 1991. Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med*, 324(17), 1156-1161.

Chaiworapongsa, T. et al., 2004. Evidence supporting a role for blockade of the vascular endothelial growth factor system in the pathophysiology of preeclampsia. Young Investigator Award. *American Journal of Obstetrics and Gynecology*, 190(6), 1541-7; discussion 1547-50.

Chang, C. & Lin, C., *LIBSVM: A Library for Support Vector Machines*, Available at: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

Christmas, R. et al., 2005. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. In pp. 12-16 . Available at: <http://educationbook.aacrjournals.org/cgi/content/full/2005/1/12>
[Accessed July 19, 2008].

Collins, F.S. et al., 2003. A vision for the future of genomics research. *Nature*, 422(6934), 835-847.

- Cook Jr, E.H. & Scherer, S.W., 2008. Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(7215), 919-923.
- Creighton, C.J., 2008. Multiple oncogenic pathway signatures show coordinate expression patterns in human prostate tumors. *PLoS ONE*, 3(3), e1816.
- Dalma-Weiszhausz, D.D. et al., 2006. The Affymetrix GeneChip® Platform: An Overview. In *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*. Academic Press, pp. 3-28. Available at: <http://www.sciencedirect.com/science/article/B7CV2-4KS57R8-1/2/9550500583b04490767160f3ad154448> [Accessed January 11, 2009].
- De Gruttola, V.G. et al., 2001. Considerations in the Evaluation of Surrogate Endpoints in Clinical Trials: Summary of a National Institutes of Health Workshop. *Controlled Clinical Trials*, 22(5), 485-502.
- De Jong, S., 1993. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3), 251-263.
- Deutsch, E. et al., *Systems Biology Experiment Analysis Management System (SBEAMS)*, Institute for Systems Biology. Available at: <http://www.sbeams.org/>.
- Diamandis, E.P., 1998. Prostate-specific Antigen: Its Usefulness in Clinical Medicine. *Trends in Endocrinology and Metabolism*, 9(8), 310-316.
- Ding, C.H. & Dubchak, I., 2001. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4), 349-358.
- Eder, I.E. et al., 2004. Genes differentially expressed in prostate cancer. *BJU international*, 93(8), 1151-5.
- Engwegen, J.Y. et al., 2006. Clinical proteomics: searching for better tumour markers with SELDI-TOF mass spectrometry. *Trends in Pharmacological Sciences*, 27(5), 251-259.
- Fan, J. et al., 2006. Illumina Universal Bead Arrays. In *DNA Microarrays, Part A: Array Platforms and Wet-Bench Protocols*. Academic Press, pp. 57-73. Available at: <http://www.sciencedirect.com/science/article/B7CV2-4KS57R8-3/2/2f4812419811fb42a1656539c8d12b31> [Accessed February 5, 2009].
- Fenske, C., 2007. *A New Sensitive and Accurate Test for Prostate Cancer*, St George's, University of London. Available at: <http://www.prostateuk.org/research/reports/R200512-05-fens->

20070220.htm [Accessed May 16, 2008].

- Feuk, L., Carson, A.R. & Scherer, S.W., 2006. Structural variation in the human genome. *Nat Rev Genet*, 7(2), 85-97.
- Fiers, W. et al., 1978. Complete nucleotide sequence of SV40 DNA. *Nature*, 273(5658), 113-20.
- Frank, R. & Hargreaves, R., 2003. Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov*, 2(7), 566-580.
- Furey, T.S. et al., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10), 906-914.
- Gelfand, M.S., 1995. Prediction of function in DNA sequence analysis. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 2(1), 87-115.
- Gentleman, R. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80.
- Gentleman, R. et al., 2005. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer-Verlag New York Inc.
- Gregory, S.G. et al., 2002. A physical map of the mouse genome. *Nature*, 418(6899), 743-750.
- Grönberg, H., 2003. Prostate cancer epidemiology. *The Lancet*, 361(9360), 859-864.
- Gunderson, K.L. et al., 2004. Decoding randomly ordered DNA arrays. *Genome Research*, 14(5), 870-7.
- Hakonarson, H. et al., 2005. Effects of a 5-Lipoxygenase-Activating Protein Inhibitor on Biomarkers Associated With Risk of Myocardial Infarction: A Randomized Trial. *JAMA*, 293(18), 2245-2256.
- Hardiman, G., 2004. Microarray platforms--comparisons and contrasts. *Pharmacogenomics*, 5(5), 487-502.
- Helgadottir, A. et al., 2004. The gene encoding 5-lipoxygenase activating protein confers risk of myocardial infarction and stroke. *Nat Genet*, 36(3), 233-239.
- Hewitt, S.M., Dear, J. & Star, R.A., 2004. Discovery of Protein Biomarkers for Renal Diseases. *J Am Soc Nephrol*, 15(7), 1677-1689.

- Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.
- Illumina, 2009a. About Illumina - Technology: BeadArray Technology - Illumina, Inc. Available at: <http://www.illumina.com/pages.ilmn?ID=5> [Accessed February 5, 2009].
- Illumina, 2009b. Products & Services - Whole-Genome Genotyping & CNV Analysis: Overview - Illumina, Inc. Available at: <http://www.illumina.com/pages.ilmn?ID=40> [Accessed February 5, 2009].
- Jansen, B.J.H. & Schalkwijk, J., 2003. Transcriptomics and proteomics of human skin. *Brief Funct Genomic Proteomic*, 1(4), 326-341.
- Jones, A.R. et al., 2007. The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotech*, 25(10), 1127-1133.
- Jr, G.W. et al., 1999. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer and Prostatic Diseases*, 2(5/6), 264-276.
- Karas, M., Bachmann, D. & Hillenkamp, F., 1985. Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Anal. Chem.*, 57(14), 2935-2939.
- Karp, N.A. & Griffin, J.L., 2005. Application of partial least squares discriminant analysis to two-dimensional difference gel studies in expression proteomics. *PROTEOMICS*, 5(1), 81-90.
- Kehrer-Sawatzki, H., 2007. What a difference copy number variation makes. *BioEssays*, 29(4), 311-313.
- Kong, F. et al., 2006. Using proteomic approaches to identify new biomarkers for detection and monitoring of ovarian cancer. *Gynecologic Oncology*, 100(2), 247-253.
- Kunjara, S. et al., 2000. Inositol Phosphoglycans and Signal Transduction Systems in Pregnancy in Preeclampsia and Diabetes: Evidence for a Significant Regulatory Role in Preeclampsia at Placental and Systemic Levels. *Molecular Genetics and Metabolism*, 69(2), 144-158.
- Lee, J.A. & Lupski, J.R., 2006. Genomic Rearrangements and Gene Copy-Number Alterations as a Cause of Nervous System Disorders. *Neuron*, 52(1), 103-121.

- Li, J. et al., 2002. Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer. *Clin Chem*, 48(8), 1296-1304.
- Li, X., *PROcess: Ciphergen SELDI-TOF Processing*,
- Lilja, H., Ulmert, D. & Vickers, A.J., 2008. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nat Rev Cancer*, 8(4), 268-278.
- Liotta, L.A., Ferrari, M. & Petricoin, E., 2003. Clinical proteomics: Written in blood. *Nature*, 425(6961), 905.
- Lipshutz, R.J. et al., 1999. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(1 Suppl), 20-4.
- Lukes, L. et al., 2009. The Origins of Breast Cancer Prognostic Gene Expression Profiles. *Cancer Res*, 69(1), 310-318.
- Lupski, J.R. & Stankiewicz, P., 2005. Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes. *PLoS Genet*, 1(6), e49.
- Lupski, J.R., 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends in Genetics*, 14(10), 417-422.
- MacDonald, K.L. et al., 1989. Performance characteristics of serologic tests for human immunodeficiency virus type 1 (HIV-1) antibody among Minnesota blood donors. Public health and clinical implications. *Annals of Internal Medicine*, 110(8), 617-21.
- MacKay, D.J.C., 2002. *Information Theory, Inference & Learning Algorithms* 1st ed., Cambridge University Press.
- Malecha, M., Bessant, C. & Saini, S., 2003. Identification of oil contaminants on polymer coated beverage cans using fluorescence spectroscopy. *Applied Spectroscopy*, 57(8), 1042-8.
- Maraganore, D.M. et al., 2005. High-resolution whole-genome association study of Parkinson disease. *American Journal of Human Genetics*, 77(5), 685-93.
- Mardia, K.V., Kent, J.T. & Bibby, J.M., 1980. *Multivariate Analysis*, Academic Press.
- Marengo, E. et al., 2008. Application of partial least squares discriminant analysis and variable selection procedures: a 2D-PAGE proteomic study. *Analytical and Bioanalytical Chemistry*, 390(5), 1327-1342.
- Marzolf, B. et al., 2006. SBEAMS-Microarray: database software supporting

