



Issues in Toxicology

# Metabolic Profiling

Disease and Xenobiotics

Edited by Martin Grootveld



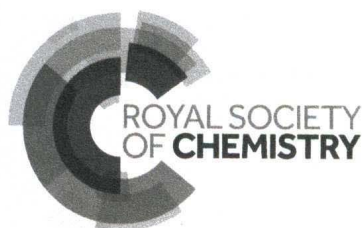
# ***Metabolic Profiling Disease and Xenobiotics***

Edited by

**Martin Grootveld**

*De Montfort University, Leicester, UK*

*Email: [mgrootveld@dmu.ac.uk](mailto:mgrootveld@dmu.ac.uk)*



Issues in Toxicology No. 21

Print ISBN: 978-1-84973-163-8

PDF eISBN: 978-1-84973-516-2

ISSN: 1757-7179

A catalogue record for this book is available from the British Library

© The Royal Society of Chemistry 2015

*All rights reserved*

*Apart from fair dealing for the purposes of research for non-commercial purposes or for private study, criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulations 2003, this publication may not be reproduced, stored or transmitted, in any form or by any means, without the prior permission in writing of The Royal Society of Chemistry or the copyright owner, or in the case of reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the licences issued by the appropriate Reproduction Rights Organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to The Royal Society of Chemistry at the address printed on this page.*

The RSC is not responsible for individual opinions expressed in this work.

Published by The Royal Society of Chemistry,  
Thomas Graham House, Science Park, Milton Road,  
Cambridge CB4 0WF, UK

Registered Charity Number 207890

For further information see our website at [www.rsc.org](http://www.rsc.org)

# Metabolic Profiling

## Disease and Xenobiotics

## Issues in Toxicology

### *Series Editors:*

Professor Diana Anderson, *University of Bradford, UK*

Dr Michael Waters, *Integrated Laboratory Systems Inc., N Carolina, USA*

Dr Timothy C Marrs, *Edentox Associates, Kent, UK*

### *Adviser to the Board:*

Alok Dhawan, *Ahmedabad University, India*

### *Titles in the Series:*

- 1: Hair in Toxicology: An Important Bio-Monitor
- 2: Male-mediated Developmental Toxicity
- 3: Cytochrome P450: Role in the Metabolism and Toxicity of Drugs and other Xenobiotics
- 4: Bile Acids: Toxicology and Bioactivity
- 5: The Comet Assay in Toxicology
- 6: Silver in Healthcare
- 7: *In Silico* Toxicology: Principles and Applications
- 8: Environmental Cardiology
- 9: Biomarkers and Human Biomonitoring, Volume 1: Ongoing Programs and Exposures
- 10: Biomarkers and Human Biomonitoring, Volume 2: Selected Biomarkers of Current Interest
- 11: Hormone-Disruptive Chemical Contaminants in Food
- 12: Mammalian Toxicology of Insecticides
- 13: The Cellular Response to the Genotoxic Insult: The Question of Threshold for Genotoxic Carcinogens
- 14: Toxicological Effects of Veterinary Medicinal Products in Humans: Volume 1
- 15: Toxicological Effects of Veterinary Medicinal Products in Humans: Volume 2
- 16: Aging and Vulnerability to Environmental Chemicals: Age-related Disorders and their Origins in Environmental Exposures
- 17: Chemical Toxicity Prediction: Category Formation and Read-Across
- 18: The Carcinogenicity of Metals: Human Risk through Occupational and Environmental Exposure

- 19: Reducing, Refining and Replacing the Use of Animals in Toxicity Testing
- 20: Advances in Dermatological Sciences
- 21: Metabolic Profiling: Disease and Xenobiotics

*How to obtain future titles on publication:*

A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

*For further information please contact:*

Book Sales Department, Royal Society of Chemistry, Thomas Graham House,  
Science Park, Milton Road, Cambridge, CB4 0WF, UK

Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247

Email: [booksales@rsc.org](mailto:booksales@rsc.org)

Visit our website at [www.rsc.org/books](http://www.rsc.org/books)



