

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Metabolomics Analysis by Liquid Chromatography coupled to High
Resolution Mass Spectrometry:

Optimizing data processing for untargeted workflows and targeted
analysis of carotenoids in algal samples

DISSERTATION

PRESENTED IN PARTIAL FULFILLMENT

OF MASTER OF SCIENCE IN CHEMISTRY

BY

ATEFEH RAFIEI

June 2015

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

Analyse Métabolomique par Chromatographie en Phase Liquide couplée
à la Spectrométrie de Masse à Haute Résolution:

Optimisation de traitement des données pour les
analyses métabolomique non ciblées et l'analyse ciblée des caroténoïdes
dans les échantillons d'algues

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN CHIMIE

PAR
ATEFEH RAFIEI

Juin 2015

ACKNOWLEDGMENT

Here, I sincerely thank Dr. Lekha Sleno for her supervision, invaluable guidance and encouragement, which have been inspirational in the successful completion of this thesis. I am thankful for her support in developing the practical ideas in each step of this research.

It was my pleasure to be a member of the Bio-analytical Mass Spectrometry laboratory, and I am sincerely thankful to the members of the lab for their help and assistance in handling the lab affairs during my M.Sc. program. My very special thanks to Leanne Ohlund, Andre LeBlanc, Makan Golizeh, Dr. Michel Wagner and Biao Ji for their outstanding collaboration in developing new ideas and techniques and for their tremendous discussion and support.

I would like to thank the financial support offered by Université du Québec à Montréal (UQAM), in the form of different awards and fellowships during the M.Sc. program. I am thankful to all the support staff at UQAM specially Ms Sonia Lachance from the chemistry department for steering me through the logistics of my M.Sc. program. I am also thankful to Dr Laura Pirastru for supplying algal samples for study of carotenoids presented in chapter 4. I would also like to thank Dr René Roy and Tze Chieh Shiao for their assistance in providing the synthetic compounds used for the carotenoid study.

Above all, I would like to express my deep gratitude to my parents Mr. Hassan Rafiei and Mrs. Tahereh Mohamadi and my spouse Mr. Amir Sanati-Nezhadi for their encouragement and support during my M.Sc. and my whole life.

DEDICATION

*Dedicated to my beloved parents,
and to my beloved spouse, Amir.*

TABLE OF CONTENTS

LIST OF FIGURES	viii
LIST OF TABLES	viii
LIST OF ABBREVIATIONS AND ACRONYMS	xv
RÉSUMÉ	xvii
ABSTRACT	xviii
CHAPTER I	
INTRODUCTION TO METABOLOMICS	1
1.1 Metabolomics definitions	1
1.1.1 Metabolome	3
1.1.2 Metabolomics and other omics	4
1.2 Different types of metabolomics approaches	5
1.3 Metabolomics platforms	6
1.3.1 NMR spectroscopy	7
1.3.2 FT-IR spectroscopy	7
1.3.3 Mass spectrometry	8
1.4 HPLC-MS based metabolomics workflow	9
1.4.1 Sample preparation	9
1.4.2 HPLC-MS analysis	10
1.4.3 Data processing	15
1.4.4 Metabolite identification	23
1.5 Research objectives	25

1.6 Thesis outline	26
CHAPTER II	
COMPARISON OF PEAK PICKING WORKFLOWS FOR UNTARGETED LC-HRMS METABOLOMICS DATA ANALYSIS	28
2.1 Abstract	28
2.2 Introduction	29
2.3 Experimental	31
2.3.1 Materials	31
2.3.2 Sample preparation	32
2.3.3 HPLC-MS analysis	33
2.3.4 Data Processing	33
2.4 Results and discussion.....	37
2.4.1 Standard mixture.....	37
2.4.2 Biological samples (bile and urine)	39
2.5 Conclusions	48
Supplementary data for chapter 2	49
CHAPTER III	
ENHANCING UNTARGETED METABOLOMIC DATA ANALYSIS BY A NOVEL DATA REDUCTION WORKFLOW	63
3.1 Abstract	63
3.2 Introduction	63
3.3 Material and Methods.....	66
3.3.1 MATLAB processing	66
3.4 Results and discussion.....	68

3.5 Conclusion.....	74
Supplementary data for chapter 3	75
DataReduction Matlab script	75

CHAPTER IV

CAROTENOID QUANTITATION IN ALGAL SAMPLES BY LIQUID CHROMATOGRAPHY-HIGH RESOLUTION MASS SPECTROMETRY	84
--	----

4.1 Abstract	84
4.2 Introduction	84
4.3 Experimental	88
4.3.1 Materials	88
4.3.2 Algal samples	88
4.3.3 Carotenoid extraction and sample preparation	889
4.3.4 Standard mixture.....	889
4.3.5 HPLC-UV-MS analysis	900
4.3.6 Data Processing	911
4.4 Method development.....	922
4.5 Results	100
4.6 Conclusions	1088

CHAPTER V

CONCLUSION.....	110
REFERENCES	112

LIST OF FIGURES

Figure 1.1 Genomics, transcriptomics, proteomics and metabolomics study genome (DNA), transcriptome (RNA), proteome (proteins) and metabolome (metabolite) content of biological samples.....	2
Figure 1.2 A simple schematic of HPLC components.....	11
Figure 1.3 A simple diagram representing the main parts of mass spectrometer (the source is not under vacuum for LC-MS systems, since these use atmospheric pressure ionization techniques)	12
Figure 1.4 Schematic representation of an electrospray ionization (ESI) source. Reprinted from (Hoffmann and Stroobant 2007) with permission	13
Figure 1.5 Schematic of a QqTOF hybrid instrument. Reprinted from (Hoffmann and Stroobant 2007) with permission	15
Figure 1.6 Peak detection process used to extract bounded information of mass signal (m/z), retention time (RT) and intensity of detected ions	18
Figure 2.1 Method used in this study to compare four peak picking workflows. Raw LC-MS data were processed with one of the following software: MetabolitePilot, MarkerView, PeakView and XCMS online. The overlaps between results were then found using an in-house “Venn-pro” MATLAB script followed by METLIN online metabolomics database searching	36
Figure 2.2 The overlaps between detected standard metabolites (confirmed by MS/MS spectral matching) using four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV) and XCMS online) for a standard mixture of 84 compounds in positive and negative ionization modes.....	38
Figure 2.3 Total ion chromatogram (TIC) for bile in positive and negative modes (A and B respectively) and urine in positive and negative modes (C and D respectively). For added clarity, TICs from bile samples were scaled down 3-fold from 18-22 minutes in the above chromatograms	40
Figure 2.4 Number of peaks found by different software: PeakView (PV), MarkerView (MV), MetabolitePilot (MP) and XCMS online for each sample type (bile, urine and standard mixture) in both positive and negative modes	42
Figure 2.5 Venn diagram representation of the average percent of overlaps between the results of four peak picking workflows; MetabolitePilot (MP), MarkerView (MV), PeakView (PV) and XCMS online	42
Figure 2.6 Average % of peaks with at least one hit in METLIN database (metabolites) or at least one hit with MSMS spectra (MSMS metabolites) for peaks in	

each region of Venn diagrams (average includes data from bile and urine in both positive and negative modes). Acronyms shown indicate MetabolitePilot (MP), MarkerView (MV), PeakView(PV) and XCMS online (XC)	44
Figure S1 Venn diagram representation of the overlaps between the results from four peak picking software (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online) used to filter LC-MS data from bile and urine in positive and negative modes	51
Figure S2 Percentage of metabolome database matching for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from bile and urine sample in positive and negative modes, peaks with at least one hit in METLIN database are shown as metabolites and those with at least one MSMS spectra (to be inspected in Metlin) are presented as MS/MS met.....	52
Figure S3 Venn diagram representation of the overlaps between the results of MS/MS spectrum match (METLIN score>60) from four peak picking software MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online on raw LC-MS data from bile and urine in positive and negative modes.....	53
Figure S4 The average percent of peaks found by different number of workflows in each region of chromatogram	62
Figure 3.1 " <i>DataReduction</i> " MATLAB script was used in this study to identify and remove isotope peaks, radical ions, adducts and in-source fragments	67
Figure 3.2 Number of peaks found by MATLAB data reduction script for different workflows: PeakView (PV), MarkerView (MV), MetabolitePilot (MP) and XCMS online for each sample type (bile, urine and compound mixture) in both positive and negative modes.....	69
Figure 3.3 Venn diagram representation of the results of comparison between ¹³ C isotope peaks found by MarkerView (MV) and developed <i>DataReduction</i> (DR) MATLAB script.....	70
Figure 3.4 Pie chart representation of the percentage of redundant peaks found by <i>DataReduction</i> MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from bile sample in positive and negative modes.....	71
Figure 3.5 Pie chart representation of the percentage of redundant peaks found by <i>DataReduction</i> MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from urine sample in positive and negative modes.....	72

Figure 3.6 Pie chart representation of the percentage of redundant peaks found by <i>DataReduction</i> MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XCMS)) from the standard mixture (Std mix) sample in positive and negative modes	73
Figure S5 Sample excel sheet containing peak detection information to be imported to MATLAB <i>DataReduction</i> workflow	83
Figure 4.1 Four carotenoid compounds studied in this work.....	87
Figure 4.2 Sample preparation used for carotenoid quantification in algal samples. It starts with filtration of the algal culture followed by bead beating extraction of carotenoids. Less solvent consumption and faster sample preparation was achieved with this optimized extraction method.....	90
Figure 4.3 Gradient elution used for HPLC-UV-MS analysis. Mobile phase A was (90% MeOH/10% H ₂ O/0.1% FA) and B was (85% MTBE/15% MeOH/0.1% FA) .	91
Figure 4.4 UV trace at 450 nm (top) and extracted ion chromatograms (bottom) of standards used in this study: 1) echinenone, 2) astaxanthin, 3) lutein, 4) canthaxanthin, 5) β -apo-8'-carotenal (IS), 5') methylated β -apo- β '-carotenal, 6) echinenone, 7) β -carotene (Meier 2012, unpublished data).....	93
Figure 4.5 Candidate internal standard compounds tested for this study including α -Tocopherol (vitamin E), Menaquinone (vitamin K2) and β -apo-8'-carotenol. β -apo-8'-carotenol was the internal standard previously used for quantification of carotenoids in algal samples	94
Figure 4.6 Evaluation of purity and coelution of β -apo-8'-carotenol (50 μ g/ml) to be used as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 8.3 to 8.6 min), C) Mass spectrum (from 9.2 to 9.3 min)	95
Figure 4.7 Evaluation of purity and co-elution of vitamin E (50 μ g/ml) tested as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 10.8 to 11.0 min)	96
Figure 4.8 Evaluation of purity and coelution of vitamin K2 (50 μ g/ml) to be used as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 10.2 to 10.4 min), C) Extracted ion chromatograms (XICs) which shows perfect co-elution of two observed ions (protonated molecule and ammonium adduct)	97
Figure 4.9 Peak areas detected for internal standard when it was added (to the sample in different stages of sample preparation, during filtration, during extraction and right before the injection to HPLC-HRMS system (n=9)	98
Figure 4.10 Evaluating the influence of sample preparation on peak area for four standard compounds, A) astaxanthin, B) lutein, C) β -Carotene, D) canthaxanthin.	

Three different sets of standard mixtures were prepared starting with i) filtration ii) extraction or iii) serial dilution of stock solution of compounds to obtain desired concentration (without filtration and extraction). 100

Figure 4.11 Extracted ion chromatograms of standards used in this study: 1-astaxanthin, 2-lutein, 3-canthaxanthin, 4-Vitamin K2 (IS), 5- β -carotene. Extracted m/z are also presented for each peak. 101

Figure 4.12 The change in carotenoid concentration in *Haematococcus* algae induced by stress condition (culturing time). The quantity of beta-carotene and lutein exist in *Haematococcus* green are higher than the limit of detection of our method for these compounds. The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration ($\mu\text{g}/\text{cell}$) and standard deviation are also shown in the figure ($n=3$) . 103

Figure 4.13 Change in carotenoid concentration in *Muriellopsis* algae induced by stress condition (culturing time). The quantity of cantaxanthin exist in Muriellopsis orange and also β -carotene exist in *Mureillopsis* green are higher than the limit of detection The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration ($\mu\text{g}/\text{cell}$) and standard deviation are also shown in the figure ($n=3$) . 104

Figure 4.14 The change in carotenoid concentration in *Oocystis* algae induced by stress condition (culturing time). The quantity of β -carotene, cantaxanthin and lutein in *Oocystis* orange, and also β -carotene and lutein in *Oocystis* green is higher than the limit of detection The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration ($\mu\text{g}/\text{cell}$) and standard deviation are also shown in the figure ($n=3$)..... 105

Figure 4.15 Carotenogenesis pathway. It was observed that in *Haematococcus* and *Muriellopsis* algae species, astaxanthin and cantaxanthin were up-regulated (tick up arrows) and decreases of lutein and β -carotene levels were revealed under stress treatment (tick down arrows) (adapted representation based on (Álvarez *et al.* 2006)) 106

Figure 4.16 Calibration curve obtained for Astaxanthin using linear least squares regression analysis. It covers 0.2 to 10 $\mu\text{g}/\text{ml}$ of Astaxanthin in standard solutions (IS concentration was 1 $\mu\text{g}/\text{ml}$)..... 106

Figure 4.17 Calibration curve obtained for β -Carotene using linear least squares regression analysis. It covers 0.02 to 0.2 $\mu\text{g}/\text{ml}$ of β -Carotene in standard solutions (IS concentration of 1 $\mu\text{g}/\text{ml}$)..... 107

Figure 4.18 Calibration curve obtained for Canthaxanthin using linear least squares regression analysis. It covers 0.05 to 2 $\mu\text{g/ml}$ of Canthaxanthin in standard solutions (IS concentration was 1 $\mu\text{g/ml}$)..... 107

Figure 4.19 Calibration curve obtained for Lutein using linear least squares regression analysis. It covers 0.2 to 10 $\mu\text{g/ml}$ of Lutein in standard solutions (IS concentration was 1 $\mu\text{g/ml}$)..... 108

LIST OF TABLES

Table S1 Compounds used in standard mixture for evaluating the four peak picking workflows	49
Table S2 Metabolites found by each peak picking workflow from standard mixture in positive ionization mode (confirmed with MS/MS match from METLIN database). 54	
Table S3 Metabolites found by each peak picking workflow from standard mixture in negative ionization mode (confirmed with MS/MS match from METLIN database) 55	
Table S4 Metabolites identified by targeted approach in standard mixture (pos) with METLIN MS/MS matching which had not been detected by any of the automated peak detection workflows	56
Table S5 Metabolites identified by targeted approach in standard mixture (neg) with METLIN MS/MS matching which had not been detected by any of the automated peak detection workflows	56
Table S6 Identified metabolites in urine (pos) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites, all of which are presented here). Metabolites found in standard mixture with matching retention times are assigned with a star	57
Table S7 Identified metabolites in urine (neg) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here). Metabolites found in standard mixture with matching retention times are assigned with a star	59
Table S8 Identified metabolites in bile (pos) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here). Matched metabolite with standard mixture is assigned with a star	60
Table S9 Identified metabolites in bile (neg) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here)	61
Table 3.1 Comparison of the total number of ¹³ C isotope peaks and the overlap between two filtering algorithms: ("DataReduction" and MarkerView) in bile and urine sample in both positive and negative modes	70

Table 4.1 Accurate masses and retention times of carotenoids observed in algae extracts by HPLC-UV-ESI-MS/MS	102
--	-----

Table 4.2 Summarized changes in carotenoid content from control sample to stressed algae samples. Based on our experimental results (fold change and t-test) <i>Muriellopsis</i> and <i>Haematococcus</i> showed an up-regulation (Up) of astaxanthin and canthaxanthin while β -carotene and lutein were down-regulated (Down). All four studied carotenoids were down-regulated in <i>Oocystis</i> (fold changes are shown in the table and p-values were all below 0.05).....	105
--	-----

LIST OF ABBREVIATIONS AND ACRONYMS

AMP	Adenosine MonoPhosphate
API	Atmospheric Pressure Ionization
CWT	Continuous Wavelet Transform
EI	Electron (Impact) Ionization
ESI	Electrospray Ionization
FA	Formic Acid
FT-IR	Fourier Transform Infrared spectroscopy
FT-ICR	Fourier Transform-Ion Cyclotron Resonance
GC	Gas Chromatography
GMP	Guanosine MonoPhosphate
HPLC	High Performance Liquid Chromatography
HRMS	High Resolution Mass Spectrometry
HMDB	Human Metabolome Database
IS	Internal Standard
IT	Ion Trap
LC	Liquid Chromatography
LC-MS	Liquid Chromatography - Mass Spectrometry
LC-MS/MS	Liquid Chromatography - tandem mass spectrometry
LLE	Liquid-Liquid Extraction
MMD	Manchester Metabolomics Database
MV	MarkerView
MS	Mass Spectrometry
MP	MetabolitePilot
MeOH	Methanol
MTBE	Methyl Tert-Butyl Ether
NMR	Nuclear Magnetic Resonance
PV	PeakView
QqTOF	Quadrupole-Time-Of-Flight mass analyzer
SNR	Signal to Noise Ratio
SPE	Solid Phase Extraction
TOF	Time-Of-Flight mass analyzer
QqQ	Triple quadrupole analyzers

UHPLC	Ultra-High Performance Liquid Chromatography
UV	Ultraviolet
XC	XCMS online

RÉSUMÉ

La métabolomique est une science d'omique récente visant à étudier les changements quantitatifs de métabolites causés par la maladie, les changements environnementaux, *etc.* Aux flux de travail non ciblées, l'acquisition d'une vue globale de tous les métabolites d'un échantillon biologique est souhaitée. En raison de la grande complexité des échantillons biologiques, des données brutes doivent être traitées soigneusement pour arriver aux résultats significatifs. La première étape de l'analyse non ciblée de données métabolomique est de générer des pics de premières données LC-MS. Dû à l'application des plusieurs algorithmes avec les différents flux de travail pour la sélection de pic, les résultats peuvent varier largement les uns des autres. L'autre défi est également à filtrer les pics redondants tels que ^{13}C isotopes, les produits d'addition et des fragments de source provenant de métabolites lors de l'analyse MS. Et malheureusement la plupart logiciels automatisé pour ramasser les pics sont incapables de détecter ces pics redondants. En ce travail, nous avons étudié systématiquement l'effet d'employer les différents flux de travail des pics pour les mêmes ensembles de données brutes. Un mélange standard (84) et composés de deux échantillons biologiques (biliaires et urine) ont été analysés par HPLC-MS-QqTOF aux deux modes positif et négatif. Les données brutes LC-MS ont été traitées avec quatre flux de travail différents pour gérer le pic, y compris Peakview®, Markerview™, MetabolitePilot™ et XCMS Online. Ensuite, les chevauchements entre les résultats des flux de travail pour gérer le pic ont été obtenus pour chaque ensemble de données en appliquant un code basé sur MATLAB. Enfin, les métabolites potentiels identifiables ont été étudiés en utilisant la base de données en ligne METLIN. Dans un autre effort pour améliorer la performance de l'analyse des données non ciblée de la métabolomique, un flux de travail de réduction de données basé sur MATLAB a été développé pour identifier et supprimer les isotopes ^{13}C , ions radicaux, adduits et et-source-fragments. D'un autre projet, une approche métabolomique ciblée a été développée pour quantifier la modification introduite au contenu caroténoïde des échantillons d'algues par le stress.

Mots-clés: métabolomique, la spectrométrie de masse, LC-MS, cueillette Peak, caroténoïdes

ABSTRACT

Metabolomics is a recent omics science aiming to study quantitative changes in metabolites caused by disease, environmental change, *etc.* In untargeted workflows, acquiring a global view of all metabolites present in a biological sample is desired. Due to the high complexity of biological samples, raw data should be processed carefully to yield meaningful results. The first step in untargeted metabolomics data analysis is to generate peaks from raw LC-MS data. Due to the use of various algorithms by different peak picking workflows, results can differ widely from each other. The other challenge is also to filter out redundant peaks such as ^{13}C isotopes, adducts and in-source fragments originating from metabolites during MS analysis and unfortunately, most automated peak picking software are unable to combine all signals belonging to a single metabolite. In this work, we systematically investigated the effect of employing different peak picking workflows for the same raw data sets. A standard mixture (84 compounds) and two biological samples (bile and urine) were analyzed by HPLC-QqTOF-MS in both positive and negative modes. Raw LC-MS data were processed with four different peak generating workflows including *Peakview*®, *Markerview*™, *MetabolitePilot*™ and *XCMS Online*. Then the overlaps between the results of peak generating workflows for each data set were obtained using a custom-built MATLAB-based code. Finally, the potential identifiable metabolites were investigated using the online METLIN database. In another effort for enhancing the performance of untargeted metabolomics data analysis, a MATLAB-based data reduction workflow was developed to identify and remove ^{13}C isotopes, radical ions, adducts and in-source-fragments. In a separate project, a targeted metabolomics approach was developed to quantify the change introduced to carotenoid content of algal samples by stress.

Keywords: Metabolomics, Mass Spectrometry, LC-MS, Peak picking, Carotenoids

CHAPTER I

INTRODUCTION TO METABOLOMICS

1.1 Metabolomics definitions

Analysis of biological samples for identification and quantification of small molecules has been done for many years, *e.g.* measurement of glucose for diabetes (Group 1979) and plasma homocysteine for vascular disease (Elevated *et al.* 1997). These studies were initially limited to small number of target compounds and far from what we currently define as metabolomics. The fact that individual molecules in biological samples are part of a large network of metabolic pathways magnified the need for a more comprehensive and global approach toward analysis of biological specimens (Ryan and Robards 2006).

Metabolomics is defined as the quantitative characterization of small molecules (metabolites) present in a biological sample (Lindon *et al.* 2011). This kind of terminology arises from other “omic” sciences such as genomics, transcriptomics and proteomics, in which genome, transcriptome and proteome content of living organisms are studied, respectively (Figure 1.1). Accordingly, metabolomics refers to the study of metabolome, the word first suggested at 1998 by Stephen Oliver (University of Manchester, UK; <http://www.man.ac.uk/>), assigned to the set of all low-molecular-mass compounds synthesized by an organism (Oliver *et al.* 1998). Soon afterward, a detailed proposal review on this subject was presented to scientific community by Oliver Fiehn (Max Plank Institute, Golm, Germany; <http://www.mpg.de>) (Fiehn 2002). It should be mentioned that one of the first metabolite profiling experiments had been performed long before by Linus Pauling and colleagues in 1971 in which metabolite content of human urine vapor and breath of subjects were analyzed by gas chromatography (Pauling *et al.* 1971),

though the more serious effort for growing this branch of omic sciences was observed only in the last decade.