- genomic expression analyses for systems biology. *BMC Bioinformatics*, 7(1), 286.
- Maynard, S.E. et al., 2003. Excess placental soluble fms-like tyrosine kinase 1 (sFlt1) may contribute to endothelial dysfunction, hypertension, and proteinuria in preeclampsia. *The Journal of Clinical Investigation*, 111(5), 649-58.
- Mootha, V.K. et al., 2003. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3), 267-273.
- Musumarra, G. et al., 2005. Genome-based identification of diagnostic molecular markers for human lung carcinomas by PLS-DA. *Computational Biology and Chemistry*, 29(3), 183-195.
- Mutter, W.P. & Karumanchi, S.A., 2008. Molecular mechanisms of preeclampsia. *Microvascular Research*, 75(1), 1-8.
- Nakabachi, A. et al., 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science (New York, N.Y.)*, 314(5797), 267.
- Nimeus-Malmstrom, E. et al., 2008. Gene expression profiling in primary breast cancer distinguishes patients developing local recurrence after breast-conservation surgery, with or without postoperative radiotherapy. *Breast Cancer Research*, 10(2), R34.
- Nixon, R.G. et al., 1997. Biological Variation of Prostate Specific Antigen Levels In Serum: An Evaluation of Day-to-Day Physiological Fluctuations in a Well-Defined Cohort of 24 Patients. *The Journal of Urology*, 157(6), 2183-2190.
- Noble, W.S., 2006. What is a support vector machine? *Nat Biotech*, 24(12), 1565-1567.
- Otto, M., 1999. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH.
- Pease, A.C. et al., 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 91(11), 5022-6.
- Petricoin III, E.F. et al., 2002. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359(9306), 572-577.
- Pipinikas, C.P. et al., 2007. Measurement of blood E2F3 mRNA - in prostate cancer by quantitative RT-PCR: a preliminary study. *Biomarkers*, 12(5), 541.

- Pirooznia, M. et al., 2008. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(Suppl 1), S13.
- Potosky, A.L. et al., 1995. The role of increasing detection in the rising incidence of prostate cancer. *JAMA*, 273(7), 548-552.
- R Development Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at: <http://www.r-project.org/> [Accessed March 4, 2008].
- Rai, A.J. et al., 2002. Proteomic approaches to tumor marker discovery. *Archives of Pathology & Laboratory Medicine*, 126(12), 1518-26.
- Rich, S., 2008. Approaching Biomarker Discovery through Genomics. *Journal of Cardiovascular Translational Research*, 1(1), 21-24.
- Ridley, M., 2000. *Genome: The Autobiography of a Species in 23 Chapters* Abridged e., HarperCollins.
- Rocha, D. et al., 2006. Seventh international meeting on single nucleotide polymorphism and complex genome analysis: 'ever bigger scans and an increasingly variable genome'. *Human Genetics*, 119(4), 451-6.
- Salama, G. et al., 2007. Seasonality of Serum Prostate-Specific Antigen Levels: A Population-Based Study. *European Urology*, 52(3), 708-714.
- Schena, M. et al., 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(20), 10614-9.
- Segal, E. et al., 2005. From signatures to models: understanding cancer using microarrays. *Nature Genetics*, 37 Suppl, S38-45.
- Selley, S. et al., 1997. Diagnosis, management and screening of early localised prostate cancer. *Health Technology Assessment (Winchester, England)*, 1(2), i, 1-96.
- Shah, A.R. et al., 2007. Enabling high-throughput data management for systems biology: The Bioinformatics Resource Manager. *Bioinformatics*, 23(7), 906-909.
- Shannon, P. et al., 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, 13(11), 2498-2504.
- Shen, R. et al., 2005. High-throughput SNP genotyping on universal bead arrays. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*,

573(1-2), 70-82.