# Preface

This book represents the culmination of at least several years' relatively intensive work, and provides an in-depth and sometimes highly critical review of research investigations performed in the metabolomics research area and, more generally, that concerning the 'omics' fields in general (for example, proteomics and genomics, *etc.*). My major objective was primarily to provide valuable advice from my own original, basic grounding in the statistical analysis of datasets with a biomolecular focus or otherwise. However, as the volume of work progressed, it became clearer to me that more and more researchers involved in these areas are, at least some of the time, keen to experience a revelation of some kind, and are utilising the wide range of methods and techniques developed in order to achieve a rapid research impact 'hit' without bearing out the consequences of their outputs in terms of both short- and long-term applications of their often dedicated bioanalytical chemistry and multivariate (MV) data analyses work. Indeed, particularly clear is the knowledge that, despite the polynomially-increasing number of publications available in this research area, there appear to be very few which actually manifest themselves into relatively simple diagnostic tools or probes for the diagnosis of the diseases which they were originally designed to investigate and perhaps also monitor. Part of this problem arises from the apparent inabilities of researchers to transform their findings into a clinically or diagnostically significant context (and/or the professional and financial constraints associated with this process), and there remains the potential hazard that, if taken out of context, such results may serve to provide some confusion and perhaps even misinformation. A further component (if you'll excuse the poor choice of words!) is derived from the high costs of performing such multicomponent analysis and the associated valid metabolomic/statistical interpretation of datasets acquired therefrom. Moreover, an additional major barrier is provided by the severe lack of

---

Issues in Toxicology No. 21

Metabolic Profiling: Disease and Xenobiotics

Edited by Martin Grootveld

© The Royal Society of Chemistry 2015

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

statistical validation and cross-validation techniques employed by such researchers in order to evaluate the reliabilities and reproducibilities of the methods that they have developed, *i.e.* so that they may provide a sound foundational basis for the results acquired in their experiments (such concerns are rigorously discussed in Chapters 1 and 2). However, not seeing these connections directly is not the same as not realising that they might be there!

Of much critical importance to the performance of many multivariate (MV) analyses of high-dimensional, high-throughput datasets is the satisfaction of, in many cases, essential assumptions for the effective operation of such models, and in both Chapters 1 and 2 the authors provide relevant information regarding these requirements, and also demonstrate their clear violation when an experimental dataset is subjected to a series of statistical tests for their satisfaction (including those concerning assumptions for normality, homoscedasticity and also the detection of statistical outliers, albeit in a univariate context), observations which are consistent with the very few of those made available by other researchers. In this manner, researchers should always question the validity of many MV analysis techniques which are applicable to such datasets. This problem is absolutely rampant in published work available in which the researchers involved have only employed univariate analysis methods such as t-tests, or one- or two-classification ANOVA (*i.e.* completely randomised or randomised block designs, respectively, for the latter), for example their almost complete lack of consideration for the intra-sample variance homogeneity (homoscedasticity) assumption when testing for significant differences ‘Between-Classifications’, and which relatively simple log- or square root-transformations of the dataset would, at least in some cases, cure. Hence, we can imagine the many problems to be encountered by workers challenged by multidimensional ‘omics’ research problems in this manner!

In Chapter 3 I also review and provide examples of the applications of additional MV analysis techniques which are already available, but nevertheless to date have only been applied to the metabolomic profiling, metabolomics and/or genomics areas in a limited (or very limited) manner. These include canonical correlation analysis (CCorA), and both the k-means and agglomerative hierarchal (AHC) clustering techniques, which have been previously extensively employed in alternative research areas such as ecology and environmental science. Such applications serve as an adjunct to the methods commonly employed in our field of interest. Although these methodologies are not proposed to serve as the first choice of MV analysis for such multidimensional datasets, they can, however, represent valuable strategies or aids for application in particular ‘omics’ investigations or circumstances, for example the use of the CCorA and canonical correspondence analysis (CCA) techniques in order to explore and evaluate any significant linkages, and also the level of dimensionality, between two separate dataset tables (or, for that matter, components or factors derived therefrom, one of which may represent biofluid or tissue biopsy metabolite



levels monitored with one technique, the other perhaps a series of latent, potentially related variables such as age, gender, family status, body mass index, blood pressure components, *etc.*).

Also noteworthy is the essential knowledge that many frequently employed or employable MV analysis techniques available are critically dependent on simple linear (Pearson) correlations between the 'predictor' (X) variables acquired in such model systems; however, in view of this, such models are fraught with many difficulties in view of (1) many potential non-linear, polynomial or otherwise, relationships existing between such variables (many metabolic pathway analyses involved or implicated are either clearly or conceivably of a 'non-linear' nature), and (2) corrections for the influence of further cross-correlated variables (a problem which is resolvable *via* the computation of *partial* correlation coefficients where only a small number, say 2–5, of variables are involved in simple multiple linear regression, partial correlation and discriminatory analysis models), which may exert a major influence on a critical dependent (Y) variable, binary, ordinate, continuous or otherwise. Fortunately, recent developments in the metabolomics research area have served to provide at least some viable means of overcoming these problems, specifically the independent component analysis (ICA) and Gaussian Graphical Models (GGMs) approaches (the former making allowances for potential polynomial relationships between such putative predictor variables, the latter targeted at the consideration of the most important *partial* correlations between them).