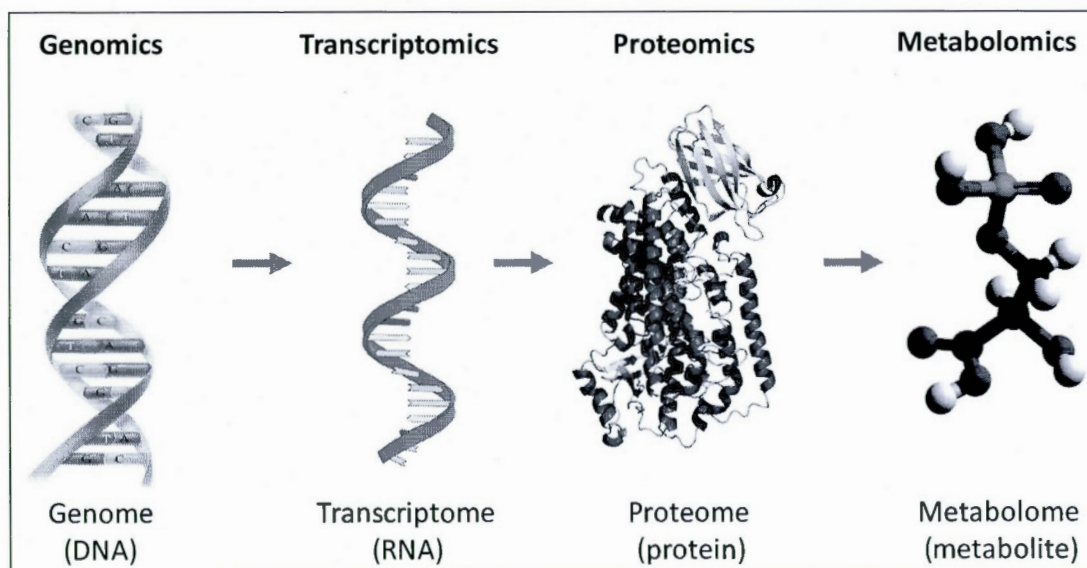


Figure 1.1 Genomics, transcriptomics, proteomics and metabolomics study genome (DNA), transcriptome (RNA), proteome (proteins) and metabolome (metabolite) content of biological samples

Another term used along with metabolomics that creates confusion in the corresponding literature, is metabonomics. Initially, metabolomics referred to the measurement of the pool of cell metabolites (Nicholson *et al.* 1999), while metabonomics was defined as the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification (Nicholson *et al.* 1999; Beger *et al.* 2010). These two terms are often used interchangeably, however both procedure and bulk of literature support metabolomics as more comprehensive study of the metabolome (Ryan and Robards 2006).

Metabolomics investigations are used in different research areas such as drug discovery (Wishart 2008), medical diagnosis and therapeutic monitoring (Gowda *et*

al. 2008), toxicology (Ramirez *et al.* 2013) as well as food science (Wishart 2008), agriculture (Dixon *et al.* 2006) and environmental studies (Ramirez *et al.* 2013).

1.1.1 Metabolome

The metabolome is defined as the set of small molecular mass organic compounds, metabolites, found in a given biological sample. Small peptides are considered as metabolites while polymerized structures such as proteins and DNA are beyond the accepted definition for metabolites. Considering the important biochemical roles of metabolites as intermediates of biochemical reactions, their quantitative level (concentration) in living cells can be affected by different processes such as regulation of transcription and translation or protein-protein interactions. Hence, studying metabolite levels has great potential to inform us about cellular function and its response to various genetic or environmental changes (Roux *et al.* 2011).

Metabolites are generally divided into two groups, based on their origin being either exogenous or endogenous (Roux *et al.* 2011). Endogenous metabolites are either primary, which are common organic molecules found in broad category of living cells, or secondary metabolites, referred to the species-specific compounds. The first group has a direct contribution to essential life processes such as growth and maintenance, *e.g.* molecules such as amino acids or glycolysis intermediates. On the other hand, secondary metabolites have limited distribution among living organisms and metabolites belonging to this group have more specific biological functions, *e.g.* hormones in mammals and alkaloids in plants (Herbert 1989).

Exogenous metabolites are the product of biotransformation of exogenous compounds caused by phase I or phase II metabolism. In phase I, the original exogenous molecule is modified by introducing small polar functional group(s),

while phase II represents the formation of a conjugation product (Shargel *et al.* 2005).

The complexity of metabolomics analysis is due to the diverse chemical properties as well as wide concentration range, estimated to be 7-9 orders of magnitude (pmol - mmol) (Dunn and Ellis 2005). Most importantly, the large number of metabolites makes an analytical approach much more complicated. For instance, estimates include >1000 metabolites present in *Escherichia coli* (Feist *et al.* 2007), >4000 for human serum (Psychogios *et al.* 2011), and between 5000 and 25000 for higher plants (Trethewey 2004).

1.1.2 Metabolomics and other omics

Although the metabolome is a complex system, it is still smaller than the proteome and genome of living cells (Watkins and German 2002). In addition, the change and variation in metabolome is more associated with altered phenotype which affects growth, development and health; while the change in proteome and genome does not always result in biochemical change. Thus, it is believed that metabolomics has the potential to provide the most functional information of all omic science (Sumner *et al.* 2003).

On the other hand, in genomics and proteomics, complete or near complete assessment of related biological content (genome and proteome) is normally achieved, while metabolomics is still far behind them from this aspect (Bouatra *et al.* 2013). For instance, publications on human metabolomics studies by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) instruments, often contains identification of fewer than 100 metabolites (Metz *et al.* 2007; Rousu *et al.* 2009; Lim *et al.* 2010; Zhang *et al.* 2010) representing a tiny fraction (less than 1%) of the human metabolome (Wishart *et al.* 2012). In order to cover a larger

portion of the metabolome, several systematic efforts were made for detailed analysis of human biofluids, such as cerebrospinal (Wishart *et al.* 2008; Mandal *et al.* 2012), saliva (Takeda *et al.* 2009), serum (Psychogios *et al.* 2011), plasma (Lawton *et al.* 2008) and urine (Bouatra *et al.* 2013).

1.2 Different types of metabolomics approaches

Metabolomics investigations are divided to three broad categories, including targeted, fingerprinting and profiling approaches (Ryan and Robards 2006). The last two are also referred to as untargeted metabolomics.

In targeted metabolomics, a pre-defined list of compounds is quantified in samples. For instance, quantification of selected carotenoid compounds in algal samples is a targeted metabolomics approach to study changes as a result of a specific stress condition (Chu *et al.* 2011).

In a fingerprinting approach, a global view of all spectral features is obtained for samples with different biological conditions, (*e.g.* samples from healthy and diseased individuals), followed by applying statistical methods to identify metabolites with significant differences in concentration levels among studied samples. Identifying the biomarker molecules for early detection of breast cancer is defined in this category (Nam *et al.* 2009).

Another type of metabolomics investigation is metabolite profiling, which involves the identification and quantification of predefined set of metabolites of known or unknown identity, related to a metabolomic pathway or a class of compounds (Dettmer and Hammock 2004; Dunn and Ellis 2005), for instance, identification and quantification of all amino acids. Although this approach is the oldest and the most established type of metabolomics (Ryan and Robards 2006), it suffers from

the disadvantage of not being universal or a “real” omic science (Dettmer and Hammock 2004).

Thus, the major difference between targeted and untargeted approaches, is in the stage which identification of metabolites is performed. In a targeted approach, the investigation is on pre-defined metabolites whose identity is known, while only metabolites with significant differences are identified in the final steps of data analysis in an untargeted workflow.

It seems that no common agreement is made in the literature regarding the classification of metabolomics studies yet. Some reviewers exclude metabolomics profiling due to not being universal and or a “real” omic science (Griffiths *et al.* 2010; Preet *et al.* 2012; Varghese *et al.* 2012) and some others ignore targeted approach in metabolomics classifications (Dettmer and Hammock 2004) and refer to it as multi-analyte methods instead (Theodoridis *et al.* 2012).

Based on sample types investigated in metabolomics, footprinting metabolomics is also defined which by definition is "*the measurement of metabolites secreted from the intracellular complement of an organism (or biological system) into its extracellular medium or matrix.*" (Tugizimana *et al.* 2013). This approach is commonly used in microbiology (Mapelli *et al.* 2008), tissue engineering (Seagle *et al.* 2008) and stem cell studies (Turner *et al.* 2008).

1.3 Metabolomics platforms

Since the metabolome consists of a wide variety of metabolites with different physicochemical properties, it is impossible to use one single technique to analyze the entire metabolome content. Several analytical techniques have been used, including three main platforms, proton nuclear magnetic resonance (^1H NMR),

mass spectrometry (MS) and fourier transform infrared (FT-IR) spectroscopy. In addition, chromatographic techniques, such as gas chromatography (GC) and liquid chromatography (LC), are often coupled to mass spectrometry for further separation of compounds present in a complex sample (Varghese *et al.* 2012).

1.3.1 NMR spectroscopy

Nuclear magnetic resonance (NMR) is one of the major analytical tools used in metabolomics studies since 1990 (Lindon *et al.* 2003). The non-destructive and non-discriminative nature of this method, as well as fast and robust analytical performance (Roux *et al.* 2011) make it a suitable platform for metabolomics investigation. Another advantage of NMR is that minimal sample preparation is needed, hence, there is less chance for metabolites to be changed or lost during sample preparation. However, NMR suffers from low sensitivity and only medium to high abundance metabolites will be detected by this technique. In addition, identification of individual metabolites is very challenging in complex mixtures since signals from different metabolites could overlap (Dettmer *et al.* 2007; Lawton *et al.* 2008), however, some efforts were done for developing mathematical platforms for quantification of metabolites using NMR spectroscopy (Weljie *et al.* 2006). Moreover, the sensitivity of NMR to chemical environment (pH, ionic strength, temperature, *etc.*) and differential sensitivity of molecules to such changes is considered as a major downfall for this technique (Weljie *et al.* 2006).

1.3.2 FT-IR spectroscopy

Fourier transform-infrared spectroscopy (FT-IR) is also used for analyzing biological samples for metabolomics, nevertheless the number of publications on this subject is much less than NMR and MS-based metabolomics. It offers advantages such as low cost, simplicity of sample preparation and low sample

volume needed (Harrigan *et al.* 2004), however, this method suffers from lack of reproducibility, as sample preparation could cause changes in the continuous intensity data (Roscini *et al.* 2010). Another disadvantage of this method is that signal interference due to a strong absorbance band of water, makes analysis of aqueous solutions problematic. Although, attenuated total reflectance sampling tools and short path-length transmission cells are employed for analysis of water-based samples, these methods only minimize water signal interference without completely alleviating the problem, therefore this could still cause serious errors in the detection of some metabolites (Botros *et al.* 2008).

1.3.3 Mass spectrometry

Mass spectrometry-based methods are widely used for metabolomics studies, and are often coupled to gas chromatography or liquid chromatography. High sensitivity, accuracy and coverage has made this technique a promising tool for metabolomics investigations (Varghese *et al.* 2012). Analysis of the metabolome with MS-based techniques provides the possibility of identification of individual metabolites (Want *et al.* 2007), the task which is much more complicated in other type of instruments, such as NMR and FT-IR. In addition, the number of MS facilities worldwide in comparison to high field NMR instruments is higher, partly because of it being less expensive instrumentation. Furthermore, many more experts are working in this area compared to specialists operating state-of-the-art NMR facilities (Theodoridis *et al.* 2012).

Gas chromatography was the first chromatographic method to be coupled to MS detection and has been used as far back as 1960's for metabolomics applications (Brooks *et al.* 1968). GC-MS provides high resolution separations and reproducible EI spectra, which facilitates identification of metabolites by database searching of known compound spectra. Meanwhile, this technique is only applicable to

molecules which are volatile and thermostable, or need to be derivatized (Roux *et al.* 2011). The major challenge in the hyphenation of liquid chromatography to mass spectrometry due to the large difference in operating pressures was resolved by the introduction of atmospheric pressure-based ionization methods (API), such as electrospray ionization (ESI) (Plumb *et al.* 2004) and atmospheric pressure chemical ionization (APCI) (Huang *et al.* 1990). In LC-MS, lower temperature is needed in comparison to GC-MS and metabolites don't need to be volatile, hence less sample preparation is usually necessary. In addition to the high dynamic range and sensitivity which are considered as main advantages of LC-MS systems (Roux *et al.* 2011), soft ionization techniques provide information on the intact molecular mass of metabolites (Roux *et al.* 2011), compared to mostly fragment ions seen in EI spectra from most GC-MS systems.

1.4 HPLC-MS based metabolomics workflow

A typical HPLC-MS based metabolomics pipeline typically consist of four steps, sample preparation, HPLC-MS analysis, data processing and metabolite identification. Each step is explained in detail below.

1.4.1 Sample preparation

For metabolomics studies by LC-MS, minimum sample preparation is typically performed in order to prevent unwanted change or removal of metabolites. For most non-pharmaceutical experiments such as plant, microbial or mammalian biomarker research, intracellular extraction and/or protein precipitation is performed followed by dilution in a suitable solvent (Dunn and Ellis 2005). Additional sample preparation could be employed, including solid phase extraction (SPE), liquid-liquid extraction (LLE) or supercritical fluid extraction, and have been used in pharmaceutical applications (Rossi and Sinz 2001; Bakhtiar *et al.* 2002; Bamba *et*

al. 2008). Unlike GC-MS studies, which require derivatization for adding suitable functional groups to molecules to make them more volatile, derivatization is not necessary in LC-MS-based studies, however, it could be used in certain cases to enhance sensitivity and chromatographic resolution (Leavens *et al.* 2002).

1.4.2 HPLC-MS analysis

1.4.2.1 HPLC

High performance liquid chromatography (HPLC) is one of the leading methods used for separation of different compounds present in solution. This method works based on the interaction between a liquid (mobile phase) and a solid or fixed gel (stationary phase). Firstly, a small volume of the sample containing the analytes, is introduced into the mobile phase. Then, the mobile phase is pumped through a chromatographic column filled with small sorbent particles. Based on the type of HPLC method used, different affinities of the compounds being analyzed with mobile and stationary phase cause their separation. For example, for reversed phase chromatography, a polar solvent and non-polar stationary phase are used, resulting in differentiation between different class of molecules present in sample based on their polarity (or hydrophobicity). In this case, more polar compounds elute first, while non-polar compounds are retained more within the stationary phase and will elute later. Mobile phase normally consists of a mixture of solvents, and elution is either isocratic (constant ratio of solvents) or gradient (changing composition over time) for improving separation efficiency. A simple diagram of a HPLC system is presented in Figure 1.2. Various detection methods could be used for identification of separated analytes eluting from the chromatographic column, one of the most sensitive detectors being a mass spectrometer.

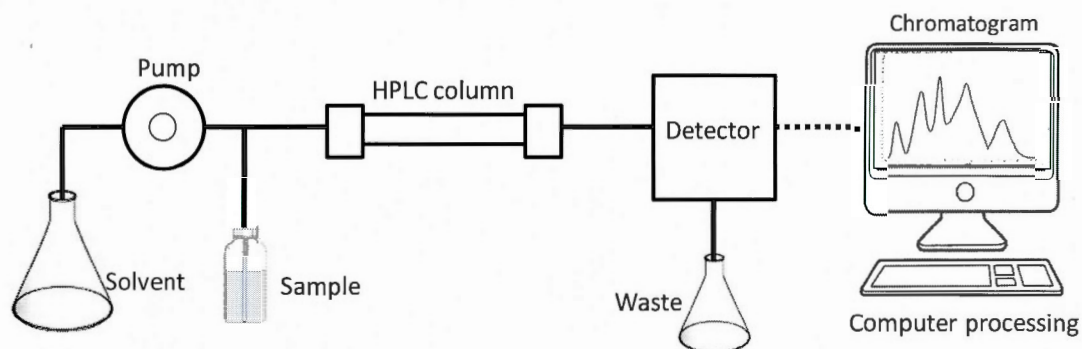


Figure 1.2 A simple schematic of HPLC components

For metabolomics studies, chromatographic separation is typically carried out using columns between in 2–4.6 mm internal diameter (i.d.), with lengths ranging from 5 to 25 cm, and packed with particles from 3–5 μm i.d. (Lindon *et al.* 2011), though smaller particles are being used more and more for increased efficiencies (with corresponding increase in operating pressures). Reversed-phase chromatography is a popular method for metabolomics investigation, however, in order to cover the large diversity of metabolites present in biological samples, other types of chromatography can be used with different stationary phases *e.g.* hydrophilic interaction chromatography (HILIC). Due to the complementary nature of HILIC and reversed phase chromatography, they could be combined in two dimensional applications (Huang *et al.* 1990). It should be mentioned no standard LC-MS method is currently recommended for profiling the complete metabolome due to the chemical diversity of metabolites (Theodoridis *et al.* 2008).

1.4.1.2 Mass spectrometry

Mass spectrometry is a powerful analytical technique widely used for biochemical applications. The main steps of it are ionizing compounds present in a sample, separating the resulting ions based on their mass-to-charge ratio and finally detecting and reporting their abundance. These steps performed by an ionization

source, mass analyzer and detector, respectively. A simple diagram of the mass spectrometer's components is given in Figure 1.3.

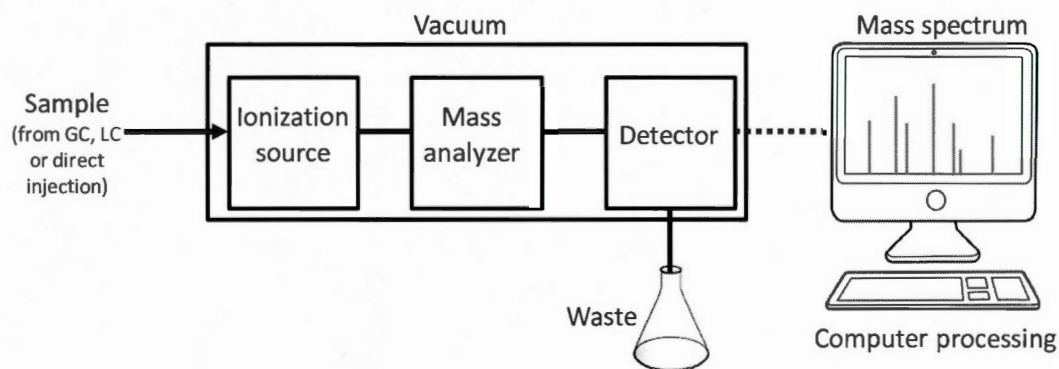


Figure 1.3 A simple diagram representing the main parts of mass spectrometer (the source is not under vacuum for LC-MS systems, since these use atmospheric pressure ionization techniques)

In the ion source, molecules of analytes undergo ionization to produce gas phase ions. Although different ionization sources are available, electrospray ionization (ESI) is the most frequently used method for LC-MS based metabolomics. In this method, the chromatographic eluent passes through a capillary nebulizer tube that is connected to a strong electric field. This field causes charge accumulation at the surface of liquid placed at the end of tube, transferring electrical charge to droplets leaving this tube. Charged droplets then lose the remaining solvent by evaporation using heat and inert gas flow. The desorption of ions from the surface of droplets will occur when the solvent is evaporated and electrical charge is large enough at the surface of tiny droplets, to produce a Coulombic explosion into individual gas-phase ions. Produced ions are then guided, by differential potentials, toward the mass analyzer (Hoffmann and Stroobant 2007). A simple representation of an ESI source is shown in Figure 1.4.

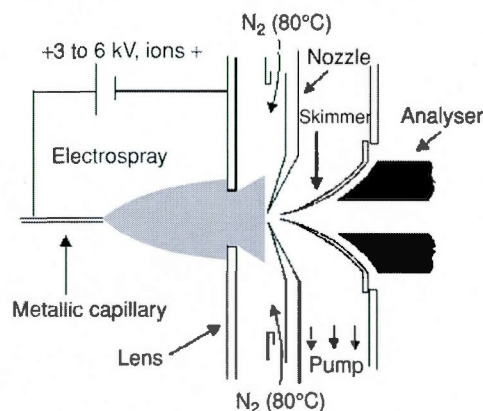


Figure 1.4 Schematic representation of an electrospray ionization (ESI) source. Reprinted from (Hoffmann and Stroobant 2007) with permission

In order to cover a greater portion of metabolites by mass spectrometry, analysis is usually performed in both positive and negative ionization modes. Some metabolites are detected in only one ionization mode (Dunn and Ellis 2005) while some metabolites could be detected in both ionization modes. Ions are created by protonation $(M+H)^+$ (in positive mode) or deprotonation, $(M-H)^-$ (in negative mode) and also possibly adduct ions are formed, as well as in-source fragment ions. Fragmentation is referred to as the dissociation of molecules to smaller parts, and fragments (or product ions) will be detected afterwards. Adducts are the result of the addition of sodium, potassium, ammonium, chloride, acetate or other ionic species to the molecule. There is also the possibility for multiply charged species to be formed as well as clusters (dimers, trimers, *etc.*), all of which can have the effect of adding complexity to the mass spectra.

Produced gas phase ions need to be separated based on mass-to-charge ratio (m/z). A wide variety of analyzers can be used for LC-MS based metabolomics including time-of-flight (TOF), Orbitrap, Fourier transform-ion cyclotron resonance (FT-ICR), ion trap (IT), and triple quadrupole (QqQ) analyzers. In the first three, high resolution and accurate mass measurements are possible. In addition, hybrid

analyzers such as quadrupole-time of flight (QqTOF) systems, provide high resolution measurements as well as the potential for tandem mass spectrometry (MS/MS). Accurate mass measurement and MS/MS analysis help in the structural elucidation of metabolites (Roux *et al.* 2011). Since the MS used in this thesis is a QqTOF system, its main operating concepts will be explained in more detail.