Sherry, S.T. et al., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1), 308-11.

Sibai, B.M., Caritis, S. & Hauth, J., 2003. What we have learned about preeclampsia. *Seminars in Perinatology*, 27(3), 239-46.

Simpkins, F. et al., 2005. SELDI-TOF mass spectrometry for cancer biomarker discovery and serum proteomic diagnostics. *Pharmacogenomics*, 6(6), 647-653.

Sorich, M.J. et al., 2003. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. *Journal of Chemical Information and Computer Sciences*, 6(43), 2019-2024.

Sparkman, O.D., 2000. *Mass Spectrometry Desk Reference*, Global View Pub.

Stajich, J.E. et al., 2002. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Res.*, 12(10), 1611-1618.

Statnikov, A. et al., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5), 631-643.

Stemers, F.J. & Gunderson, K.L., 2007. Whole genome genotyping technologies on the BeadArrayTM platform. *Biotechnology Journal*, 2(1), 41-49.

The International HapMap Consortium, 2003. The International HapMap Project. *Nature*, 426(6968), 789-796.

The MathWorks, 2009. Neural Network Toolbox. Available at: <http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/> [Accessed February 6, 2009].

Tjoa, M.L., Levine, R.J. & Karumanchi, S.A., 2007. Angiogenic factors and preeclampsia. *Frontiers in Bioscience: A Journal and Virtual Library*, 12, 2395-402.

Vapnik, V.N., 1998. *Statistical Learning Theory*, Wiley-Interscience.

Velculescu, V.E. et al., 1997. Characterization of the Yeast Transcriptome. *Cell*, 88(2), 243-251.

Venkatesha, S. et al., 2006. Soluble endoglin contributes to the pathogenesis of

- preeclampsia. *Nat Med*, 12(6), 642-649.
- Voshol, H. et al., 2005. Evaluation of Biomarker Discovery Approaches to Detect Protein Biomarkers of Acute Renal Allograft Rejection. *Journal of Proteome Research*, 4(4), 1192-1199.
- Walker, J.J., 2000. Pre-eclampsia. *The Lancet*, 356(9237), 1260-1265.
- Weston, A.D. & Hood, L., 2004. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3(2), 179-96.
- White, A.M. et al., 2005. Automated Microarray Image Analysis Toolbox for MATLAB. *Bioinformatics*, 21(17), 3578-3579.
- Wigle, D.T. et al., 2008. Role of hormonal and other factors in human prostate cancer. *Journal of toxicology and environmental health. Part B, Critical reviews*, 11(3-4), 242-59.
- Williams, P.J. et al., 2007. Inositol phosphoglycan P-type in preeclampsia: a novel marker? *Hypertension*, 49(1), 84-9.
- Yeang, C. et al., 2001. Molecular classification of multiple tumor types. *Bioinformatics*, 17(suppl_1), S316-322.
- Zarrinkar, P.P. et al., 2001. Arrays of arrays for high-throughput gene expression profiling. *Genome Research*, 11(7), 1256-61.
- Zhang, M.Q., 2002. Computational prediction of eukaryotic protein-coding genes. *Nature Reviews. Genetics*, 3(9), 698-709.
- Zhou, Y. et al., 2002. Vascular endothelial growth factor ligands and receptors that regulate human cytotrophoblast survival are dysregulated in severe preeclampsia and hemolysis, elevated liver enzymes, and low platelets syndrome. *The American Journal of Pathology*, 160(4), 1405-23.