A further very important aspect of such investigations involves the consideration of potentially a multitude of interactions between variables involved in the statistical processing of MV bioanalytical datasets (such as those encountered in factorial ANOVA experimental designs), and although this is possible for relatively small numbers of *lateral* variables such as those noted above (including clinically relevant indices, where appropriate), it remains an overwhelming challenge to deal with those arising in MV datasets consisting of hundreds or even thousands of potential predictor variables! For current considerations, however, I and my co-authors merely focus on the applications of techniques (and related examples) which combat and effectively deal with the former (much simpler) task, *i.e.* those concerning the applications of the ANOVA-Simultaneous-Component-Analysis (ASCA) method (which permit exploration of ANOVA-derived orthogonal effect matrices for underlying intra-metabolite relationships and correlations), which is described in my own Chapter 3, and, in a more problem-targeted context, in Chapter 4 by Westerhuis *et al.*, the latter also involving Multi-Level Partial Least Squares-Discriminatory Analysis (ML-PLS-DA). Indeed, in Chapter 4, the authors provide valuable information regarding the development and application of this novel technique, in particular its employment for the solution of two challenging time-series metabolomics tasks, the first investigating the differential treatments applied to a plant species, the second a polyphenolic interventional study in human participants.

Since many of the complete variances of datasets acquired in frequently conducted metabolomics investigations are accountable by variations in sample-donor identities, the time-points at which samples are collected, and also a possible range of further (albeit lateral 'independent' X variables), this relatively recent advance into the metabolomics research area serves to effectively circumvent the confounding effects of such interfering variables, and hence permit researchers to focus on the significance of the main factor(s) of interest following their removal, specifically those observed 'Between-Disease or -Treatment Classifications' as appropriate. A range of researchers have focused on isolating and determining the significance of a range of variance components in complex factorial experimental designs for very many years (although perhaps only in a univariate context), and hence it is a little surprising that metabolomics researchers in general have only recently got round to the idea that it would be highly advantageous also to perform this procedure in a corresponding MV model manner!

Professor Dziuda's contribution in Chapter 5 reveals and outlines metabolomics methods available for the analysis of datasets which have larger numbers of potential predictor (X) variables than there are samples available for analysis. This consideration is of critical importance to the great majority of scientists involved in the metabolomics and further 'omics' research areas, especially those which, in view of advice provided to them (or alternatively their viewpoint), are generally limited to the applications of conventional MV analytical techniques such as PCA or PLS-DA, which are clearly restricted or limited in the context of their applications to such ( $n < P$  or  $n \ll P$ ) datasets, especially the latter method!

This contributor also discusses the application of some commonly employed and well-established data-mining methods to such cases, and also rises to this challenge in his outline and critical appraisal of some new techniques targeted at overcoming this  $P \gg n$  problem encountered in many metabolomics investigations. Primarily, this author focuses on the methods and approaches which are appropriate for the analysis of high-throughput, multidimensional 'omics' datasets, and also provides much useful information regarding some common misconceptions and pitfalls in this area. He also provides guidance concerning when exactly to employ such methods. One major point of interest and importance arising from this work is the rather severe lack of considerations for biomolecular feature selection available in the current literature. Indeed, as he states, this is, after all, the most important aspect of biomarker discovery! He then further delineates the critical importance of presenting new frontiers regarding the sensible MV statistical analysis of such complex and challenging datasets, specifically those involving selected supervised 'learning' algorithms which, when coupled to powerful feature selection methods, can serve to provide a wealth of information regarding MV biomarker identification processes. This chapter also focuses on the extreme importance of considerations for the biological interpretation and significance of the biomarkers selected (together with the critical requirement for their correct validation), plus a



novel data-mining technique that permits their efficient, robust, parsimonious and biologically and/or clinically interpretable discovery.