QqTOF systems are in fact similar to triple quadrupole, in which a time-of-flight (TOF) analyzer replaces the “third” quadrupole. In QqTOF analyzers, two quadrupoles are operated in series, namely Q1, q2 followed by a time of flight tube. A q0 quadrupole (or multipole) is often added before the Q1 ion filter to provide collisional cooling and focusing of the ions (Chernushevich *et al.* 2001).. A schematic representation of a typical QqTOF is shown in Figure 1.5

Two main types of experiments can be done with QqTOF analyzers, TOF-MS or MS/MS analysis. In TOF-MS mode, the Q1 works in rf-only mode, meaning that it transmits all ions to the high resolution TOF analyzer to be separated. In the TOF, entering ions are separated based on their velocity, since all ions enter with the same kinetic energy, giving information regarding their m/z . On the other hand, in MS/MS analysis, Q1 will act as an ion filter, only passing a specific precursor ion to the collision cell (q2), where it is fragmented by collisional-induced dissociation (CID). The product (or fragment) ions are then sent to the TOF to obtain product ion separation (Chernushevich *et al.* 2001).

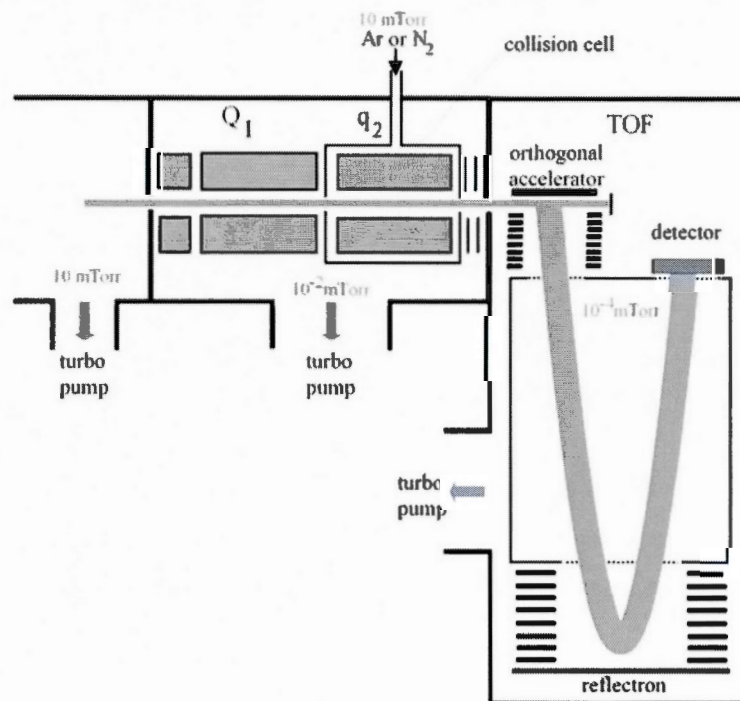


Figure 1.5 Schematic of a QqTOF hybrid instrument. Reprinted from (Hoffmann and Stroobant 2007) with permission

1.4.3 Data processing

Data treatment workflows for MS-based metabolomics consist of several stages of processing including: noise filtering, peak detection, ion annotation, alignment and normalization, and statistical analysis, followed by metabolite identification (Castillo *et al.* 2011).

Noise filtering is used primarily to eliminate the background signals and instrumental interferences from true biologically-related signals (Castillo *et al.* 2011). Peak detection is referred to as the representation of ion signals as “features” with specific m/z , retention time and peak area information (Varghese *et al.* 2012). De-isotoping and ion annotation is used afterward to cluster peaks related to the same metabolites such as isotopes, adducts, and in-source-fragment ions (Varghese

et al. 2012). For quantification purposes (or comparing peak intensities between different samples), normalization of signals is also required to prevent errors caused by instrumental or sample preparation-related variations (Katajamaa and Orešič 2007). Statistical analysis is then performed to select signals representing significant differences among sample groups, *e.g.* for biomarker discovery. Identification of compounds is the next step, which is usually challenging and time-consuming (Scalbert *et al.* 2009; Hall 2011; Zhou *et al.* 2012). This pipeline represents the usual workflow of data processing in fingerprinting and profiling metabolomics. Depending on the specific type of study, whether it is targeted, fingerprinting or profiling, some steps would not be necessary or could be modified. For example, for metabolomics fingerprinting, statistical analysis to find discriminative signals is followed by identification of corresponding metabolites (Roux *et al.* 2011). Each of these steps will be explained in more detail in the following sections.

1.4.3.1 Noise filtering

The very first step in the treatment of LC-MS data, whether it is used for metabolomics or proteomics, is to filter the noise and baseline correct the data. This step has the potential to improve the quality of peak detection by reducing detection of false positive features (Castillo *et al.* 2011), since raw LC-MS data suffers from both chemical and random noise (Katajamaa and Orešič 2007). Chemical noise is normally very evident at the beginning and end of the elution gradient and often originates from molecules of solvents and buffers used for sample preparation or chromatographic separation, as well as column bleed (Hilario *et al.* 2006). Random noise is generally caused by imperfect detector function (Zhang *et al.* 2009).

Several different methods are used for this purpose including Savitzky-Golay type of local polynomial fitting (Wang *et al.* 2003) and wavelet transformation (Li *et al.*

2005) which are both applied in m/z direction or filtering with moving averages in chromatographic trend (Radulovic *et al.* 2004).

Baseline correction is usually performed by finding the baseline shape and subtracting it from LC-MS raw data (Katajamaa and Orešič 2007). Several approaches have been done namely low-order polynomial Savitzky-Golay filter (Wang *et al.* 2003), linear regression for lowest point of smoothed spectrum (Haimi *et al.* 2006) or iterative asymmetric least-squares estimation (Eilers and Boelens 2005), which are one-dimensional background estimations.

The filtering and baseline removal is implemented in some peak detection software such as XCMS (Smith *et al.* 2006), MAVEN [21] and apLCMS [18], while some other software such as OpenMS (Sturm *et al.* 2008) offer several filters for the user to choose from.

1.4.3.2 Peak detection

Peak detection, also known as feature detection or peak picking, is the process of extracting signals of MS peaks (m/z) and chromatographic signal (retention time) as well as peak area or intensity measurement of all detected peaks (Figure 1.6) (Varghese *et al.* 2012).

From a signal processing perspective, peak detection is carried out based on one or more of the following parameters: signal-to-noise ratio (SNR), intensity threshold, slopes of peaks, local maximum, shape ratio, ridge lines, model-based criterion and peak width (Yang *et al.* 2009). Typically, a combination of methods is used in order to increase the quality of peak detection and lower the chance for identification of false positive peaks. For instance, the basic version of XCMS [33] bins the data to 0.1 m/z windows, then by considering the maximum intensity at each RT, it

identifies the signal in each slice. A second filtering criteria based on peak shape is used followed by the final selection of peaks using signal-to-noise ratio cut-off. Detailed information about peak detection algorithms is beyond the scope of this chapter. For more informatics content, you can refer to the comprehensive review article by Zhang *et al.* (Zhang *et al.* 2009).

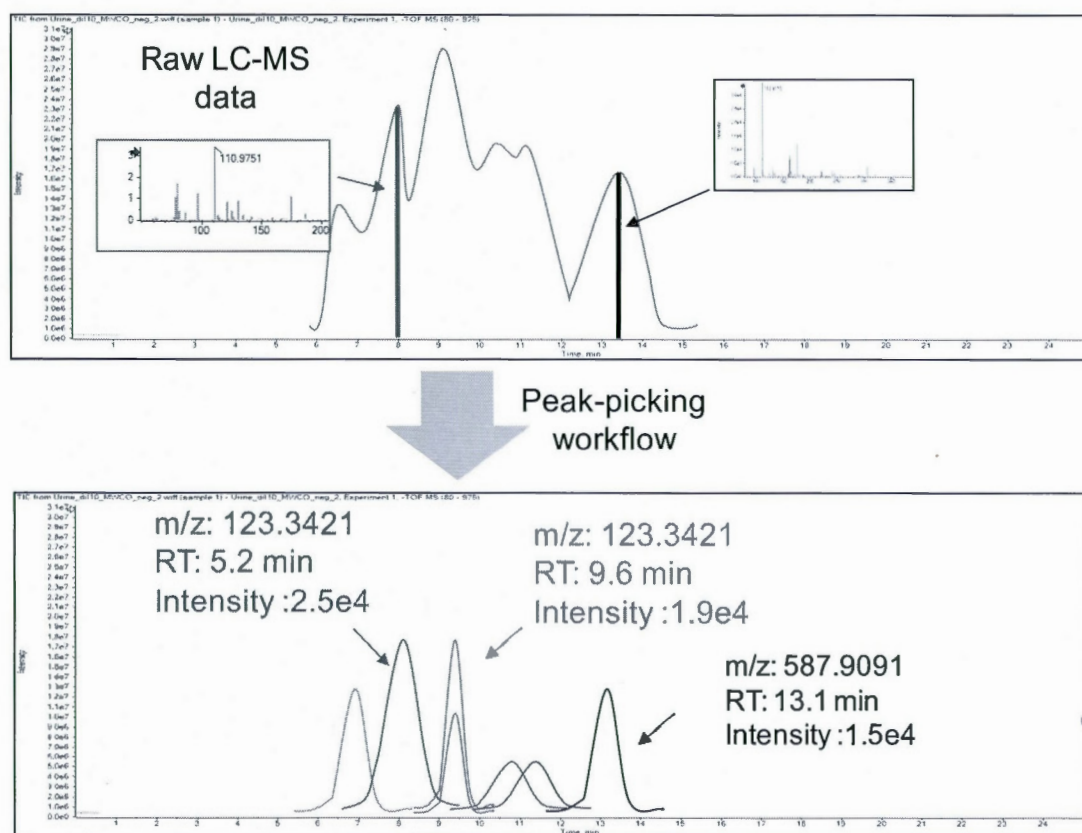


Figure 1.6 Peak detection process used to extract bounded information of mass signal (m/z), retention time (RT) and intensity of detected ions

Two general types of software and workflows are available for this purpose, including commercial and freely-available software. Commercial software are usually provided by MS instrumentation vendors, such as MarkerView (AB Sciex), PeakView (AB Sciex), MarkerLynx (Waters), SIEVE (Thermo), MassProfiler Professional (Agilent), or ProfileAnalysis (Bruker). The underlying operating

algorithms for these software is often not clear due to the commercial considerations. In addition, there are some free and/or open access software such as MetaboAnalyst (Xia *et al.* 2009; Xia *et al.* 2012), MZmine (Katajamaa *et al.* 2006), XCMS (Smith *et al.* 2006; Tautenhahn *et al.* 2008) and MetAlign (Lommen 2009). In the case of some open sources such as for XCMS (Smith *et al.* 2006; Tautenhahn *et al.* 2008) and MZmine (Katajamaa *et al.* 2006), the operating algorithm is accessible and could be modified by the user. Katajamaa *et al.* provided the lists of commercial and freely-available software used for metabolomics applications (Katajamaa and Orešič 2007). Furthermore, a list of freely available software and the codes used for different LC-MS data processing is provided at the address of (<http://www.ms-utils.org/>). These codes commonly work in computer programming environment such as Java, Matlab, C or R.

Although there is no limitation on the number of software and workflows available for feature detection, there are some challenges in this area. For instance, the final results of different platforms could differ widely due to different algorithms used. Hence, the decision for choosing the right platform becomes critical in metabolomics data analysis. Usability, documentation and easy visualization of the results are the main factors for selection of appropriate workflow used by common users, especially, those unfamiliar with programming languages. The ability of algorithm to distinguish between low intensity peaks and noise is also important factor to be considered. Another main criteria is the coverage of software on different aspects of the data processing workflow, starting from noise filtering and baseline correction to be done automatically along with feature detection (Castillo *et al.* 2011).

1.4.3.3 Ion annotation

LC-MS-based metabolomics experiments can result in a huge number of peaks, only a portion of which are related to true biological metabolites. The fact that each single metabolite can also give rise to several ions, namely adducts, in-source fragments and isotopes, makes data processing challenging. Therefore, if only mass-based search is carried out for peak detection, false identification of peaks is normal (Varghese *et al.* 2012). Ion annotation is the method used for assigning all redundant peaks corresponding to ions arising from the same species and grouping them together to reduce the complexity of data for further biological interpretation.

Since most of the elements exist in more than one naturally-occurring isotopic form, molecules containing different isotopes have different masses, detected as their isotopic pattern in the mass spectra (Jaitly *et al.* 2009). In-source fragment ions result from dissociation of the intact ionized molecules before they enter the mass analyzer, and although ESI is considered as soft ionization source, this phenomenon still can occur for certain compounds. Losses of water, ammonia and/or CO₂ are common in-source fragments in metabolomics data. The third type of derivative peaks comes from adduct ions. An adduct ion is, by definition, "*an ion formed by the interaction of two species, usually an ion and a molecule, and often within the ion source, to form an ion containing all the constituent atoms of one species as well as an additional atom or atoms*" (McNaught and Wilkinson 2000). Sodium and potassium adducts are common species observed in MS of small molecules. For lists of common adducts seen in MS experiment, you can refer to articles by Haung *et al.* and Keller *et al.* (Huang *et al.* 1999; Keller *et al.* 2008).

The ion annotation employs two clues for assigning redundant peaks: 1) the mass difference between two peaks should match with related isotopes, adducts or

fragment ions, 2) the similarity between the extracted ion chromatogram of two peaks as they have the same elution profile (Varghese *et al.* 2012).

Numerous efforts have been made for developing of either an independent ion annotation workflow (working on the results of feature detection) or implemented algorithms within feature selection software. For example, CAMERA imports the results of R-based XCMS and annotates peaks in two steps. First, peaks are grouped based on retention time and similarity between peaks, and then the difference between their m/z values is compared with a list of normally occurring adducts and in-source fragments for any possible relationship (Kuhl *et al.* 2011). Another workflow uses pre-defined m/z differences, chromatographic elution and intensity correlation (for isotopic peaks) to assign redundant peaks, resulting in 50% data reduction (Brown *et al.* 2009). In addition, there are some commercial software such as ACD/IntelliXtract (a part of the ACD/MS workbook suite) which works based on the given rule table (ACD/IntelliXtract 2007). PUTMEDID-LCMS is a public tool which imports raw LC-MS data and group peaks originating from the same metabolites by mass difference, retention time and peak area matching (Brown *et al.* 2011). Peak area correlation is employed for confirmation criteria of the isotopic peaks. IDEOM, free implementation for Microsoft Excel assigns ESI redundant peaks as well as FT or ringing signals by employing RT, peak shape and intensities and difference in m/z values (Creek *et al.* 2012).

1.4.3.4 Alignment, normalization and statistical analysis

Alignment is a crucial step for metabolomics analysis on more than one sample or more than one run, since small variations are often seen in retention time and m/z values of the same metabolite across different samples (Podwojski *et al.* 2009). Alignment algorithms either work based on raw LC-MS data or features found previously by peak picking tools. Moreover, some peak picking packages such as

XCMS (Smith *et al.* 2006) or MarkerView software incorporate the alignment method as well. The fact that elution differences among samples may be non-linear and multiple pairs of feature may be found as matching peaks, makes alignment a challenging process and thus needs to be performed with caution (Castillo *et al.* 2011). Podwojski *et al.* studied three different algorithms including linear regression, loss regression and local vectors and their results showed the importance of considering non-linear deviation in proteomics data (Podwojski *et al.* 2009). A detailed review on available alignment algorithms used for LC-MS is provided by Vandenbogaert *et al.* (Vandenbogaert *et al.* 2008).

Normalization of peak intensities may be performed specially for quantitative measurement, or when metabolite fingerprinting is performed. The unwanted systematic bias in LC-MS data, resulting from several sources such as experimental difference, could overshadow the real biological difference in concentration of metabolites. Hence, peak intensities should be corrected before doing statistical analysis between samples.

Two approaches are dominant for normalization including statistical and standard addition method. The first one is based on whole dataset for instance, normalization by unit norm of intensities (Scholz *et al.* 2004), the maximum likelihood method (Orešič *et al.* 2004) or median of intensities (Wang *et al.* 2003). However, this method is unable to assign absolute concentration of metabolites (Katajamaa *et al.* 2006). The second method is to use one or more standard compounds as a reference for normalization and absolute quantitation. Due to the large number of metabolites studied in metabolomics investigations, the selection of proper standard compounds for normalization is ambiguous when more than one standard is used. Similarity between elution behavior of analyte and standard compound is used for selection of appropriate standard compound for each analyte.

Statistical analysis is generally used for two purposes. It can be used for sample classification when limited information about samples is available (*e.g.* identification of silent mutation phenotypes in offspring). In this case, unsupervised statistical methods are employed such as hierarchical clustering analysis (HCA), principal component analysis (PCA), or independent component analysis (ICA). In addition, statistical analysis is used to find metabolites with great differences in intensities among different classes of samples, related to the studied condition, such as health, diet or exposure to toxins. Using supervised mathematical methods like partial least squares (PLS) or soft-independent method of class analogy (SIMCA) can be used for this purpose (Dettmer *et al.* 2007).

1.4.4 Metabolite identification

Identification of metabolites is the last step in the metabolomics pipeline before any biological information could be interpreted regarding biochemical pathways or biomarker discovery. It is performed in all metabolomics studies, no matter what the type of the investigation. For untargeted metabolomics, it is performed for all peaks detected as monoisotopic ions, while the effort for fingerprinting studies is to identify peaks with significant differences in different samples. For targeted metabolomics, identification is most likely done prior to LC-MS analysis, and absolute quantitation is instead the last step.

Two main strategies are used widely for identification of compounds by mass spectrometry. Accurate mass measurement of selected ions acquired by high resolution MS systems (HRMS) could yield elemental formulae for the chemical characterization of metabolites. In addition, tandem mass spectrometry results, yielding the fragmentation pattern, could be used for structural elucidation of metabolites. For the final confirmation of metabolites, standard solutions of compounds should be analyzed with the same instrument using the same method, if

available, to ensure chromatographic elution and mass spectral similarity (Bueschl *et al.* 2013).

Database searching is one of the most time consuming parts of a metabolomics workflow, and can also have some difficulties. For instance, the collision energy in which MS/MS spectra for standard compounds is acquired could be different from that of the experiment, resulting to differences in fragment ion patterns. In addition, lack of MS/MS spectra for many compounds in metabolomics databases makes the identification of metabolites a highly labor-intensive task (Bueschl *et al.* 2013).

In addition to relying on the high efficiency of MS instruments, more classical chemical methods can be used in metabolomics workflows. For example, chemical derivatization can help identify the functional groups and differential labeling can aid in the relative quantitation of metabolites by observing a specific mass shift and intensity ratio (Dettmer *et al.* 2007). Hydrogen/deuterium exchange methods also provide information on the number of exchangeable protons for identification of some functional groups like alcohols, amines, carboxylic acids, *etc.* (Dettmer *et al.* 2007).

1.4.4.1 Databases

The number of metabolomics databases and their metabolite content is still limited in comparison with genomics and proteomics databases due to being relatively newer, and less straightforward. PubChem (Wang *et al.* 2009), METLIN (Smith *et al.* 2005) and KEGG (Kanehisa *et al.* 2004) provide simple mass-based searches while HMDB (<http://www.hmdb.ca>) provides clinical and molecular biology data as well.

Other useful metabolomics databases are Manchester metabolomics database (MMD) (Brown *et al.* 2009) that is created from other sources like genome scale metabolic models (Herrgård *et al.* 2008), Human Metabolome Database (HMDB) (Wishart *et al.* 2007), Lipid Maps (Sud *et al.* 2007), BioCyc (Karp *et al.* 2005) and DrugBank (Wishart 2008).

1.5 Research objectives

LC-MS based metabolomics experiments produce a large amount of data which needs special care for data processing. As it was shown in previous section, data processing has several steps, where peak picking and ion annotation have great importance. As different peak picking software use various algorithms, the final results provided by these software may be different and directly affect metabolomics results. Hence, two objectives were defined for this research as follows:

1. To compare four different peak picking software for untargeted metabolomics applications with the aid of MATLAB programming. The studied peak detection workflows include three commercial packages, PeakView®, Markerview™, MetabolitePilot™, (all provided by AB Sciex), and freely available, XCMS online workflow.
2. To develop MATLAB-based code for ion annotation of peak picking results.

These two approaches were designed to help the improvement of LC-MS based metabolomics as relatively new science to provide better performance in addressing biologically-important questions. In addition, in a separate project, a targeted metabolomics assay was improved by presenting the modifications on sample preparation procedure which covers the third main objective of the MSc thesis.

3. Quantitation of four carotenoid compounds in algal samples using HPLC-HRMS system. The aim of this work was to present a simple and fast extraction method for analyzing the carotenoid content of algal solutions in response to introduced stress conditions.

1.6 Thesis outline

This thesis has been organized as follows:

Chapter 1 introduces metabolomics concepts, different branches of this new *omic* science, as well as an overview of the experiment pipeline. This chapter also presents the research objectives and layout of the thesis.

Chapter 2 describes the research on comparison of peak picking workflows for metabolomics profiling research on two biological samples. Four different peak detection software were compared including three commercial software from AB Sciex (Peakview®, Markerview™, MetabolitePilot™) and XCMS Online (open-source web-based software). Raw LC-MS data from two biological sample types (bile and urine) as well as a standard mixture of 84 compounds were processed with same criteria. Then, the overlaps between the results were investigated by a MATLAB script developed for this purpose. Finally, the resulting lists of potential metabolites from each workflow were investigated using the METLIN database based on accurate precursor ion mass and MS/MS spectral matching. The performance of these four peak picking workflows was also evaluated with a custom standard mixture of 84 biologically-relevant small molecules. Work presented in this chapter is the subject of a published peer-reviewed article in *Rapid Communications in Mass Spectrometry* (Rafiei and Sleno 2015).

Chapter 3 presents the MATLAB-based ion annotation workflow developed to filter out redundant peaks from peak picking results. This code was designed to import data from any peak generating workflow in excel format, perform different stages of filtering and results in more condensed peak lists by removing redundant peaks. After evaluation of the performance of this filtering method, the performance of four peak picking workflows, namely Peakview®, Markerview™, MetabolitePilot™ and XCMS online, were evaluated in terms of number of peaks found as redundant peaks by our newly developed “*DataReduction*” workflow.

In chapter 4, a targeted metabolomics approach was developed for the absolute quantification of the changes in carotenoid content of three algal samples under stress conditions. Three different algae species are *Haematococcus*, *Oocystis*, and *Muriellopsis*. Carotenoid separation and subsequent analysis were done on a UHPLC instrument coupled to a hybrid quadrupole time-of-flight mass spectrometer. An online UV detector was also used for further confirmation of the studied compounds. Based on exact mass measurements, four carotenoids were quantified in control and stressed-algal samples.

CHAPTER II

COMPARISON OF PEAK PICKING WORKFLOWS FOR UNTARGETED LC-HRMS METABOLOMICS DATA ANALYSIS

This work has been published in a journal paper of *Rapid communication in mass spectrometry* with my contribution as first author (Rafiei and Sleno 2015).

2.1 Abstract

Data analysis is a key step in mass spectrometry-based untargeted metabolomics, starting with the generation of generic peak lists from raw LC-MS data. Due to the use of various algorithms by different workflows, the results of different peak picking strategies often differ widely. Raw LC-HRMS data from two types of biological samples (bile and urine) as well as a standard mixture of 84 compounds, were processed with four peak picking softwares: Peakview®, Markerview™, MetabolitePilot™ and XCMS Online. The overlaps between the results of each peak generating method were then investigated. To gauge the relevance of peak lists, a database search using METLIN online database was performed to determine which features had accurate masses matching known metabolites as well as a secondary filtering based on MS/MS spectral matching. In this study, only a small proportion of all peaks (less than 10%) were common to all four software programs. Comparison of database searching results showed peaks found uniquely by one workflow have less chance of being found in the METLIN metabolomics database and even less likely to be confirmed by MS/MS. It was shown that the performance of peak generating workflows has a direct impact on untargeted metabolomics results. As it was demonstrated that the peaks found in more than one peak detection workflow have higher potential to be identified by accurate mass as well as MS/MS spectrum matching, it is suggested to use the overlap of different peak picking workflows as preliminary peak lists for more rugged statistical analysis in global metabolomics investigations.

2.2 Introduction

Although there is debate on the terminology related to metabolomics (Villas-Bôas *et al.* 2005), it can be defined as the quantitative characterization of small molecules (metabolites) present in a biological sample (Lindon *et al.* 2011). It is a relatively new “omic” science, with potential applications in many research areas, such as drug discovery (Wishart 2008), oncology (Spratlin *et al.* 2009), medical diagnosis and therapeutic monitoring (Gowda *et al.* 2008) as well as food science (Wishart 2008) and agriculture (Dixon *et al.* 2006). Metabolomics studies are divided into two main categories: targeted and untargeted. While in a targeted workflow, a pre-defined list of metabolites is surveyed, untargeted metabolomics aims to obtain a global overview of as many metabolites as possible in the sample and to monitor changes caused by disease, drug treatment, *etc.*

Biological samples studied in metabolomics are often very complex. For instance, estimates include >1000 metabolites present in *Escherichia coli* (Feist *et al.* 2007), >4000 for human serum (Psychogios *et al.* 2011), and between 5000 and 25000 for higher plants (Trethewey 2004). Working with this large number of compounds in untargeted studies requires special consideration when processing high resolution mass spectrometry (HRMS) data.

Employing peak-picking workflows to filter raw LC-MS data is the first step in MS-based untargeted metabolomics data analysis. There are a wide variety of software packages available for this purpose. Some software are provided by commercial MS instrument vendors, such as MarkerView (AB Sciex), PeakView (AB Sciex), MarkerLynx (Waters), MassProfiler Professional (Agilent), SIEVE (Thermo) or ProfileAnalysis (Bruker). There are also freely available open or close source workflows, *e.g.* MZmine (Katajamaa and Orešič 2005), XCMS (Smith *et al.* 2006), and MetAlign (Lommen 2009). Although there is no shortage of software available, various algorithms are used by different peak picking workflows, hence

final processed results can differ widely from each other. Peak picking workflows filter results based on one or more of the following parameters: signal-to-noise ratio (SNR), intensity threshold, slopes of peaks, local maximum, shape ratio, ridge lines, model-based criterion and peak width (Yang *et al.* 2009). Zhang *et al.* have reviewed this subject in detail (Zhang *et al.* 2009).

The effect of employing different peak picking algorithms on the same LC-MS data was investigated by Bauer *et al.* for proteomics applications (Bauer *et al.* 2011). Three peak detection algorithms including signal-to-noise ratio (SNR), template-based peak detection and Continuous Wavelet Transform (CWT) were evaluated for protein analysis. By employing a defined set of reference peaks, sensitivity and specificity of peak picking algorithms were compared. Their results show that performance of SNR algorithms depends highly on data quality, while template-based peak detection algorithms may ignore asymmetrical peaks. However, the latter showed robust performance for lower noise levels. The CWT method showed good performance for even relatively high noise but tuning the algorithm is difficult due to the high number of parameters involved. By employing both simulation data as well as real data, CWT method showed the best performance. In another study, three peak detection packages were tested including msInspect, MZmine, as well as an algorithm described in VIPER software and the effect of various peak-picking criteria was evaluated for each package (Zhang *et al.* 2009). The challenge is not exclusive to GC- and LC-MS based data. MALDI MS data was also subjected to the investigation on different peak detection algorithms including Cromwell, CWT, LMS, LIMPIC and PROcess (Yang *et al.* 2009).

In this work, different peak picking software were compared rather than peak detection algorithms for two reasons. First, some workflows might use more than one algorithm for differentiating between peaks and noise, hence pure comparison of algorithms would be less useful. Secondly, peak picking algorithms used by

software may remain unknown by end-users using these commercial packages. Therefore, a practical approach of evaluating and comparing peak picking workflows for metabolomics applications is presented. Studied workflows include the use of three commercial software from AB Sciex (Peakview®, Markerview™, MetabolitePilot™) as well as XCMS Online (an open-source web-based software). Raw LC-MS data from two biological sample types (bile and urine) were processed with the four different workflows. In order to show differences between the performance of each software, the overlaps between the results were then investigated (using a “VennPro” MATLAB script). Finally, the resulting lists of potential metabolites from each workflow were investigated using the METLIN database based on accurate precursor ion mass and MS/MS spectral matching. The performance of these four peak picking workflows was also evaluated with a custom standard mixture of 84 biologically-relevant small molecules.

2.3 Experimental

2.3.1 Materials

Cholic acid, deoxycholic acid, tryptophan methyl ester, sodium diclofenac, ibuprofen, *S*-benzyl-cysteine, 17 α -ethylestadiol, canthaxanthin, 3-hydroxyanthranilic acid, kynurenine, kynurenic acid, formic acid and HPLC grade methanol were obtained from Sigma-Aldrich (Oakville, ON, Canada). Atrazine and anthranilic acid were from Fluka (Oakville, ON, Canada) and acetonitrile was obtained from Caledon (Georgetown, Ontario, Canada). Sodium hydroxide was purchased from Anachem (Lachine, QC, Canada). Ultrapure water was supplied by a Synergy UV purification system from Millipore (Billerica, MA, USA). PM1 to 5 MicroPlates™ were bought from Biolog (Hayward, CA, USA) as a source for many standard compounds. Urine and bile samples from individual healthy untreated dog (Beagle) were obtained from CiToxLAB (Laval, QC, Canada).

2.3.2 Sample preparation

2.3.2.1 Standard mixture

Initially, 84 stock solutions of known compounds with different physicochemical properties and molecular weight ranging from 88 to 564 g/mol were prepared. The detailed information of the sample preparation and list of metabolites (Table S1) used in this compound mixture is presented at the end of this chapter (supplementary data for chapter 2). Each of the 84 compounds was directly injected into the mass spectrometer in both positive and negative mode, without chromatographic separation. From these direct (loop) injections, protonated ions (MH^+) were observed for 78 metabolites and deprotonated ions (MH^-) were observed for 73 metabolites. The mixture of 84 compounds was prepared by mixing each standard solution with a final concentration ranging from 1-500 μM for all molecules.

2.3.2.2 Biological samples

Urine and bile samples from untreated dog were subjected to molecular weight cut-off (MWCO) filtering to reduce metabolite loss from the biological samples. Samples (undiluted urine and 5-fold diluted bile) were filtered using 0.45 μm regenerated cellulose spin filters (Canadian Life Sciences, Peterborough, ON, Canada) at 1250 rpm for 2 minutes, to remove any insoluble material, and then by 5 kDa MWCO regenerated cellulose spin filters (Amicon, Oakville, ON, Canada) for 20 minutes at 14,000 rpm, thus removing any large molecules (*e.g.* proteins) from samples prior to analysis. Resulting samples were then diluted 10-fold prior to HPLC-MS/MS analysis. The same procedure was performed with ultrapure water as a control (blank) to filter out any contaminant peaks resulting from filters, tubes or LC-MS system.

2.3.3 HPLC-MS analysis

Samples (10 μ l) were injected three times each onto a BetaBasic C18 column (2.1 \times 150 mm), with 3 μ m particles (Thermo Scientific, Canada) using a Nexera® UHPLC system (Shimadzu, Columbia, MD). Liquid chromatographic separation was performed with mobile phases of 0.1% formic acid in water (A) and 0.1% formic acid in MeOH (B), with an initial hold at 3% for 2 min, followed by a gradient of 3–50% B in 15 min, to 90% B at 20 min, held until 25 min, with a flow rate of 300 μ l/min at 40 °C.

All MS spectra were acquired on a high-resolution hybrid quadrupole-time-of-flight (QqTOF) TripleTOF® 5600 mass spectrometer (AB Sciex, Concord, ON, Canada) equipped with a DuoIonSpray source, in positive and negative electrospray mode. The instrument performed a survey TOF-MS acquisition from m/z 80-800 with an accumulation time of 300 ms, followed by MS/MS on the four most intense ions from m/z 80-800 using information-dependent acquisition (IDA) with dynamic background subtraction (DBS). Each MS/MS had an accumulation time of 150 ms and collision-offset voltage of 30 ± 10 V. TOF-MS and MS/MS were automatically calibrated every four injections with an in-house standard mix (m/z 119-966 in negative mode and m/z 121-922 in positive mode).

2.3.4 Data Processing

Data processing was performed in three steps: peak picking, MATLAB processing followed by searching for potential metabolites in METLIN database using accurate mass and MS/MS spectral matching. The LC-MS data from bile, urine and a standard mixture in both positive and negative ionization modes were processed with PeakView, MetabolitePilot™, MarkerView and XCMS online by employing identical parameters (± 5 ppm mass accuracy, 500 cps threshold and minimum peak width of 5s).

2.3.4.1 Peak Picking

Raw LC-MS data was processed with four peak picking software: MetabolitePilot™ 1.4, MarkerView™ 1.2, PeakView® 2.0 (AB SCIEX), as well as XCMS online (<https://xcmsonline.scripps.edu/>). The criteria used by each are given below.

MetabolitePilot Minimum peak width: 5s, minimum chromatographic intensity: 500 cps, smoothing before peak finding, sample/control ratio greater than 5. MS m/z tolerance: 10 ppm, minimum MS peak intensity: 500 cps, maximum number of metabolites: 1000. MetabolitePilot has a limit of 1000 peak/run for generic peaks, therefore mass range windows were set as narrow as necessary to have peak numbers not exceeding this limitation. Generated peaks were then visually inspected and peaks resulting from background noise were removed directly in software.

MarkerView A feature peak list was created directly from raw data (.wiff) files with subtraction offset of 10 scans, minimum spectral peak width: 10 ppm, minimum RT peak width: 5 scans, signal-to-noise threshold of 5. Then LC-MS peak lists (*.peaks) from multiple samples were imported into MarkerView using the following criteria: retention time tolerance: 0.33 min, mass tolerance: 10 ppm, intensity threshold: 500. A *t*-test was then performed to compare samples (three replicates) with controls (three replicates) peaks with <5 fold increase compared to blank samples were then removed from peak list.

PeakView Extracted ion chromatograms were visualized with a width of 0.02 Da, an intensity threshold of 500 cps and peak detection sensitivity at medium. Generated extracted ion chromatograms (XICs) were visually inspected and irregular peaks were removed. Sample XIC lists were then investigated in blank

samples and peaks with <5 fold signal intensity compared to blank were removed by employing a simple MATLAB script.

XCMS online Raw LC-MS data (.wiff) files were first converted to mzXML using ProteoWizard 3.0.3548 (Chambers *et al.* 2012). After uploading mzXML data to the XCMS website, a CentWave method was used for peak picking with the following parameters.: maximal m/z tolerance of 10 ppm, and peak widths from 5 to 30 s, $mzwid$: 0.015, $minfrac$: 0.5, $bandwidth$: 5. The resulting text file was exported into excel and all peaks with less than 5x for sample/control ratio were removed.

2.3.4.2 MATLAB processing

MATLAB R2012a (MathWorks, 2012) was used for processing peak picking results. A "*VennPro*" MATLAB-based workflow was developed to find overlaps between the results of the four tested workflows. It imports peak lists (in excel), and finds all possible overlaps between different groups. It can be used to find similar peaks across samples or across the results of different peak finding algorithms. It used a 10 ppm m/z window and 0.15 min difference in retention time to identify similar peaks. The results of this MATLAB script were used to draw venn-diagrams for overlaps between peak picking results.

2.3.4.3 METLIN database search

METLIN web-based metabolomics database (<http://metlin.scripps.edu/index.php>) was used for tentative identification of metabolites. Database matching was performed in two steps, including accurate mass and MS/MS spectral matching. For accurate mass filtering, 5 ppm mass tolerance was used for MH^+ or MH^- ions in positive and negative modes, respectively. The results were saved in .CSV format

and, employing a MATLAB script, the total number of m/z with at least one hit in the database was calculated as well as the number of metabolites with at least one available MS/MS spectrum in METLIN. For MS/MS matching, the information-dependent acquisition (IDA) spectra having a quality score > 60 were used. After initial visual inspection, the most probable matches were evaluated by the "MS/MS spectrum match" option of METLIN by importing the 30 most intense peaks of each MS/MS spectrum into METLIN, using 10 ppm tolerance for precursor ion and 0.05 Da for MS/MS. The results with match score > 60 were reported as matched.

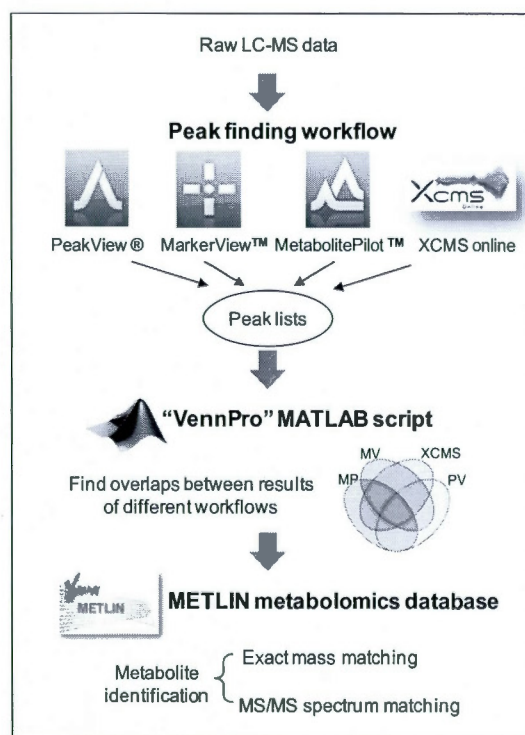


Figure 2.1 Method used in this study to compare four peak picking workflows. Raw LC-MS data were processed with one of the following software: MetabolitePilot, MarkerView, PeakView and XCMS online. The overlaps between results were then found using an in-house "Venn-pro" MATLAB script followed by METLIN online metabolomics database searching

2.4 Results and discussion

The LC-MS data from standard mixture, bile and urine in both positive and negative ionization modes were processed with four peak detection workflows including PeakView, MetabolitePilot, MarkerView and XCMS online by employing similar criteria (5 ppm mass accuracy, 500 cps threshold and minimum peak width of 5 s). The workflow employed in this study is illustrated in Figure 2.1. For instance, MarkerView and XCMS represent the isotopic peaks among the results while PeakView and MetabolitePilot didn't indicate the isotopic peaks and may exclude them initially. In this work, all two dimensional results obtained directly from peak detection workflows Various definition have been used for the definition of "peak" in different software packages. MarkerView and XCMS represent the isotopic peaks while PeakView and MetabolitePilot did not indicate the isotopic peaks. In this work, all two dimensional results obtained directly from peak detection workflows are referred as "peak".

2.4.1 Standard mixture

A standard mixture of 84 known compounds was used in order to evaluate the performance of the different peak picking workflows compared in this study and also in order to compare the results of untargeted and targeted approaches. First, LC-MS data in positive and negative modes were processed with PeakView, MarkerView, MetabolitePilot and XCMS online. Then the overlaps between the results were found using "Venn-Pro" MATLAB script. METLIN search results for all peaks were investigated, using the molecular formula of known compounds. This resulted in 24 and 28 metabolites in positive and negative mode, respectively, found by at least one peak picking workflow and having a MS/MS spectrum available in the METLIN database. A second LC-MS/MS analysis was performed to acquire MS/MS spectra of these metabolites based on their accurate precursor ion

masses of protonated and deprotonated molecules using an inclusion list as MS/MS triggering (IDA) criteria. Figure 2.2 represents the Venn diagram for the results of MS/MS matched standard compounds. It was observed that a) none of the peak picking strategies resulted in 100% recovery of standard metabolites, with only 24 metabolites being found in positive mode and 13 in negative mode, b) the capability for detecting standard metabolites varied widely between workflows. For instance, MarkerView found the highest number of metabolites with a recovery ratio of 23/24 in positive mode. PeakView showed good performance (20/24), especially considering its much shorter preliminary peak list. MetabolitePilot and XCMS online gave similar results both with a recovery ratio of 18/24. In negative mode, metabolites were identified based on the results of MarkerView and XCMS, while this number was 10 for PeakView and 11 for MetabolitePilot. It also shows the complementary nature of different peak generating algorithms. For instance, from 24 detectable metabolites in positive mode, only 13 metabolites are present in all four workflows. The lists of compounds from each region of the Venn-diagrams are presented in supplementary data (Tables S2 and S3).

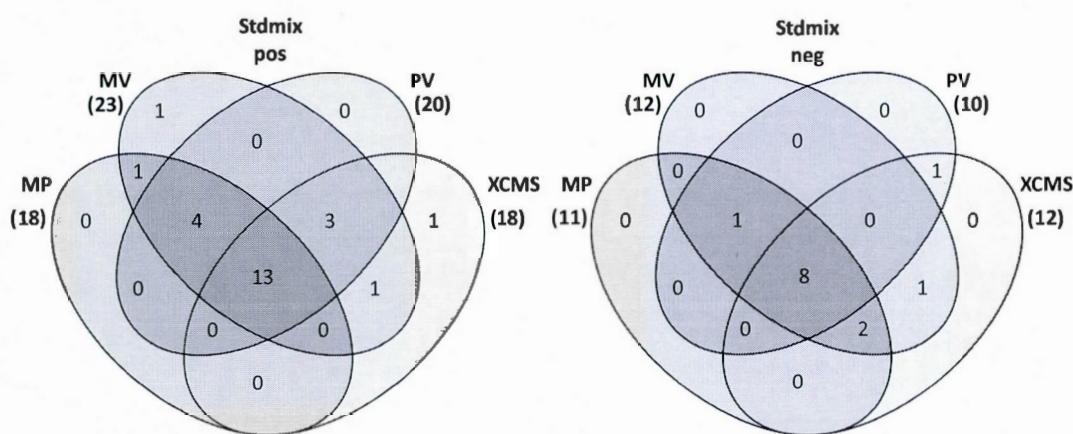


Figure 2.2 The overlaps between detected standard metabolites (confirmed by MS/MS spectral matching) using four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV) and XCMS online) for a standard mixture of 84 compounds in positive and negative ionization modes

A targeted metabolomics analysis was also performed using PeakView for the standard mixture in two steps: accurate mass matching and MSMS spectrum matching. In first step, extracted ion chromatogram survey (5 ppm mass accuracy) was done for protonated and deprotonated ions for positive and negative ionization modes, respectively. Manual analysis confirmed the presence of 40 compounds in positive mode, while 30 compounds were detected in negative ionization mode. At the consequent step, IDA results were compared with the MS/MS spectrum results of database. In this step 36 and 22 metabolites were completely matched to the METLIN data base results in positive and negative modes. The comparison of targeted analysis with the total number of peaks detected by four peak detection workflows in untargeted manner shows that all untargeted results are covered with manual analysis. There are 12 and 9 metabolites in targeted analysis results that are not found by any of used peak detection workflows (Table S4 and S5).

Several reasons may cause incomplete coverage of all 84 metabolites either by manual analysis or automatic peak detection software: a) The experimental setup (*e.g.* chromatographic column and elution gradient) limits to observe a number of metabolites, b) Incomplete coverage of METLIN data base, 3/4 of metabolites had MSMS spectrum in each ionization modes, and c) Unavailable IDA experimental results prevents to confirm MS/MS matching for a number of metabolites. d) Not all of the compounds are observable in both ionization modes.

2.4.2 Biological samples (bile and urine)

Bile and urine represent complex biological samples with very different metabolic profiles. In Figure 2.3, TICs for both sample types in positive and negative modes are displayed and exhibit the contrast in polarity of the majority of metabolites present which can be detected by LC-MS. Also, bile interestingly shows much higher intensity in negative mode, presumably due to the presence of hydrophobic bile acids in this biofluid. As shown in Figure 2.4, the total number of peaks found

by each software varied greatly, *e.g.* MetabolitePilot found 9879 peaks for the urine data set in negative mode, Markerview and XCMS found 11553 and 11708 peaks, respectively, while PeakView found only 2015 peaks. PeakView, with the lowest number of generated peaks, showed a weak performance from this point of view in comparison with other workflows. Although XCMS found the largest peak lists for bile and the standard compound mixture in both positive and negative modes, the total number of peaks found by MarkerView was larger in the urine sample. These results indicate the importance of selecting an appropriate peak-picking step in untargeted metabolomics studies, since this procedure could impact directly on the final results of study. Nevertheless, it is not always the workflow yielding the highest number of peaks, which should be automatically deemed the best. This, of course, will depend on the quality of the resulting peak lists, for defining the metabolome without too many extraneous features being monitored, which could undoubtedly misguide the outcome of statistical analysis when different sample groups are to be compared in metabolomics fingerprinting studies.