These points are also critically considered in my own Chapters 1–3, the third of which provides full details and an application example of Dr Magidson's recently developed Correlated Component Regression (CCR) technique, which can be applied to such  $n \ll P$  datasets. As noted above, a further critically important reason for necessarily implementing the application (and hopefully routine future usage) of such forms of data analysis *via* the now commonly employed  $^1\text{H}$  NMR or LC-MS techniques, for example, is the high cost of performing such investigations. Indeed, for the purposes of one grant application which I recently submitted in conjunction with clinical colleagues, the rate for the collection of blood plasma samples for one particular clinical study performed at a single UK Health Service provider was approximately £200 per collection, and this without the additional costings required for the essential provision of associated high-resolution  $^1\text{H}$  NMR analysis and subsequent MV explorations of the datasets acquired!

Chapter 6 by Dr Rick Dunn and co-workers outlines the diverse applications of differing mass spectrometric platforms to the biological and metabolomics research areas, and here the authors focus on the series of advantages offered by these systems, particularly those concerning their specificities, sensitivities and the established potentials and applications of these techniques for the multicomponent analysis of biofluids and tissues (linked with the capacity to classify the identities of thousands of metabolites present in a single sample). The applications of such methodologies will undoubtedly continue to expand, and may also give rise to novel discoveries relating to human health and diseases, together with the subsequent potential development of novel and challenging therapeutic interventional strategies.

Recent developments regarding the applications of data classification - algorithms, firstly those involving unsupervised PCA and cluster analysis techniques, and secondly supervised methods such as Linear Discriminant Analysis (LDA), PLS-DA, Soft Independent Modelling of Class Analogy (SIMCA), Artificial Neural Network (ANNs), SVM machine-learning and Bayesian classification systems to the detection and characterisation of the 'biomarker' roles of metabolites in both soft and hard tissues, together with biofluids collected from humans, are outlined by Kenichi Yoshida and myself in Chapter 7. Indeed, Professor Yoshida's investigations have revealed much valuable metabolic information regarding the ability of these MV analysis techniques to distinguish between healthy and cancerous tissues collected from humans. The application of ongoing technologies for the detection and identification of biomarker patterns which are distinctive for various tumours are also discussed, as is the requirement for the performance of multiple experiments for these purposes.

In Chapter 8, Professor Adamec introduces and discusses the applications of Group-Specific Internal Standard Technology (GSIST) as a newly

developed, novel and highly sensitive LC-MS method that permits the analysis of biomolecules at sensitivities required for the life science research areas. Indeed, novel derivatisation reagents and methods serve to provided major benefits regarding the LC-ESI-MS analysis of metabolites, specifically those involving enhancements of detection sensitivity, attenuations of the hydrophobicities/hydrophilicities of analytes, and their retention times, and chromatographic band-spreading patterns (processes which increase the resolution and rapidity of the separation techniques involved), and also an increased efficacy of both comparative recovery and quantification processes, the latter including the employment of isotopic adducts of selected derivatisation reagents.

Uniquely, Professor Dzeja and colleagues of the Mayo Clinic (USA) outline the value of applying stable isotope  $^{18}\text{O}$ -assisted  $^{31}\text{P}$  NMR and mass spectrometric analyses in order to permit the simultaneous monitoring of high-energy phosphate metabolite levels and their rates of turnover in blood and tissue specimens (Chapter 9). This novel technological breakthrough has given rise to the synchronous monitoring of both ATP synthesis and its utilisation, in addition to the detection of phosphotransfer fluxes involved in the glycolytic, and adenylate and creatinine kinase pathways. Moreover, the status of mitochondrial nucleotides, which are implicated in the Krebs cycle and its dynamics, together with the glycogen turnover process therein, can also be determined. One major advantage offered by this  $^{18}\text{O}$ -based technology is that it has the ability to monitor virtually all phosphotransfer reactions occurring within cells (including those associated with small pool signalling molecule turnovers), and also the dynamics involved in such energetic signal communications. These investigators therefore provide much valuable information concerning the phosphometabolomic/fluxomic profiling of transgenic human disease models which explore trans-systems metabolic network adaptations, and also the potential detection and monitoring of biomarkers which may be related to the effectiveness of treatments for human diseases and/or drug toxicology.

Chapter 10 by Dr Chris Silwood and myself focuses on the application of both conventional and more recently developed methods for the MV analysis of multianalyte human biofluid datasets, the latter involving the Self-Organising Maps (SOMs, both supervised and unsupervised approaches) technique, and their applications have served to provide useful information concerning the ability of an oral rinse product added *in vitro* to exert an influence on the  $^1\text{H}$  NMR metabolic profile of human saliva. Indeed, these methods readily facilitated the detection of perturbations mediated by the oxidation of critical salivary biomolecule scavengers by the actions of an active oxyhalogen agent in the product tested.