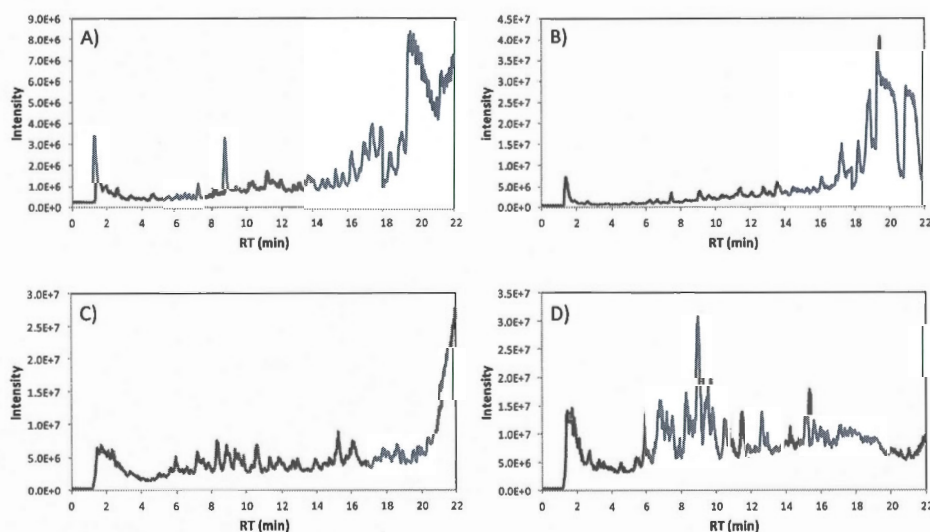


Figure 2.3 Total ion chromatogram (TIC) for bile in positive and negative modes (A and B respectively) and urine in positive and negative modes (C and D respectively). For added clarity, TICs from bile samples were scaled down 3-fold from 18-22 minutes in the above chromatograms

Regarding the total number of peaks found, the results of different peak picking algorithms differ widely from each other. In order to find overlapping peaks, peak lists from each of the four peak generating workflows were imported into a custom-built "*VennPro*" MATLAB script (using 5 ppm mass and 0.15 min retention time tolerance, for assigning similar peaks). The average percent of peaks found in each overlap is represented in Figure 2.5. Results of this MATLAB processing, from each sample type in positive and negative modes illustrated separately, are also presented in supplementary data (Figure S1). It was found that an average of 41.1% of total features are detected only by XCMS online without any overlap with other software. This weak overlap of XCMS with other software could actually be introducing more "noise" into the statistical analysis. Another criteria assessed was based on which software yielded the most "repeatability" with found peaks from other workflows. MarkerView showed the best results in this comparison being involved in the highest 2-way (9.9%) and 3-way (10.6%) overlaps. The higher performance of MarkerView is coherent with previous results (Figure 2.2) of detecting higher number of metabolites in standard mixtures. All four workflows yielded an overlap of 7.7% of all found peaks. This comparison led to the most certainty in peaks found by MarkerView.

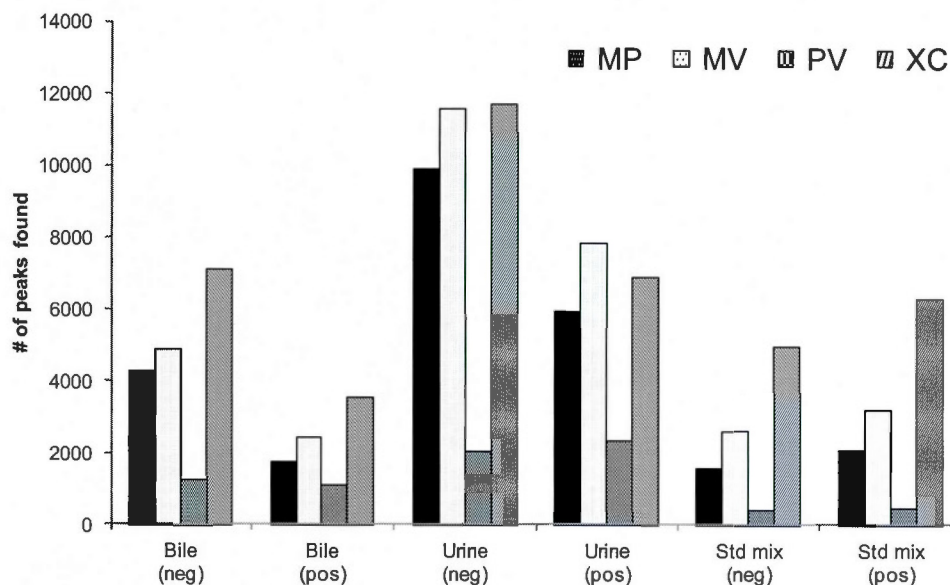


Figure 2.4 Number of peaks found by different software: PeakView (PV), MarkerView (MV), MetabolitePilot (MP) and XCMS online for each sample type (bile, urine and standard mixture) in both positive and negative modes

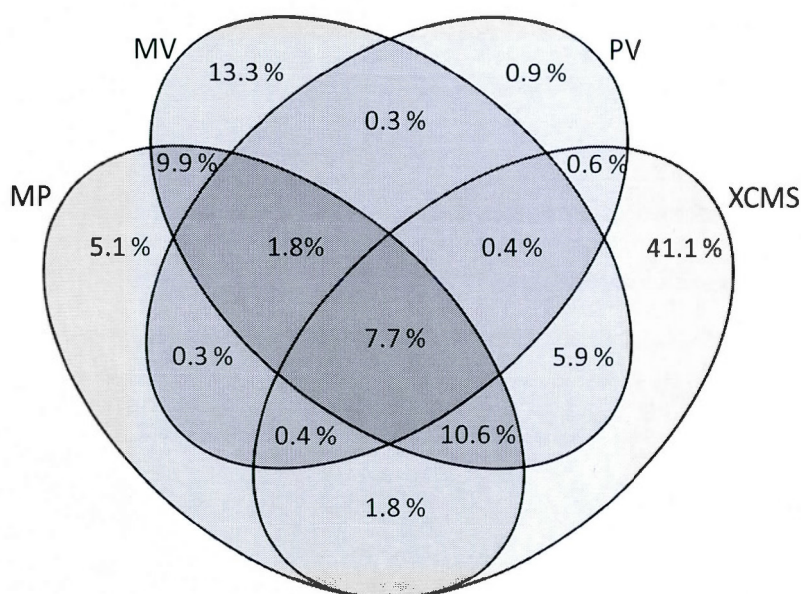


Figure 2.5 Venn diagram representation of the average percent of overlaps between the results of four peak picking workflows; MetabolitePilot (MP), MarkerView (MV), PeakView (PV) and XCMS online

Precursor ion m/z values (MH^+ or MH^-) of all peaks were searched using the METLIN online metabolomics database (features with at least one hit in database (within 5 ppm) labeled as “metabolites” and at least one MS/MS spectra in Metlin labeled as “MSMS metabolites”). The average percent of “metabolites” and “MSMS metabolites” are presented in Figure 2.6. Features found only by XCMS online, had the lowest percentage of metabolites in METLIN, with an average of 33% while peaks found only by PeakView and MetabolitePilot both had 53% with at least one corresponding metabolite (within 5 ppm) in the database. Although it was expected that the percent of metabolites and MSMS metabolites increase by the number of overlaps (between two, three or four workflows), no evident relationship was found for accurate mass matching. It is probable that not all metabolites detected in these samples are present in the METLIN database. It was observed that the overlap between MetabolitePilot and PeakView had the highest percentage of found metabolites and “MSMS metabolites”. This could indicate that their implemented algorithms for peak finding are the most similar.

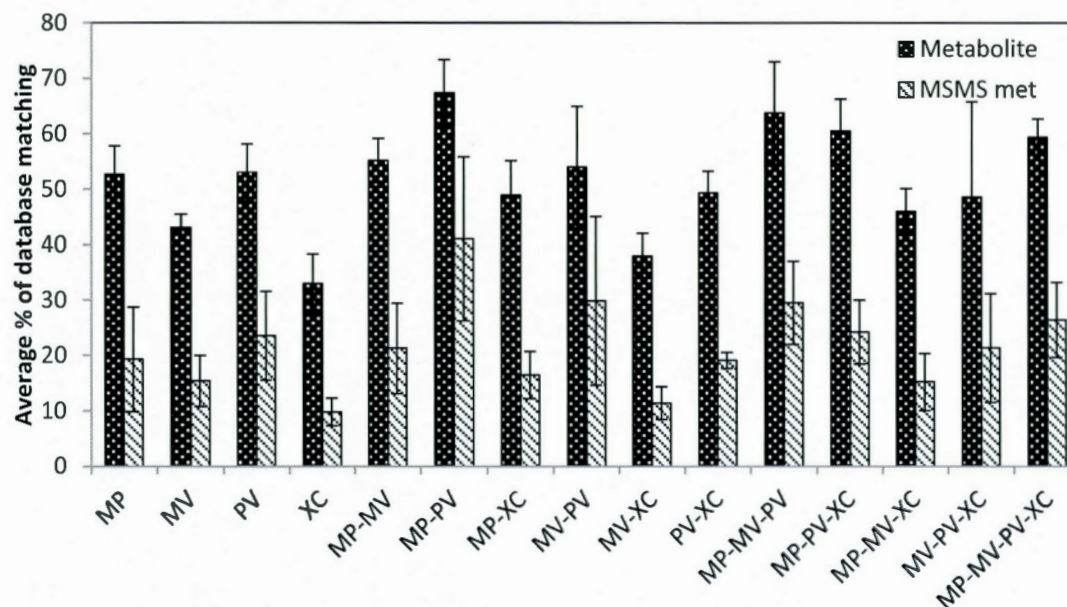


Figure 2.6 Average % of peaks with at least one hit in METLIN database (metabolites) or at least one hit with MSMS spectra (MSMS metabolites) for peaks in each region of Venn diagrams (average includes data from bile and urine in both positive and negative modes). Acronyms shown indicate MetabolitePilot (MP), MarkerView (MV), PeakView(PV) and XCMS online (XC)

The percentages of tentatively identified metabolites for different samples from each workflow are shown as supplementary data (Figure S2). It is demonstrated that PV yielded the highest % database matching (but with the shortest peak list). It also shows that a higher proportion of peaks are found as metabolites and MSMS metabolites in positive mode, potentially as a function of more data being accumulated in positive mode and therefore more chance of these metabolites being present in metabolomics databases, such as METLIN. It is also seen that the urine sample contains more identifiable metabolites, compared to the bile sample.

In a subsequent step, MS/MS spectral matching was performed from the results of information-dependent acquisition (IDA) triggered precursor ions. Confirmed metabolites having MS/MS spectral matching METLIN score > 60 are shown in supplementary data (Figure S3). It was observed that almost all tentatively

identified metabolites are found either by MetabolitePilot or MarkerView (with or without overlap with others). These results also indicated a higher probability of MS/MS matching for peaks present in the 4-way overlap of Venn diagrams. The results of MS/MS spectral matching correlates with accurate mass matching results in terms of higher number of identified metabolites in urine versus bile, and in positive mode versus negative mode. It must be noted that the generic IDA criteria used for triggering MS/MS resulted in a low % of peaks found to have high quality MS/MS spectra associated to them. A higher proportion of fragmentation spectra would be possible if inclusion lists were used for subsequent analysis based on m/z values found following peak picking workflows. This would however necessitate an extra LC-MS injection to acquire this data. Tentatively identified metabolites from biological samples (urine and bile) have also been compiled into Supplemental Tables S6-S9.

In a separate analysis the results of peak picking workflows for both biological samples and standard mixture were evaluated to investigate the possible effect of retention time in peak detection efficiency. The number of peaks found by different peak detection workflows was calculated for five equal portions of chromatogram (0-5 min, 5-10 min, *etc.*). Although the percentage of the peaks detected in different parts of chromatogram was sample dependent, no significant difference among peak detection workflow was observed for the percentage of peaks found in any parts of chromatogram. In a parallel analysis, the peaks located in different portions of venn diagrams was evaluated in terms of any special trend in their retention time information. It was observed that the peaks found in the 20-25 minutes retention time windows have slightly higher probability to be found by only one workflow (Figure S4). It could be due to the column bleeding which occurs more in this region of the chromatogram.

To have a better comparison between the tested workflows, the practical advantages and disadvantages of each were also considered. MarkerView and XCMS online yield results in the fastest time. PeakView and MetabolitePilot allow visual inspection of extracted ion chromatograms for all detected peaks. XCMS online also provide visual inspection of peaks, however, it is very time consuming, as each XIC graphics is stored in separate files. MarkerView and XCMS online are able to identify isotopes and remove them easily. XCMS online has the great advantage of being free and open source, however, this does leave the user at the mercy of proper updating and maintenance of the system. MetabolitePilot had a limitation of 1000 peaks for each peak generating run and can therefore be very time-consuming, especially when dealing with complex biological samples. This is likely a function of the fact that MetabolitePilot was developed mainly for investigating metabolism of drugs or other xenobiotics where criteria can be added for filtering metabolites of interest based on structural similarities to the parent compound and also controls can be used to easily remove “interfering” endogenous metabolites from the resulting lists in drug metabolism workflows. MarkerView, MetabolitePilot and XCMS have the ability to directly filter the resulting peak lists compared to controls in terms of fold change and statistical significance in a streamlined manner while manual comparison between sample and control is necessary in PeakView. An ideal workflow would integrate the ability to filter out peaks based on possible in-source adducts and fragment ions, neither of which was possible easily with the tested software.

In this study, main processing parameters including m/z tolerance, retention time window and intensity threshold were selected identical to fairly perform the comparison. Each peak detection workflows also have its own parameters (depending on the processing algorithm) which were selected as defaults values. Hence, the performance of peak detection workflows may be slightly higher that

what is shown in this study as processing parameters used may differ from optimized values.

The challenge for selection of appropriate peak detection could be approached in different ways. For instance, an ideal situation would be to use peak detection workflow with highest performance however the comparison of all peak detection workflows is not readily available and not all peak picking workflows offer similar processing options needed for researcher. In addition, compatibility of the data format and peak detection workflow, availability (in case of commercial software) as well as the need for special knowledge (for example, programming skills in R based XCMS), and friendly graphic interface may affect the workflow selected by researcher. Another approach would be to combine peak detection data from multiple tools and assign peaks with statistical scores based on the number of tools detect them. The results of our study shows that the peaks detected by more than one workflow have higher potential to be identified by accurate mass as well as MS/MS spectrum matching. On the other hand, peaks that are detected by only one peak detection workflow still may contain important biological information although with lower chance.

Among the studied peak picking workflows, MarkerView showed a better recovery ratio for standard compounds as well as having larger overlaps with other peak generating workflows for complex biological samples. MarkerView's performance could be due to employing three replicate samples to perform t-test, (to be similar with other software compared in this paper). This result is in agreement with the importance of alignment process which found matching peaks through multiple samples in metabolomics studies.

2.5 Conclusions

Four peak generating software were evaluated for untargeted peak picking in a metabolomics workflow. The performance of peak picking workflows was shown to have a direct impact on the final results. Vast differences in resulting peak lists were observed when different peak picking strategies on identical LC-HRMS data from complex urine and bile samples as well as a standard metabolite mix was used. Among the studied peak picking workflows, MarkerView showed a better recovery ratio for standard compounds as well as having larger overlaps with other peak generating workflows for complex biological samples. In addition, a targeted approach was performed on the standard mixture and it was shown that there were a number of metabolites undetectable by all used peak detection workflows.

Supplementary data for chapter 2

Table S1 Compounds used in standard mixture for evaluating the four peak picking workflows

#	Compound	Formula	Exact mass (Da)	#	Compound	Formula	Exact mass (Da)
1	Pyruvic acid	C ₃ H ₄ O ₃	88.0160	43	Methionine sulfoxide	C ₅ H ₁₁ NO ₃ S	165.0460
2	Putrescine	C ₄ H ₁₂ N ₂	88.1000	44	Phenylalanine	C ₉ H ₁₁ NO ₂	165.0790
3	Alanine	C ₃ H ₇ NO ₂	89.0477	45	Cysteic acid	C ₃ H ₇ NO ₅ S	169.0045
4	Lactic Acid	C ₃ H ₆ O ₃	90.0317	46	Pyridoxine	C ₈ H ₁₁ NO ₃	169.0739
5	Acetoacetic acid	C ₄ H ₆ O ₃	102.0317	47	α -Glycerol-phosphate	C ₃ H ₉ O ₆ P	172.0137
6	Serine	C ₃ H ₇ NO ₃	105.0426	48	Arginine	C ₆ H ₁₄ N ₄ O ₂	174.1117
7	Cytosine	C ₄ H ₅ N ₃ O	111.0433	49	Citrulline	C ₆ H ₁₃ N ₃ O ₃	175.0957
8	Histamine	C ₅ H ₉ N ₃	111.0796	50	Inositol	C ₆ H ₁₂ O ₆	180.0634
9	Uracil	C ₄ H ₄ N ₂ O ₂	112.0273	51	Tyrosine	C ₉ H ₁₁ NO ₃	181.0739
10	Proline	C ₅ H ₉ NO ₂	115.0633	52	Phosphoserine	C ₃ H ₈ NO ₆ P	185.0089
11	Valine	C ₅ H ₁₁ NO ₂	117.0790	53	Kynurenic acid	C ₁₀ H ₇ NO ₃	189.0426
12	Succinic Acid	C ₄ H ₆ O ₄	118.0266	54	Glycyl-aspartic acid	C ₆ H ₁₀ N ₂ O ₅	190.0590
13	Threonine	C ₄ H ₉ NO ₃	119.0582	55	Quinic acid	C ₇ H ₁₂ O ₆	192.0634
14	Phenethylamine	C ₈ H ₁₁ N	121.0891	56	Phosphothreonine	C ₄ H ₁₀ NO ₆ P	199.0246
15	Nicotinamide	C ₆ H ₆ N ₂ O	122.0480	57	Spermine	C ₁₀ H ₂₆ N ₄	202.2157
16	Nicotinic acid	C ₆ NH ₅ O ₂	123.0320	58	Tryptophan	C ₁₁ H ₁₂ N ₂ O ₂	204.0899
17	Thymine	C ₅ H ₆ N ₂ O ₂	126.0429	59	Ibuprofen	C ₁₃ H ₁₈ O ₂	206.1307
18	Pyroglutamic acid	C ₅ H ₇ NO ₃	129.0426	60	Kynurenine	C ₁₀ H ₁₂ N ₂ O ₃	208.0848
19	Agmatine	C ₅ H ₁₄ N ₄	130.1218	61	Phosphocreatine	C ₄ H ₁₀ N ₃ O ₅ P	211.0358
20	Hydroxyproline	C ₅ H ₉ NO ₃	131.0582	62	S-Benzyl-cysteine	C ₁₀ H ₁₃ NO ₂ S	211.0667
21	Leucine	C ₆ H ₁₃ NO ₂	131.0946	63	Atrazine	C ₈ H ₁₄ ClN ₅	215.0938
22	Methyl succinate	C ₅ H ₈ O ₄	132.0423	64	Taurocholic acid	C ₂₆ H ₄₅ NO ₇ S	515.2917
23	Asparagine	C ₄ H ₈ N ₂ O ₃	132.0535	65	Trp methyl ester	C ₁₂ H ₁₄ N ₂ O ₂	218.1055

24	Ornithine	$C_5H_{12}N_2O_2$	132.0899	66	<i>N</i> -Acetyl-mannosamine	$C_8H_{15}NO_6$	221.0899
25	Aspartic Acid	$C_4H_7NO_4$	133.0375	67	Cystathionine	$C_7H_{14}N_2O_4S$	222.0674
26	Adenine	$C_5H_5N_5$	135.0545	68	Thymidine	$C_{10}H_{14}N_2O_5$	242.0903
27	Anthranilic acid	$C_7H_7NO_2$	137.0477	69	Cytidine	$C_9H_{13}N_3O_5$	243.0855
28	Tyramine	$C_8H_{11}NO$	137.0841	70	Uridine	$C_9H_{12}N_2O_6$	244.0695
29	Glutamine	$C_5H_{10}N_2O_3$	146.0691	71	Deoxyadenosine	$C_{10}H_{13}N_5O_3$	251.1018
30	Lysine	$C_6H_{14}N_2O_2$	146.1055	72	Adenosine	$C_{10}H_{13}N_5O_4$	267.0968
31	Glutamic acid	$C_5H_9NO_4$	147.0532	73	Inosine	$C_{10}H_{12}N_4O_5$	268.0808
32	Methionine	$C_5H_{11}NO_2S$	149.0511	74	<i>N</i> -Phthaloyl-Glu	$C_{13}H_{11}NO_6$	277.0586
33	Guanine	$C_5H_5N_5O$	151.0494	75	Guanosine	$C_{10}H_{13}N_5O_5$	283.0917
34	Xanthine	$C_5H_4N_4O_2$	152.0334	76	17 α -Ethinylestradiol	$C_{20}H_{24}O_2$	296.1776
35	<i>p</i> -OH phenylacetic acid	$C_8H_8O_3$	152.0473	77	Diclofenac	$C_{14}H_{11}Cl_2NO_2$	295.0167
36	3-OH anthranilic acid	$C_7H_7NO_3$	153.0426	78	CMP	$C_9H_{14}N_3O_8P$	323.0519
37	Octopamine	$C_8H_{11}NO_2$	153.0790	79	UMP	$C_9H_{13}N_2O_9P$	324.0359
38	Histidine	$C_6H_9N_3O_2$	155.0695	80	AMP	$C_{10}H_{14}N_5O_7P$	347.0631
39	Orotic acid	$C_5H_4N_2O_4$	156.0171	81	GMP	$C_{10}H_{14}N_5O_8P$	363.0580
40	α -amino-caprylic acid	$C_8H_{17}NO_2$	159.1259	82	Deoxycholic acid	$C_{24}H_{40}O_4$	392.2927
41	Carnitine	$C_7H_{15}NO_3$	161.1052	83	Cholic acid	$C_{24}H_{40}O_5$	408.2876
42	Ethionine	$C_6H_{13}NO_2S$	163.0667	84	Canthaxanthin	$C_{40}H_{52}O_2$	564.3967

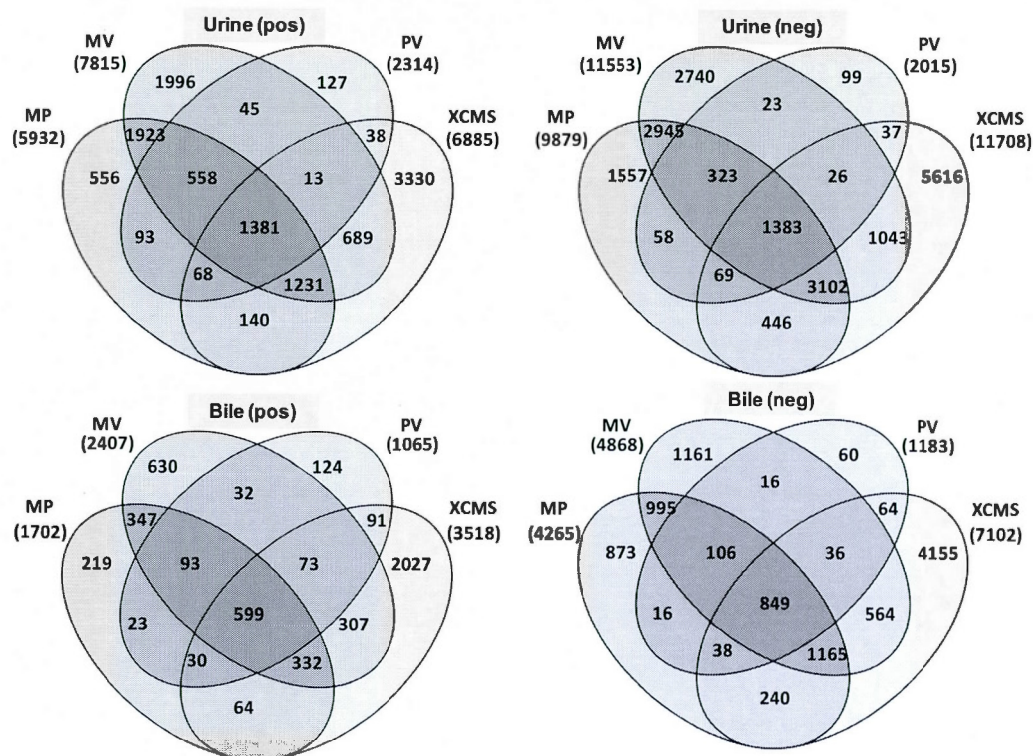


Figure S1 Venn diagram representation of the overlaps between the results from four peak picking software (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online) used to filter LC-MS data from bile and urine in positive and negative modes

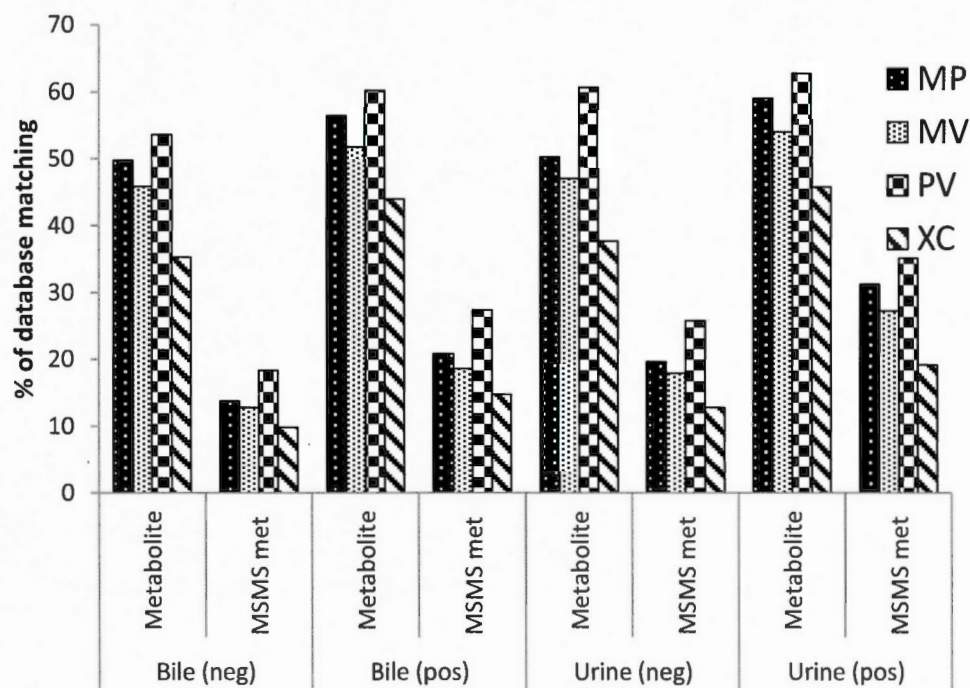


Figure S2 Percentage of metabolome database matching for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from bile and urine sample in positive and negative modes, peaks with at least one hit in METLIN database are shown as metabolites and those with at least one MSMS spectra (to be inspected in Metlin) are presented as MS/MS met

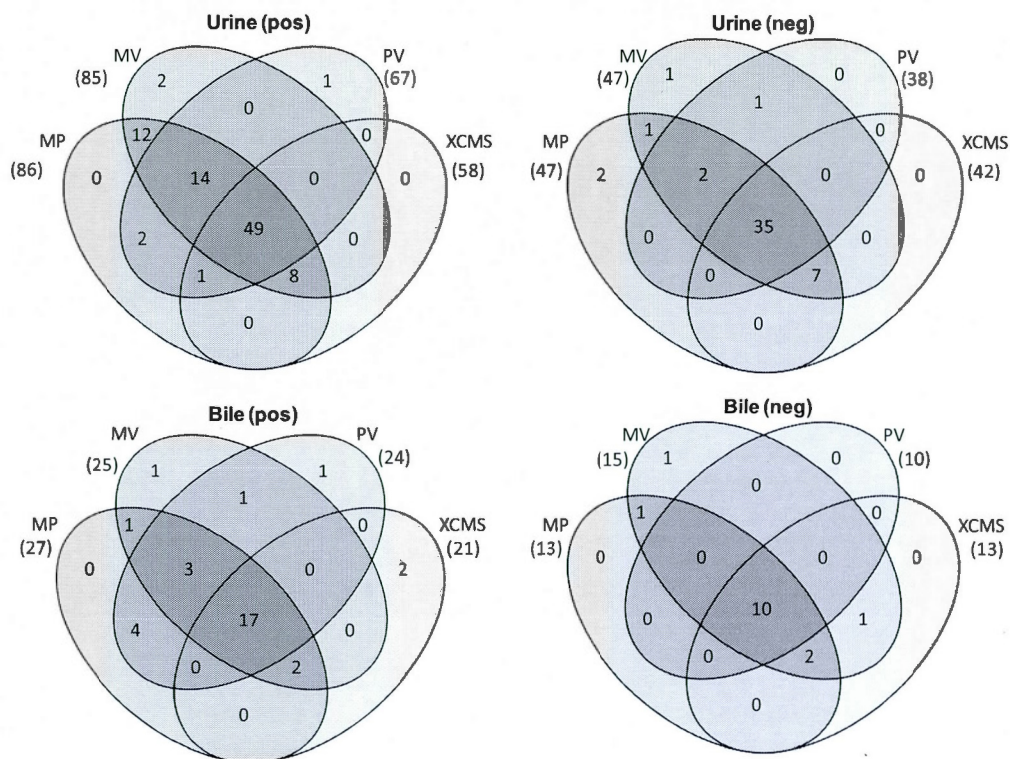


Figure S3 Venn diagram representation of the overlaps between the results of MS/MS spectrum match (METLIN score>60) from four peak picking software MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online on raw LC-MS data from bile and urine in positive and negative modes

Table S2 Metabolites found by each peak picking workflow from standard mixture in positive ionization mode (confirmed with MS/MS match from METLIN database)

Metabolite	RT (min)	Found by			
		MP	MV	PV	XC
Histidine	1.6	-	✓	-	-
Cytidine	2.6	-	-	-	✓
UMP	1.9	✓	✓	-	-
Cytosine	1.6	-	✓	-	✓
Tyramine	2.7	✓	✓	✓	-
Uridine	2.4	✓	✓	✓	-
Leucine	2.8	✓	✓	✓	-
Uracil	1.9	✓	✓	✓	-
Proline	1.6	-	✓	✓	✓
Phenylalanine	5.0	-	✓	✓	✓
Kynurenine	4.5	-	✓	✓	✓
Tryptophan	7.5	✓	✓	✓	✓
Tyrosine	2.4	✓	✓	✓	✓
Adenine	3.9	✓	✓	✓	✓
Guanine	3.7	✓	✓	✓	✓
Guanosine	3.7	✓	✓	✓	✓
Thymidine	5.8	✓	✓	✓	✓
AMP	2.3	✓	✓	✓	✓
Pyridoxine	2.2	✓	✓	✓	✓
Anthranilic acid	9.4	✓	✓	✓	✓
Kynurenic acid	9.0	✓	✓	✓	✓
Thymine	3.2	✓	✓	✓	✓
Xanthine	2.5	✓	✓	✓	✓
GMP	2.4	✓	✓	✓	✓

Table S3 Metabolites found by each peak picking workflow from standard mixture in negative ionization mode (confirmed with MS/MS match from METLIN database)

Metabolites	RT (min)	Found by			
		MP	MV	PV	XC
Inositol	1.6	-	✓	-	✓
Quinic acid	1.6	-	-	✓	✓
17 α _Ethinylestradiol	19.4	✓	✓	✓	-
Cytidine	1.8	✓	✓	-	✓
Pyridoxine	2.2	✓	✓	-	✓
Deoxycholic acid	21.8	✓	✓	✓	✓
Cholic acid	21.1	✓	✓	✓	✓
Tyrosine	2.7	✓	✓	✓	✓
Diclofenac	20.5	✓	✓	✓	✓
Uridine	2.5	✓	✓	✓	✓
Kynurenine	4.5	✓	✓	✓	✓
Thymidine	5.8	✓	✓	✓	✓
Guanosine	3.8	✓	✓	✓	✓

Table S4 Metabolites identified by targeted approach in standard mixture (pos) with METLIN MS/MS matching which had not been detected by any of the automated peak detection workflows

Metabolites	RT (min)
Glutamic acid	1.6
Carnitine	1.6
Octopamine	1.7
Valine	1.8
CMP	1.8
Methionine	2.0
Nicotinic acid	2.0
Adenosine	3.6
2-deoxyadenosine	3.9
3-OH anthranilic acid	5.6
Phenylethylamin	5.9
Atrazine	17.9

Table S5 Metabolites identified by targeted approach in standard mixture (neg) with METLIN MS/MS matching which had not been detected by any of the automated peak detection workflows

Metabolites	RT (min)
Orotic acid	1.7
CMP	1.8
AMP	2.4
GMP	2.8
Adenosine	3.4
Tryptophan	7.5
Kynurenic acid	9.1
N-phthaloyl-Glu	11.7
Ibuprofen	20.7

Table S6 Identified metabolites in urine (pos) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites, all of which are presented here). Metabolites found in standard mixture with matching retention times are assigned with a star

<i>m/z</i>	RT (min)	metabolite	<i>m/z</i>	RT (min)	metabolite
95.0161	1.8	Dimethyl sulfone	165.0542	5.9	p-Coumaric acid
100.0756	3.8	δ -Valerolactam	165.0543	2.6	p-Coumaric acid
100.1124	4.2	Cyclohexylammonium	166.0719	2.8	Methylguanine
113.0595	4.3	Sorbic acid	166.0857	5.1	Phenylalanine*
115.0752	9.9	3-Methyl-4-pentenoic acid	167.0703	10.1	3-(2-OH-phenyl)propionic acid
115.0752	9.9	4-Hydroxy hexenal	167.0703	10.1	Isopenol
115.0752	9.9	γ -Caprolactone	168.1012	4.1	O-Methyldopamine
115.0752	9.9	δ -Hexalactone	171.0647	7.9	3,4-Dihydroxyphenyl glycol
116.0703	2.0	Aminocyclobutane carboxylic acid	175.0861	5.6	Indole-3-acetamide
122.0265	9.2	Cysteine	175.0963	12.6	Suberic acid
127.0389	3.2	Larixinic acid	177.0906	17.3	5,6,7,8-Tetrahydro-2-naphthoic acid
127.0389	3.2	4-Hydroxy-6-methylpyran-2-one	180.0516	2.9	Isoxanthopterin
127.0497	1.7	Thymine	181.0606	7.5	Nicotinuric acid
130.0498	7.9	2-Pyrrolidone-5-carboxylic acid	188.0700	15.5	3-Amino-2-naphthoic acid
130.0498	7.9	Pyroglutamic acid	188.0700	15.5	Indoleacrylic acid
130.0500	3.7	2-Pyrrolidone-5-carboxylic acid	188.0700	15.5	Genkwanin
130.0500	3.7	Pyroglutamic acid	188.0700	15.5	Wogonin
130.0500	4.9	2-Pyrrolidone-5-carboxylic acid	188.0700	15.5	Glycitein
130.0500	4.9	Pyroglutamic acid	188.0700	15.5	Biochanin A
132.1015	2.8	Leucine* (Iso-, Nor-, Allo-leucine)	188.0707	7.5	Indoleacrylic acid
137.0452	2.2	Hypoxanthine	189.1225	6.3	Gly Leu/Ile (or Leu/Ile Gly)
137.0452	2.2	Allopurinol	189.1231	1.9	N-(ϵ or α)-Acetyl-lysine
137.0452	6.2	Hypoxanthine	190.071	2.4	N-Acetyl-glutamic acid
137.0452	6.2	Allopurinol	190.1178	1.7	N6-Carbamoyl-Lysine
138.0547	1.6	<i>p</i> -Aminobenzoic acid	191.1022	3.0	2,6-Diaminoheptanedioate
138.0547	1.6	2-Pyridylacetic acid	192.0655	10.6	5-Hydroxyindoleacetic acid
139.0387	5.9	3,4-	193.0495	10.9	5,7-Dihydroxy-4-

Dihydroxybenzaldehyde			methylcoumarin		
146.0594	6.3	Isoquinoline <i>N</i> -oxide	195.0651	11.1	Erbstatin analog
146.0597	10.6	Isoquinoline <i>N</i> -oxide	195.0651	11.1	Scytalone
146.0598	8.2	Isoquinoline <i>N</i> -oxide	195.0761	5.8	Aminohippuric acid
147.0439	7.6	Coumarin	206.0448	8.5	Xanthurenic acid
147.0442	10.4	Coumarin	209.0805	10.3	Dimethylcaffeic acid
149.0593	5.1	trans-Cinnamic acid	209.0916	4.5	Kynurenine*
149.0594	9.0	trans-Cinnamic acid	211.1330	11.0	Jasmonic acid
149.0958	11.3	Cuminaldehyde	215.1272	15.2	(3-Me-cyclohexane-diyl)diacetic acid
151.0747	14.3	m-Tolylacetic acid	215.1385	4.7	Pro Val
151.0747	14.3	3,4-Dimethylbenzoic acid	218.1382	3.1	Propionyl-carnitine
151.0747	14.3	2-Phenylpropionic acid	224.0914	8.0	Acetyl-tyrosine
151.0749	13.8	3,4-Dimethylbenzoic acid	231.1586	15.1	Dodecanedioic acid
151.075	13.1	3,4-Dimethylbenzoic acid	233.1133	3.3	Asp Val
153.0403	5.1	Xanthine	247.1083	12.3	N-Acetyl-tryptophan
153.0908	11.3	4-Hydroxy nonenal alkyne	247.1289	7.0	Leu/Ile Asp (or Asp Leu/Ile)
154.0492	3.5	Aminosalicylic Acid	252.1082	3.7	Deoxyadenosine
154.0496	4.7	Aminosalicylic Acid	259.0919	1.7	5-Methyluridine
154.0496	4.7	3-Hydroxyanthranilic acid	268.1034	3.4	Adenosine
154.0498	6.9	Aminosalicylic Acid	268.1034	3.4	Vidarabine
154.0498	6.9	3-Hydroxyanthranilic acid	271.0600	17.4	Galangin
154.0971	1.8	N ω -Acetylhistamine	285.0757	16.5	Prunetin
155.0697	11.5	2,6-Dihydroxy-4-methoxytoluene	285.0757	16.5	Acacetin
155.0697	11.5	3,4-Dihydroxyphenyl ethanol	290.1349	4.1	Asp Gly Val
155.0701	7.9	2,6-Dihydroxy-4-methoxytoluene	295.129	9.5	Glu Phe (or Phe Glu)
156.0766	6.5	Histidine	297.1075	5.1	Asp Tyr
157.1217	17.0	4-Hydroxy nonenal	298.1145	5.6	2-Methylguanosine
158.0965	10.1	8-Isoquinoline methanamine	298.1145	5.6	Nelarabine
160.0750	12.0	1-Acetylindole	299.1853	18.3	13,14-dihydro-15-keto-tetranor PGD2
160.0750	12.0	Indoleacetaldehyde	299.1853	18.3	13,14-dihydro-15-keto-tetranor PGE2
161.0803	9.9	3-Methyladipic acid	338.1334	8.5	Asp Gly Phe (Gly Asp Phe)
161.0803	9.9	Pimelic acid	354.1280	5.3	Gly Asp Tyr (or YDG/DYG)
161.0803	9.9	3,3-Dimethylglutaric acid	447.0910	10.8	Baicalin

Table S7 Identified metabolites in urine (neg) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here). Metabolites found in standard mixture with matching retention times are assigned with a star

<i>m/z</i>	RT (min)	Metabolite	<i>m/z</i>	RT (min)	metabolite
131.0352	5.5	Dimethylmalonic acid	172.9913	5.9	4-Hydroxybenzenesulfonic acid
131.0352	5.5	Methylsuccinic acid	173.0824	12.8	Suberic acid
131.0352	5.5	Glutaric acid	175.0611	8.9	2-Isopropylmalic acid
137.0250	7.0	Salicylic acid	181.0505	7.6	3-Methylorsellinic acid
138.0201	2.8	3 (or 6)-Hydroxypicolinic acid	181.0507	6.0	3,4-Dihydroxyphenylpropanoate
144.0460	8.3	Isoquinoline N-oxide	181.0507	6.0	Flopropione
145.0506	7.2	3-Methylglutaric acid	187.0978	17.9	Nonic acid
145.0506	7.2	Adipic acid	191.0202	2.1	Citric acid
145.0506	7.2	2,2-Dimethyl Succinic acid	193.0617	5.9	Aminohippuric acid
151.0265	2.5	Oxypurinol	194.0459	6.3	Salicyluric acid
151.0265	2.5	Xanthine	194.0467	10.5	Salicyluric acid
151.0402	6.6	4-Hydroxy-3-methylbenzoic acid	195.0659	11.7	Homoveratric acid
151.0402	6.6	Hydroxyphenylacetic acid	195.0666	9.2	Homoveratric acid
151.0402	6.6	p-Anisic acid	197.0453	6.8	Syringic acid
151.0402	6.6	3-Cresotinic acid	197.0453	6.8	2-Hydroxy-3,4-dimethoxybenzoic acid
151.0402	6.6	Mandelic acid	201.1135	17.9	Sebacic acid
155.0098	2.4	Orotic acid	203.0824	7.6	Tryptophan
157.0368	1.6	Allantoin	204.0662	13.1	Cinnamoylglycine
159.0668	10.0	Pimelic acid	204.0671	12.1	3-Indolelactic acid
159.0668	10.0	3-Methyladipic acid	212.0028	7.1	Indoxylsulfuric acid
164.0357	12.4	N-Formylanthranilic acid	213.1130	17.6	3-Me-cyclohexane-1,1-diyl diacetic acid
164.0579	2.3	Methylguanine	223.0611	9.3	Sinapic acid
164.0718	5.2	Phenylalanine	229.1442	17.5	Dodecanedioic acid
165.0559	9.7	Tropic acid	243.0616	1.9	Uridine*
165.0559	9.7	Dihydro-3-coumaric acid	243.1598	20.0	Undecanedicarboxylic acid
165.0559	9.7	Atrolactic acid	245.0935	12.5	N-Acetyl-tryptophan
165.0559	9.7	3-(2-OHphenyl)propionic acid	253.0503	16.2	Daidzein
167.0215	2.1	Uric acid	255.0663	15.1	Isoliquiritigenin
167.0349	6.8	Homogentisic acid	257.1757	20.5	Tetradecanedioic acid

167.0350	4.8	Dihydroxyphenylacetic acid	263.0227	2.7	3-OMe-4-OHphenylethylene glycol sulfate
167.0350	4.8	Homogentisic acid	296.0996	5.7	Methylguanosine
172.0981	11.4	Acetyl-Leucine	308.0984	1.6	N-Acetylneuraminic Acid

Table S8 Identified metabolites in bile (pos) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here). Matched metabolite with standard mixture is assigned with a star

<i>m/z</i>	RT	metabolite	<i>m/z</i>	RT (min)	metabolite
93.0547	1.6	Glycerol	227.1139	9.3	Carnosine
130.0500	1.9	2-Pyrrolidone-5-carboxylic acid	227.1139	9.3	His Ala
130.0500	1.9	Pyroglutamic acid	255.0660	13.5	Daidzein
136.0617	3.5	Adenine*	255.0660	13.5	3,7-Dihydroxyflavone
137.0459	3.7	Hypoxanthine	258.1083	2.3	5-Methylcytidine
137.0459	3.7	Allopurinol	269.0886	3.7	Inosine
149.0594	9.3	trans-Cinnamic acid	271.0602	14.8	Galangin
151.0750	16.5	m-Cresyl acetate	273.0754	11.1	Naringenin
151.0750	16.5	3,4-Dimethylbenzoic acid	285.0760	13.8	Prunetin
156.0761	9.3	Histidine	285.0760	13.8	Acacetin
166.0718	2.3	Methylguanine	285.0760	13.8	Genkwanin
180.0516	3.0	Isoxanthopterin	285.0760	13.8	Wogonin
180.0652	8.4	Hippuric acid	285.0760	13.8	Glycitein
183.0515	6.1	1-Methyluric acid	285.0760	13.8	Biochanin A
188.0701	12.5	3-amino-2-naphthoic acid	289.2160	16.4	trans-Dehydroandrosterone
188.0701	12.5	Indoleacrylic acid	298.1145	5.6	Methylguanosine
194.0805	9.4	Phenylacetyl glycine	298.1145	5.6	Nelarabine
194.0805	9.4	Methylhippuric acid	417.1176	11.7	Daidzin
198.0864	1.7	N-Acetyl-L-Histidine	447.0933	13.0	Baicalin
206.0452	8.6	Xanthurenic acid	449.1077	9.9	Naringenin-O-β-Glucuronide
220.1181	6.4	Pantothenic Acid	466.3146	20.4	Glycocholic Acid
225.1958	19.1	N,N'-Dicyclohexylurea	500.3033	21.9	Cholanic acid diol sulphoethylamide
227.1139	9.3	Ala His	516.2992	19.5	Taurocholic acid

Table S9 Identified metabolites in bile (neg) with METLIN MS/MS matching score of higher than 60 (each individual peak might results in several possible metabolites which are all presented here)