With regard to the toxicology research area, in Chapter 11 Wei Tang and Quiwei Xu provide detailed descriptions of drug-induced liver injury, focusing on the current views and understandings regarding the underlying mechanisms involved in these processes. These investigators also focus on



the applications of metabolomics techniques to the provision of essential biomolecular information regarding the pathogenesis of hepatotoxicity, including the seeking, identification and plausible future applications of significant biomarkers for detection, diagnosis, prevention and clinical control of this condition.

Finally, in Chapter 12 Dr Gomase evaluates the application of chemogenic techniques in order to seek chemical (specifically drug) targets within biosystems, in this case relevant proteins. Such research work can indeed serve as a valuable aid to developments in the areas of gene discovery and presents regulation, cheminformatics and molecular signalling opportunities with respect to the potential authentication of novel therapeutic agents for the treatment of chronic human diseases such as a series of cancers. Indeed, the reliable and effective prediction of interactions between specific proteins and low-molecular-mass molecules represents one of the most important phases in our capacities to elucidate the mechanisms involved in a multitude of biological processes, and may also play a crucial role in the development of future drug-discovery systems, together with its further application to the less hazardous and practical issues associated with stem cell regeneration processes.

I would like to express my sincerest thanks to all the authors who contributed chapters to this book (who unfortunately also had to put up with a number of delays with its preparation and completion). Thanks also go to a number of my research collaborators, including those based on my own university campus, namely Victor Ruiz Rodado, Dr Sundarchandran, Prof. Katherine Huddersman, Drs David Elizondo and Dr Dan Sillence, to mention but some, and those from other universities or elsewhere, in particular Prof. Richard Brereton (formerly of the University of Bristol), Prof. Frances Platt (University of Oxford), Prof. Geoffrey Hawkes (Queen Mary, University of London) and Dr Chris Silwood, some of whom have directly or indirectly contributed towards the generation of this work (*via* the kind provision of biofluid samples for  $^1\text{H}$  NMR analysis and/or clinical/clinical chemistry datasets), and sometimes also with the MV or computational intelligence analysis of datasets generated. I also wish to thank a lot of further staff at Leicester School of Pharmacy for their kind support whilst I was involved in producing this work.

Strangely, this book was written and edited, at various stages, in the USA, Brazil, Argentina, Paraguay, Crete and Spain (and sometimes also Portugal), but most especially in various regions of the UK, including North Wales, Shropshire, Manchester, London, Leicester and next to Loch Lomond in Scotland. I also wish to thank the operators of various train, plane and automobile rides which also offered ample opportunities for me to work on the manuscripts, the Black Bear pub in Whitchurch and also the (not so) Happy Friar and Fat Cat bars in Leicester, in which the bar staff did not complain too much about me writing in their 'hospitable' environments. Finally, I also thank my fantastic wife Kerry for all the help and support she

provided whilst I was working on this task (amongst many others): she really had to put up with quite a lot of difficult days involved, at least some of which were unavoidable. I also sincerely thank her for typing my many scribbled revisions to this work, and also for providing invaluable suggestions for improved ones! I hope that this book will serve as a valuable aid to both scientific and clinical researchers who wish to explore such spheres of the unknown!

# Contents

<b>Chapter 1</b>	<b>Introduction to the Applications of Chemometric Techniques in ‘Omics’ Research: Common Pitfalls, Misconceptions and ‘Rights and Wrongs’</b>	<b>1</b>
	<i>Martin Grootveld</i>	
1.1	Introduction	1
1.2	Principal Component Analysis (PCA)	2
1.2.1	Critical Assumptions Underlying PCA	4
1.2.2	Number and Significance of Explanatory Variables Loading on a PC	9
1.2.3	Number of Extractable PCs and Their Characteristics	9
1.2.4	Total Variance of the Dataset	10
1.2.5	What is an Adequate Sample Size for PCA and Further Forms of MV Analysis?	10
1.2.6	Interpretability Criteria of PCs	11
1.2.7	Varimax Rotation	12
1.2.8	Example Case Study	13
1.2.9	Examination of a Wider Range of Components	15
1.2.10	Consideration of Type I (False-Positive) Errors	16
1.2.11	Determinations of the Suitability of MV Datasets for Analysis with PCA and FA	17
1.3	Partial Least Squares-Discriminatory Analysis (PLS-DA)	18
1.3.1	Case Study Describing an Example of PLS-DA ‘Overfitting’	20
1.3.2	Permutation Testing	22