<i>m/z</i>	RT (min)	metabolite	<i>m/z</i>	RT (min)	metabolite
137.0252	6	Salicylic acid	266.0894	3.5	Vidarabine
164.0581	2.3	Methylguanine	266.0894	3.5	Deoxyguanosine
167.0216	2.1	Uric acid	267.0729	3.8	Inosine
167.0353	7.0	5-Methoxysalicylic acid	283.0611	13.9	Phycion
172.0979	11.4	Acexamic acid	283.0611	13.9	Prunetin
172.0979	11.4	Acetyl-leucine	283.0611	13.9	Glycitein
178.0372	3.0	xanthopterin	283.0611	13.9	Wogonin
191.0199	2.2	Citric acid	283.0611	13.9	Acacetin
204.0301	8.7	Xanthurenic acid	283.0611	13.9	Biochanin A
212.0023	7.3	Indoxylsulfuric acid	296.1000	5.7	Methylguanosine
225.0992	7.7	Carnosine	346.0551	2.5	Adenosine 3'-monophosphate
263.0222	2.8	3-OMe4-OHphenylethyleneglycolsulfate			

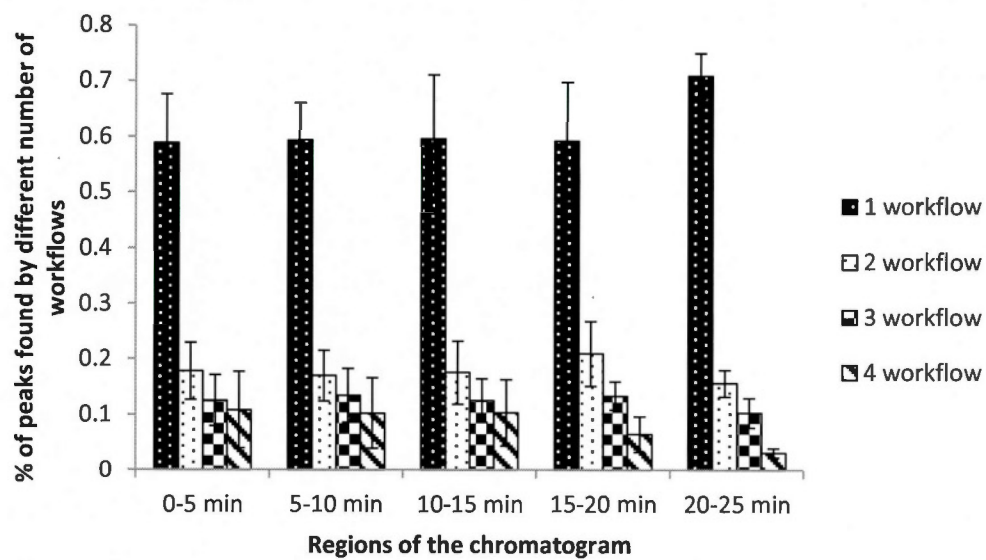


Figure S4 The average percent of peaks found by different number of workflows in each region of chromatogram

CHAPTER III

ENHANCING UNTARGETED METABOLOMIC DATA ANALYSIS BY A NOVEL DATA REDUCTION WORKFLOW

3.1 Abstract

Data analysis is a crucial step in many bioanalytical workflows, including untargeted mass spectrometry-based metabolomics. This branch of metabolomics deals with tens of hundreds (or thousands) of features from raw data with the eventual goal of detecting molecules involved in important biological pathways and/or biomarkers of disease. In untargeted metabolomics research, the data processing usually consists of two general steps, peak list generation and metabolite identification. Most peak-generating workflows are incapable of distinguishing protonated intact molecules from adducts, for example, making feature identification complicated and time-consuming. A MATLAB-based workflow designed to remove isotopes, radical ions, adducts and in-source fragments from a raw feature list in order to have a higher proportion of intact protonated ions in the resulting filtered list is presented. It imports data (in excel-compatible format) from any peak generating workflow (in both positive and negative ionization modes), applies different processing steps and results in a more condensed and reliable feature list. Four peak picking workflows, using namely PeakView®, Markerview™, MetabolitePilot and XCMS online, were evaluated in terms of number of peaks which could be filtered out, and thus be deemed as “redundant” features, using the developed *DataReduction* workflow.

3.2 Introduction

MS-based untargeted metabolomics data analysis begins by treating raw data with peak-picking algorithms (also known as peak or feature detection). There are two

main challenges for this step. The first is to select the appropriate approach, since a wide variety of software packages are available for this purpose. A comparison was done of four peak picking workflows and the results were presented in Chapter 2 of this thesis.

Due to the fact that each neutral molecule has the potential of being observed at several m/z values in MS-based metabolomics, the second challenge is to identify and remove redundant peaks from peak picking results, namely non-mono-isotopic ions, in-source fragments, multiply charged species, adduct and cluster ions (Katajamaa and Orešič 2007) (Kuhl *et al.* 2011). Thus, it should be considered that not all observed m/z values detected correspond to unique metabolites. Considering the fact that most software packages are unable to identify these types of redundant peaks, a non-filtered mass-based search from initial peak picking results can cause false identification of metabolites (Varghese *et al.* 2012). Depending on the sample type and ionization mode, the number of molecules forming multiple ion species is different. It was demonstrated by Brown *et al.*, that depending on sample type and mass spectrometry method used, between 14% and 33% of metabolites could be observed at more than one m/z . For placental footprints, 1 in 3 or 33% metabolites were detected as multiple ions in one analytical run (Brown *et al.* 2009).

Several attempts have been made to develop workflows for ion annotation. For instance, CAMERA, a freely-available R-based package, performs ion-annotation on peak picking results from R-based XCMS. It uses retention time and similarity between peak shapes to group correlated peaks. It then calculates the difference between m/z values for each peak pair within the groups and compares it to certain m/z relationships often seen during ionization (Kuhl *et al.* 2011). Pre-set m/z differences, retention time and intensity correlation were used in another study to identify redundant peaks and resulted in 50% data reduction. In this approach, R-based XCMS was used for deconvolution of raw data in combination with esi

program to write peak output files to an annotated version (Brown *et al.* 2009). There are also some commercially-available software for this, such as ACD/IntelliXtract (a part of ACD/MS work suite) based on a given rule table to annotate ion species (ACD/IntelliXtract 2007). Publicly-available software PUTMEDID-LCMS, a tool operating in Taverna environment, generates pair-wise peak correlations, annotates features to group different ion types of the same metabolite based on mass differences, similar retention times and correlation coefficient between peak responses. For this workflow, Raw data were converted to the NETCDF format and R-based XCMS was used for deconvolution of data (Brown *et al.* 2011). IDEOM is a freely-available package for Microsoft Excel for peak annotation of ESI redundant peaks as well as FT or ringing signals. It uses retention time, peak shape and correlation of peak intensities (Creek *et al.* 2012).

Although, there are a number of workflows for peak annotation, they are usually compatible only with special peak generating workflows (*e.g.* CAMERA with XCMS), or they use raw LC-MS data and perform peak picking and ion annotation in series. Hence, it would be of interest to develop a post-peak picking workflow to process peak lists from any workflow, identify redundant peaks and remove them.

In this work, we developed a MATLAB-based workflow to filter out redundant peaks from peak picking results. It imports data from any peak generating workflow, performs different filters and results in more condensed peak list by removing redundant peaks. After evaluating the performance of this filtering method, the performance of four peak picking workflows, namely PeakView®, Markerview™, MetabolitePilot™ and XCMS online were assessed in terms of the number of redundant peaks found by our developed *DataReduction* workflow.

3.3 Material and Methods

Two types of biological samples (bile and urine) as well as a standard mixture of 84 compounds were used for this study. Samples were analyzed with liquid chromatography coupled to high-resolution mass spectrometry employing a QqTOF system. Raw LC-MS data were then processed with four peak picking workflows including MarkerView, PeakView, MetabolitePilot and XCMS online followed by employing the *DataReduction* workflow for filtering of redundant peaks. Since the same datasets from Chapter 2 were used here, detailed information on materials, sample preparation, HPLC-MS analysis and peak picking criteria have been presented previously.

3.3.1 MATLAB processing

"DataReduction" MATLAB code was written precisely to find and remove peaks corresponding to ^{13}C isotopes, radical ions as well as some frequent adducts and in-source fragments present in peak lists. **"DataReduction"** used 5 ppm mass tolerance and 0.1 min difference in retention time for annotating related peaks. The whole MATLAB script is presented at the end of this chapter (supplementary data for chapter 3).

From an initial peak list (generated by peak picking workflows) exported into an excel-compatible format, the script finds ^{13}C isotopic peaks with 5 ppm mass accuracy and 0.1 min RT difference, with additional criteria that ^{13}C isotope peaks should have lower intensity (for small molecules). Assuming that the majority of observed m/z values are protonated in positive mode and deprotonated in negative mode, radical ions were identified in a subsequent filtering step. After association of peaks to ^{13}C isotopes and radical ions, they were eliminated from original peak list.

A list containing information of the most frequent adducts and in-source fragments was imported into MATLAB. The following adducts were included in this study; NH_4^+ , Na^+ , K^+ , as well as doubly charged $(\text{M}+2\text{H})^{2+}$ (positive mode) and HCO_2^- (negative mode) in addition to in-source fragments such as loss of water, formic acid (H_2CO_2) and ammonia (NH_3) (in positive mode) and carbon dioxide (CO_2) (in negative mode). This excel sheet contains the exact mass difference and charge associated to each adduct and fragment and additional cases can be added manually. This MATLAB script was used to remove redundant peaks and results in a more concise peak list with higher proportion of protonated or deprotonated molecules (intact metabolites).

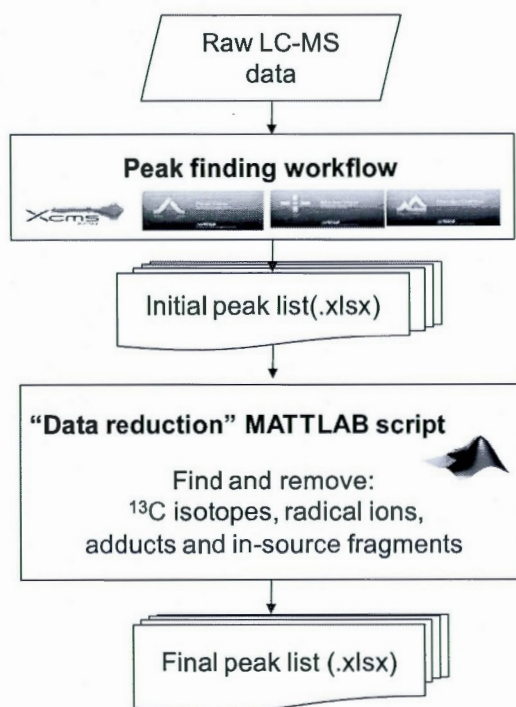


Figure 3.1 "DataReduction" MATLAB script was used in this study to identify and remove isotope peaks, radical ions, adducts and in-source fragments

3.4 Results and discussion

The "*DataReduction*" MATLAB script was designed for processing peak picking results from raw data and also for its compatibility with any peak generating workflow, as long as peak list information (m/z , RT and intensity or peak area) is exportable into an excel-compatible format. The initial peak lists obtained from peak generating workflows from two biological sample types (urine and bile) and a standard mixture of 84 compounds were investigated with the custom-built "*DataReduction*" MATLAB script, where ^{13}C isotopes, radical ions, as well as some adducts and in-source fragments were detected (Figure 3.2).

Even though this MATLAB script does not use correlation between peak shapes for peak annotation, narrow mass and retention time difference windows (5 ppm mass tolerance and 0.1 min retention time difference) were employed to ensure that the filtering step was efficient at removing redundant peaks with little chance of removing intact (protonated or deprotonated) metabolite peaks. In order to evaluate the "*DataReduction*" MATLAB script, a comparison was made between the peak lists found by the custom script and the isotopic peak detection option built into the MarkerView software (Table 3.1 and Figure 3.6). The results show that for the bile data set in negative mode, "*DataReduction*" MATLAB script and MarkerView™ found 1848 and 1806 isotopic peaks, respectively, while 1616 peaks were common between these two methods. On average, almost 85% of total isotope peaks found by MarkerView and DataReduction workflow were common to both. This indicates the good efficiency of the "*DataReduction*" MATLAB script for peak annotation using the chosen criteria, since MarkerView™ isotope peak assignment uses elution profile information (during peak generation step) and not simply ΔRT and $\Delta m/z$ of the apex of each chromatographic peak.

After confirming the performance of the DataReduction MATLAB script, the four peak picking workflows used in this study were evaluated in terms of number of

redundant peaks found by this filtering method (Figures 3.3 to 3.5). Although PeakView had the lowest number of peaks initially found, it has the greatest proportion of "non-redundant peaks" (those remaining after removing redundant peak) with an average of 90% of the original list being conserved after filtering. MetabolitePilot and PeakView represented the lowest number of ^{13}C isotopes among their results, likely due to a built-in algorithm for excluding these isotopes. Meanwhile, XCMS and MarkerView had assigned isotopic peaks for further manual removal by the user. No evident difference between different peak generating workflows was observed regarding the % peaks corresponding to radical ions, adducts and in-source fragments, indicating there would be no pre-filtering for these in any of the automated workflows tested. It would obviously be useful to have software able to generate peak lists and filter for such redundant peaks in the same step, which is possible in certain commercial software such as Agilent's Qualitative Analysis and Waters' Progenesis software.

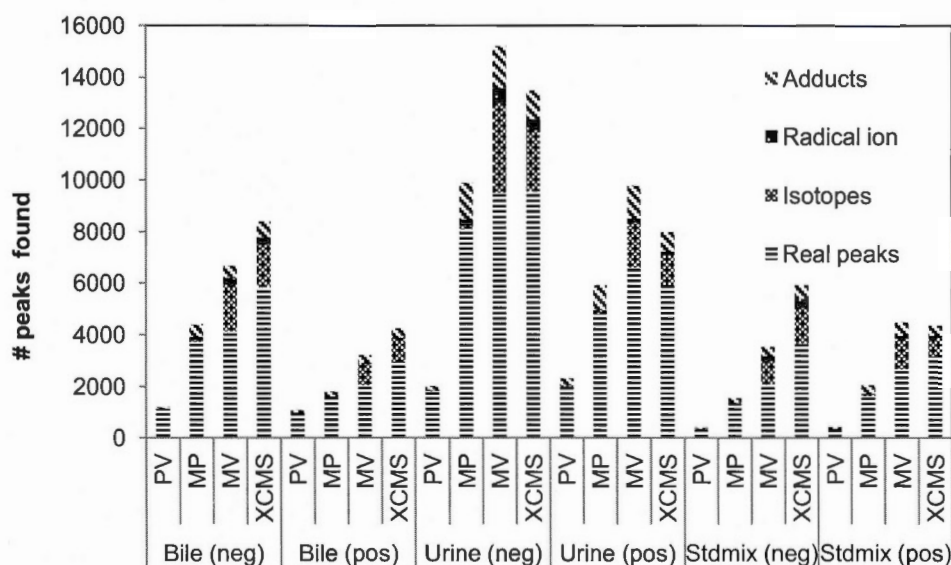


Figure 3.2 Number of peaks found by MATLAB data reduction script for different workflows: PeakView (PV), MarkerView (MV), MetabolitePilot (MP) and XCMS online for each sample type (bile, urine and compound mixture) in both positive and negative modes

Table 3.1 Comparison of the total number of ^{13}C isotope peaks and the overlap between two filtering algorithms: ("*DataReduction*" and MarkerView) in bile and urine sample in both positive and negative modes

Sample (mode)	MarkerView	DataReduction	Common
Urine (neg)	3498	3667	3085
Urine (pos)	1770	1998	1534
Bile (neg)	1806	1848	1616
Bile (pos)	805	813	714

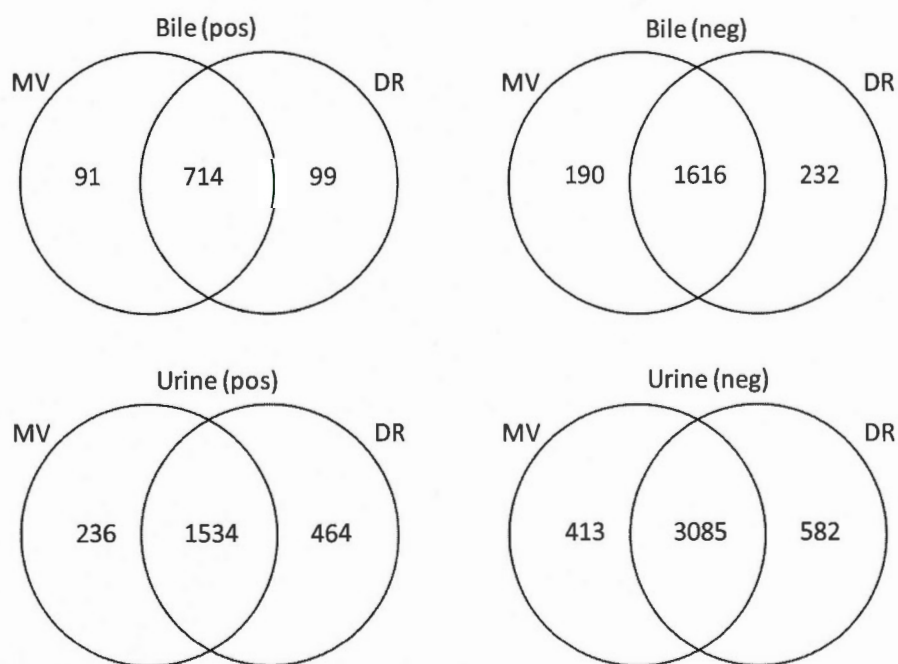


Figure 3.3 Venn diagram representation of the results of comparison between ^{13}C isotope peaks found by MarkerView (MV) and developed *DataReduction* (DR) MATLAB script.

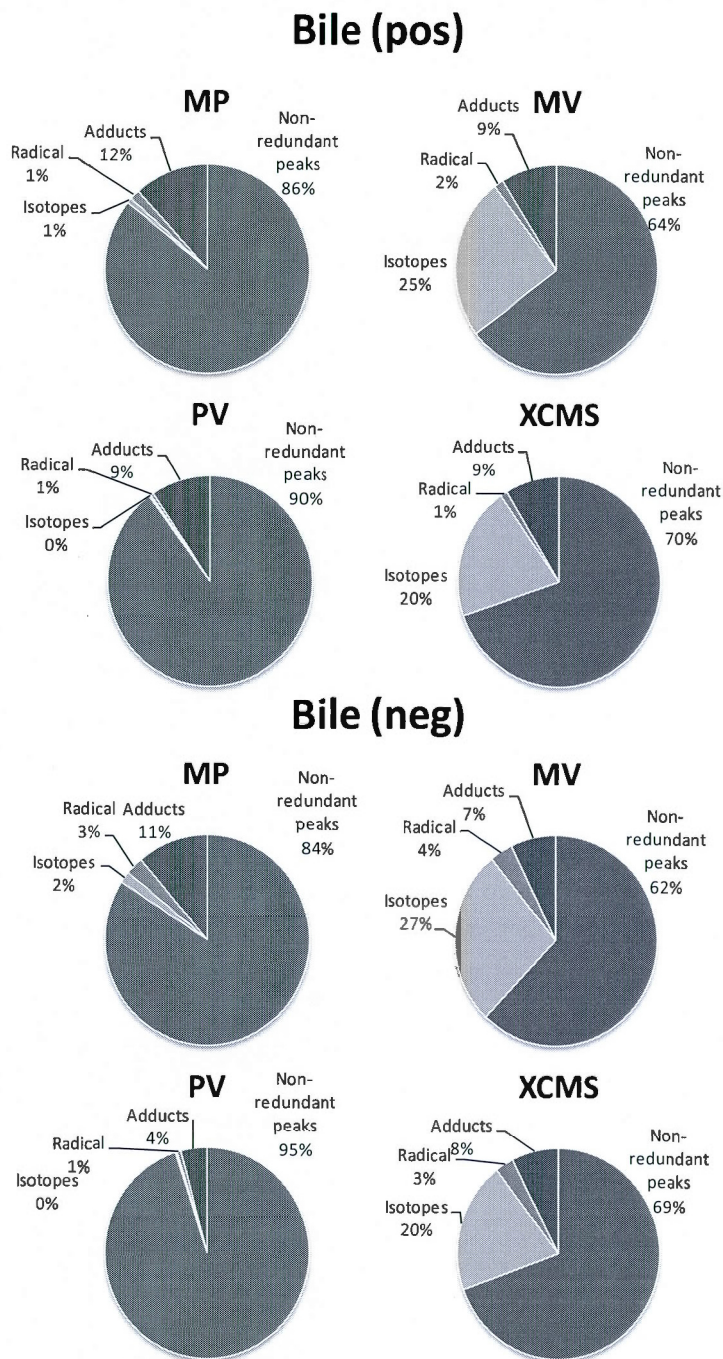


Figure 3.4 Pie chart representation of the percentage of redundant peaks found by *DataReduction* MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from bile sample in positive and negative modes

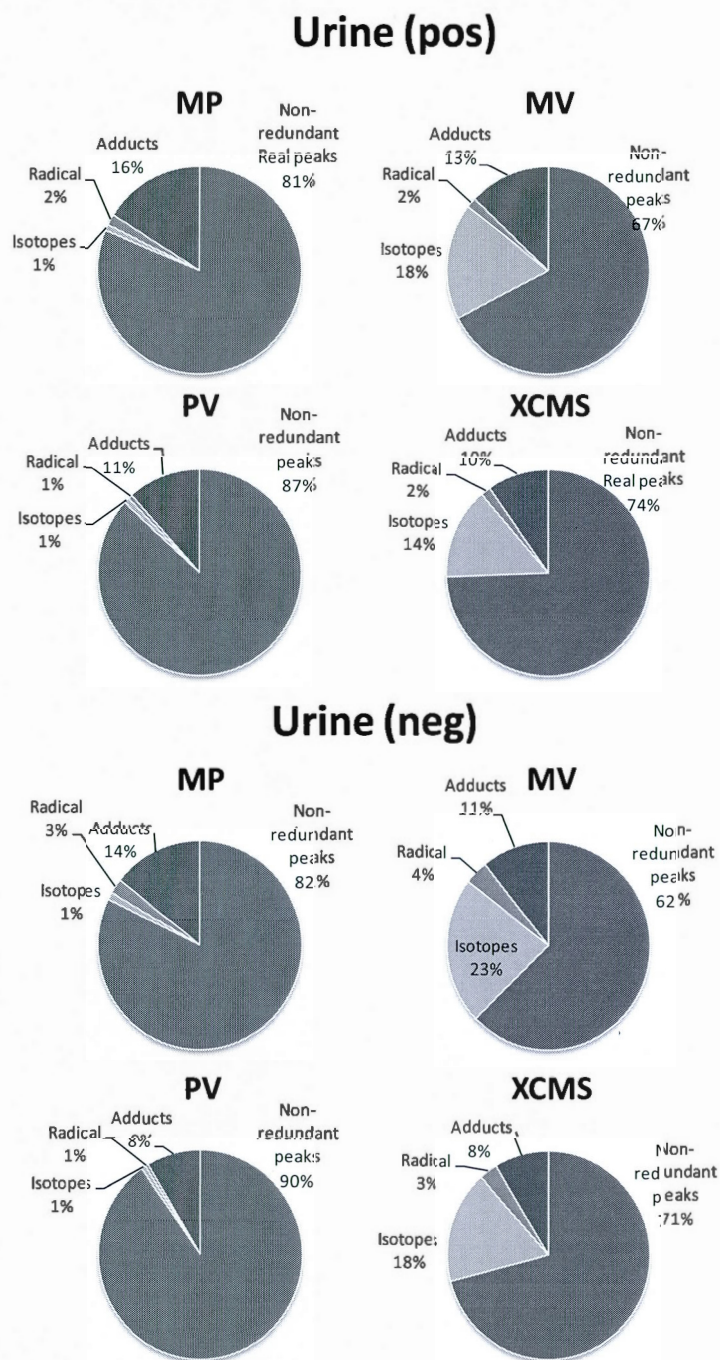


Figure 3.5 Pie chart representation of the percentage of redundant peaks found by *DataReduction* MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XC)) from urine sample in positive and negative modes