---

Issues in Toxicology No. 21

Metabolic Profiling: Disease and Xenobiotics

Edited by Martin Grootveld

© The Royal Society of Chemistry 2015

Published by the Royal Society of Chemistry, [www.rsc.org](http://www.rsc.org)

1.3.3	Procedures for the Validation and Cross-Validation of PLS-DA Models	24
1.3.4	Attainment of the Final Calibration Model	28
1.3.5	Quality Evaluation Processes	28
1.3.6	Cost-Benefit Analysis (CBA)	30
Appendix I		31
Appendix II		33
Acknowledgements		33
References		33

## **Chapter 2 Experimental Design: Sample Collection, Sample Size, Power Calculations, Essential Assumptions and Univariate Approaches to Metabolomics Analysis** 35

*Martin Grootveld and Victor Ruiz Rodado*

2.1	Introduction	35
2.2	Essential Considerations for Sample Collection	36
2.3	Raw Data Preprocessing Steps	39
2.4	Data Normalisation, Scaling and Dimensionality Reduction	42
2.5	Assumption of Normality	44
2.6	Analysis-of-Variance (ANOVA): Experimental Design and Analysis	50
2.6.1	Model I: Fixed Effects	50
2.6.2	Model II: Random Effects	53
2.6.3	Hierarchical or 'Nested' Models	54
2.6.4	Factorial/Multifactorial Models	54
2.6.5	ANOVA-Simultaneous Component Analysis	57
2.6.6	Further Considerations of Interaction Components of Variance in MV Modeling	57
2.7	Outline of the Applications of Univariate Approaches to the Analysis of Metabolomics Datasets	58
2.7.1	More on Essential Assumptions Required	60
2.7.2	Bonferroni Correction for Multiple Comparisons of Mean Values	62
2.8	Power (Sample Size) Computations for Untargeted, Univariate Investigations of Metabolomics Datasets	64
2.9	Sample Size Requirements and Statistical Power Computations for High-Dimensional, Metabolomic Datasets	67
2.10	Error Analysis	69
	Acknowledgements	69
	References	70



**Chapter 3**

**Recent Developments in Exploratory Data Analysis and Pattern Recognition Techniques**

*Martin Grootveld*

74

3.1	Introduction	74
3.2	Canonical Correlation Analysis (CCorA)	75
3.2.1	CCorA Case Study	76
3.3	Classification and Regression Tree (CART) Analysis	80
3.4	Moderated t-Statistic Methods	81
3.4.1	Significance Analysis of Microarrays (SAM)	81
3.4.2	Empirical Bayesian Approach Modelling (EBAM)	82
3.5	Machine Learning Techniques	83
3.5.1	Self-Organising Maps (SOMs)	83
3.5.2	Support Vector Machines (SVMs)	85
3.5.3	Random Forests (RFs)	86
3.6	Cluster Analysis	86
3.6.1	Agglomerative Hierarchical Clustering (AHC) Methods	89
3.6.2	Clustering Analysis Case Study	91
3.7	Novel Approaches to the Analysis of High-throughput Metabolomics Datasets	92
3.7.1	Genetic Algorithms	95
3.7.2	Gaussian Graphical Models	96
3.7.3	Independent Component Analysis (ICA)	98
3.8	Multidimensional Data ( $P > n$ ) Problems Encountered in MV Regression Modelling	101
3.8.1	Regression Regularisation	102
3.8.2	Model Tuning and Optimisation <i>via</i> an M-Fold Cross-Validation Process	102
3.8.3	Principal Component Regression (PCR)	103
3.8.4	Partial Least Squares Regression (PLS-R)	104
3.8.5	Correlated Component Regression (CCR)	104
	References	110

**Chapter 4**

**Analysis of High-dimensional Data from Designed Metabolomics Studies**

*Johan A. Westerhuis, Ewoud J. J. van Velzen, Jeroen J. Jansen, Huub C. J. Hoefsloot and Age K. Smilde*

117

4.1	Introduction	117
4.2	Case Study 1: The Effect of Jasmonic Acid on the Production of Glucosinolates in <i>Brassicaceae oleracea</i>	119
4.2.1	The ANOVA Model	121