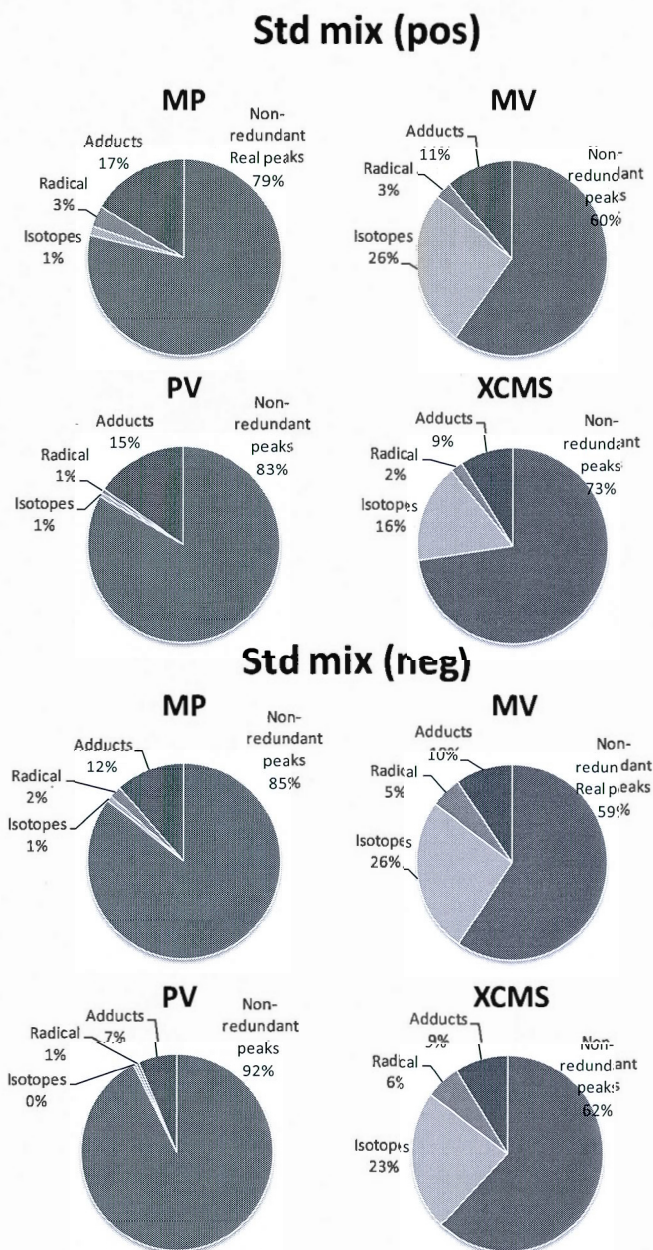


Figure 3.6 Pie chart representation of the percentage of redundant peaks found by *DataReduction* MATLAB script for the results of four peak picking workflows (MetabolitePilot (MP), MarkerView (MV), PeakView (PV), XCMS online (XCMS)) from the standard mixture (Std mix) sample in positive and negative modes

3.5 Conclusion

The "*DataReduction*" MATLAB script was developed for processing initial peak picking results. It is compatible with any peak generating workflow, as long as peak list information (m/z , RT and intensity or peak area) is exportable into excel format. It uses small m/z and RT windows for identification of redundant peaks related to isotopes, radical ions, adducts and in-source fragments. Although it doesn't employ peak shape similarity as criteria to find redundant peaks, it shows good performance and it found 85% of the isotope related peaks that were also found by MarkerView, a peak picking software able to assign isotopic peaks based on peak detection profiles.

By employing this *Data Reduction* method, four peak picking workflows were compared to each other. It was found that an isotopic detection algorithm is used in PeakView and MetabolitePilot, which resulted in the detection of almost no peaks as isotopes by our custom-built script. On the other hand, MarkerView and XCMS online show all found peaks, including isotopes, in their generated peak lists with the possibility of removing isotope peaks by the user in a subsequent step.

In this chapter, *Data Reduction* MATLAB script was presented, further investigation on its performance for identification of the redundant peaks will be performed in future works.

Supplementary data for chapter 3

DataReduction Matlab script

Datareduction MATLAB script was developed in order to find redundant peaks and remove them from original data. The whole manuscript is presented in following:

```

Clc
Clear all
Number=6
format long
%---> Read Sample files
sample=xlsread('F:\Bile_neg_MV');
filename='F:\MATLAB_Bile_neg_MV';
%---> Define ionization state (positive or negative)
chargestate=2; %pos==1, neg==2
if chargestate==1
adduct=xlsread ('final_adductslist1.xlsx',6);
end
if chargestate==2
adduct=xlsread ('final_adductslist1.xlsx',7);
end
%---> Define parameters
c1=2;
c2=2;
p1=2;
zero11cell=[0,0,0,0,0,0,0,0,0,0];
C13whole=zero11cell;
C13_arearatio=zero11cell;
Proto_whole=zero11cell;
Protominuscommon=zero11cell;
Proto_C13common=zero11cell;
a1=2;
a2=2;
a3=2;
a4=2;
a5=2;
p2=2;
p3=2;
ssrt=1;
zero4cell=[0,0,0,0];
wholeminusc13=zero4cell;
wholeminusc13M=zero4cell;
adductinfo=zero11cell;

```



```

addadd=zero4cell;
realadd=zero4cell;
realpeaks=zero4cell;
adducts=zero4cell;
c13_proto_title=zero11cell;
whole_8cell_title=zero11cell;
whole4cell_title=zero4cell;
adductinfo_title=zero11cell;
%
%---> Remove rt<1.6 ifrom sample
for s1=1:size(sample,1)
    rtsinitial=sample(s1,3);
    if rtsinitial>1.6
        isinitial=sample(s1,1);
        mzsinitial=sample(s1,2);
        intsinitial=sample(s1,4);
        samplert(ssrt,1:4)=[isinitial,mzsinitial,rtsinitial,intsinitial];
        ssrt=ssrt+1;
    end
end
%
%---> Find C13 isotopes
for i=1:size(samplert,1);
    numi=samplert(i,1);
    mhzi=samplert(i,2);
    Protoadd=mhzi+1.00335;
    for j=1:size(samplert,1);
        numj=samplert(j,1);
        mhzj=samplert(j,2);
        ppm=abs((mhzj-Protoadd)/(mhzj*10e-6));
        if ppm<5;
            rtmzi=samplert(i,3);
            rtmzj=samplert(j,3);
            delrt=abs(rtmzi-rtmzj);
            areai=samplert(i,4);
            areaj=samplert(j,4);
            arearatio=(areai/areaaj);
            if delrt<0.1
                if arearatio>1
                    C13_arearatio(c2,1:11)=[numi,mhzi,rtmzi,areai,numj,mhzj,rtmzj,areaaj,delrt,ppm,arearatio];
                    c2=c2+1;
                end
            end
        end
    end
end
end
end

```

```

%
%----> Find Radical ions
for i=1:size(samplert,1);
    numi=samplert(i,1);
    mhzi=samplert(i,2);
    Protoadd=mhzi+1.007825;
    for j=1:size(samplert,1);
        numj=samplert(j,1);
        mhzj=samplert(j,2);
        ppm=abs((mhzj-Protoadd)/(mhzj*10e-6));
        if ppm<5;
            rtmzi=samplert(i,3);
            rtmzj=samplert(j,3);
            delrt=abs(rtmzi-rtmzj);
            areai=samplert(i,4);
            areaj=samplert(j,4);
            arearatio=(areai/areaj);
            if delrt<0.1

Proto_whole(p1,1:11)=[numi,mhzi,rtmzi,areai,numj,mhzj,rtmzj,areaj,delrt,ppm,arearatio];
        p1=p1+1;
        end
    end
end
end
%
%----> Remove C13 isotopes from radical ions for positive mode (chargestate==1) since
we should erase lower mass in protonated relation while removing higher mass in c13
relation
if chargestate==1
for q=1:size(Proto_whole,1)
    numi=Proto_whole(q,1);
    t=1;
    for w=1:size(C13_arearatio,1)
        c13=C13_arearatio(w,1);
        delwi=abs(c13-numi);
        if delwi==0
            t=t+1;
        end
    end
    end
    if t==1
        mhzi=Proto_whole(q,2);
        rtmzi=Proto_whole(q,3);
        areai=Proto_whole(q,4);
        numJ=Proto_whole(q,5);
        mhzj=Proto_whole(q,6);
        rtmzj=Proto_whole(q,7);
        areaj=Proto_whole(q,8);
    end
end

```

```

delrt=Proto_whole(q,9);
ppm=Proto_whole(q,10);
arearatio=Proto_whole(q,11);

```

```

Protominuscommon(p2,1:11)=[numi,mhzi,rtmzi,areai,numJ,mhzj,rtmzj,areaj,delrt,ppm,area
ratio];

```

```

    p2=p2+1;
    end
    if t>1
        mhzi=Proto_whole(q,2);
        rtmzi=Proto_whole(q,3);
        areai=Proto_whole(q,4);
        numJ=Proto_whole(q,5);
        mhzj=Proto_whole(q,6);
        rtmzj=Proto_whole(q,7);
        areaj=Proto_whole(q,8);
        delrt=Proto_whole(q,9);
        ppm=Proto_whole(q,10);
        arearatio=Proto_whole(q,11);
    end

```

```

Proto_C13common(p3,1:11)=[numi,mhzi,rtmzi,areai,numJ,mhzj,rtmzj,areaj,delrt,ppm,area
ratio];

```

```

    p3=p3+1;
    end

```

```

end
end
%
```

```

%---> Remove assigned C13 isotopes from original peak list

```

```

for i=1:size(samplert,1);
    iorigin=samplert(i,1);
    k=1;
    for v=1:size(C13_arearatio,1);
        numc13=C13_arearatio(v,5);
        vi=abs(iorigin-numc13);
        if vi==0
            k=k+1;
        end
    end
    if k==1
        mhz=samplert(i,2);
        rt=samplert(i,3);
        int=samplert(i,4);
        wholeminusc13(a1,1:4)=[iorigin,mhz,rt,int];
        a1=a1+1;
    end
end

```

```

end
%
```

```

% ---> Remove assigned radicals from peak list

```



```

for i=1:size(wholeminusc13,1);
    iorigin=wholeminusc13(i,1);
    k=1;
    if chargestate==1
        for v=1:size(Protominuscommon,1);
            numradical=Protominuscommon(v,1);
            vi=abs(iorigin-numradical);
            if vi==0
                k=k+1;
            end
        end
    end
    if chargestate==2
        for v=1:size(Proto_whole,1);
            numradical=Proto_whole(v,5);
            vi=abs(iorigin-numradical);
            if vi==0
                k=k+1;
            end
        end
    end
    if k==1
        mhz=wholeminusc13(i,2);
        rt=wholeminusc13(i,3);
        int=wholeminusc13(i,4);
        wholeminusc13M(a2,1:4)=[iorigin,mhz,rt,int];
        a2=a2+1;
    end
end
%
%-----> Find adducts and in-source fragments
for h=1:size(adduct,1);
    ionmass=adduct(h,1);
    charge=adduct(h,2);
    for i=1:size(wholeminusc13M,1);
        i1=wholeminusc13M(i,1);
        mzi=wholeminusc13M(i,2);
        adductmass=(mzi+ionmass)/(charge);
        for j=1:size(wholeminusc13M,1);
            j1=wholeminusc13M(j,1);
            mhzj=wholeminusc13M(j,2);
            ppm=abs((mhzj-adductmass)/(adductmass*10e-6));
            if ppm<5;
                rtmzi=wholeminusc13M(i,3);
                rtmzj=wholeminusc13M(j,3);
                delrt=abs(rtmzi-rtmzj);
                if delrt<0.1

```

```

adductinfo(a3,1:11)=[i1,mzi,rtmzi,j1,mhzj,rtmzj,ionmass,charge,adductmass,delrt,ppm];
    a3=a3+1;
    end
    end
    end
end
end
%
%----> Remove assigned adducts of adducts from list of adducts (Due to complexity of
sample, some adducts represented to have adducts)
b1=1;
b2=2;
for i=1:size (adductinfo,1);
msnum1=adductinfo(i,1);
k=1;
    for j=1:size (adductinfo,1);
        msnum2=adductinfo(j,4);
        diff=abs(msnum1-msnum2);
        if diff==0
            k=k+1;
        end
    end
    if k>1
        addofaddnum=adductinfo(i,4);
        mzaddadd=adductinfo(i,5);
        rtaddadd=adductinfo(i,6);
        addadd(b1,1:4)=[addofaddnum,mzaddadd,rtaddadd,k];
        b1=b1+1;
    end
    if k==1
        realaddnum=adductinfo(i,4);
        mzadd=adductinfo(i,5);
        rtadd=adductinfo(i,6);
        realadd(b2,1:3)=[realaddnum,mzadd,rtadd];
        b2=b2+1;
    end
end
%
%----> Remove adducts from peaks list

for i=2:size(wholeminusc13M,1);
    num1=wholeminusc13M(i,1);
    n1=1;
    %peaks who are adducts
    for j=1:size(realadd,1);
        num2=realadd(j,1);
        delij=abs(num1-num2);

```

```

    if delij==0
        n1=n1+1;
    end
end
    % its not an adduct
    if n1==1
        mhz=wholeminusc13M(i,2);
        rt=wholeminusc13M(i,3);
        int=wholeminusc13M(i,4);
        realpeaks(a4,1:4)=[num1,mhz,rt,int];
        a4=a4+1;
    end
    %its an adduct
    if n1>1
        mhz=wholeminusc13M(i,2);
        rt=wholeminusc13M(i,3);
        int=wholeminusc13M(i,4);
        adducts(a5,1:4)=[num1,mhz,rt,int];
        a5=a5+1;
    end
end
%
%-----
%---> Assign title for excel sheets
sheet0='whole_file';
sheet3='C13ratio';
sheet4='Allprotonated';
sheet6='MinusC13';
sheet7='MinusC13M';
sheet8='AdductsInfo';
sheet9='RealPeaks';
sheet10='Adducts';
sheet12='sizeinfo';
sheet13='adductofadd';
sheet14='realadd';
sheet1='proto_minusc13ratio';
sheet11='proto_c13common';
%
% ---> Write down all produced data into excel sheets
xlswrite(filename,samplert,sheet0);
xlswrite(filename,C13_arearatio,sheet3);
xlswrite(filename,Proto_whole,sheet4);
xlswrite(filename,Protominuscommon,sheet1);
xlswrite(filename,Proto_C13common,sheet11);
xlswrite(filename,wholeminusc13,sheet6);
xlswrite(filename,wholeminusc13M,sheet7);
xlswrite(filename,adductinfo,sheet8);
xlswrite(filename,addadd,sheet13);
xlswrite(filename,realadd,sheet14);

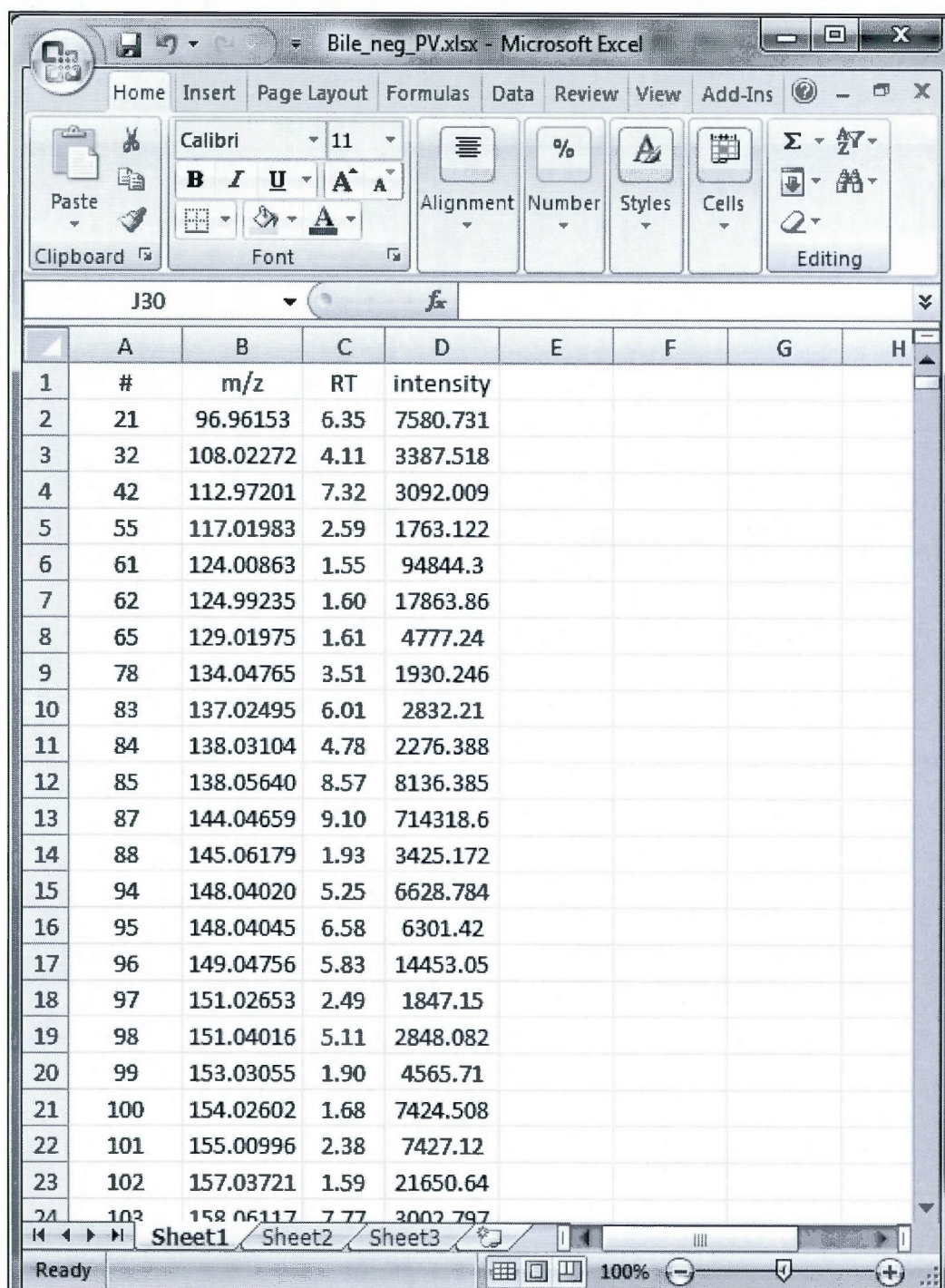
```



```

xlswrite(filename,realpeaks,sheet9);
xlswrite(filename,adducts,sheet10);
c13_proto_title=[{'i'},{'mhzi'},{'rtmzi'},{'inti'},{'j'},{'mzj'},{'rtmzj'},{'intj'},{'delrt'},{'ppm'},{'arearatio'}];
whole4cell_title=[{'iorigin'},{'mhzi'},{'rt'},{'area/intensity'}];
adductinfo_title=[{'i'},{'mzi'},{'rtmzi'},{'j'},{'mzj'},{'rtmzj'},{'ionmass'},{'charge'},{'adduct mass'},{'delrt'},{'ppm'}];
xlswrite(filename,c13_proto_title,sheet3);
xlswrite(filename,c13_proto_title,sheet4);
xlswrite(filename,whole4cell_title,sheet6);
xlswrite(filename,whole4cell_title,sheet7);
xlswrite(filename,whole4cell_title,sheet9);
xlswrite(filename,adductinfo_title,sheet8);
xlswrite(filename,whole4cell_title,sheet10);
xlswrite(filename,c13_proto_title,sheet1);
xlswrite(filename,c13_proto_title,sheet11);
xlswrite(filename,whole4cell_title,sheet13);
xlswrite(filename,whole4cell_title,sheet14);
originalsize=size(samplert,1);
c2=c2-1;
p1=p1-2;
p2=p2-2;
p3=p3-2;
a1=a1-2;
a2=a2-2;
a3=a3-2;
a4=a4-2;
a5=a5-2;
b1=b1-2;
b2=b2-2;
sizeinfo(2,1:12)=[originalsize,c2,p1,p2,p3,a1,a2,a3,a5,b1,b2,a4];
xlswrite(filename,sizeinfo,sheet12);
sizeinfoTitle=[{'original size'},{'C13ratio>1'},{'Allprotonated'},{'Protonated Minus C13ratio'},{'Protonated_C13ratio common'},{'Allpeaks minus C13ratio'},{'AllMinusC13protonated'},{'adducts pair'},{'Adducts'},{'addadd'},{'realadd'},{'non-redundant peaks'}];
xlswrite(filename,sizeinfoTitle,sheet12);
% End of the DataReduction manuscript

```



Bile_neg_PV.xlsx - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Add-Ins

Clipboard Font Alignment Number Styles Cells Editing

J30

	A	B	C	D	E	F	G	H
1	#	m/z	RT	intensity				
2	21	96.96153	6.35	7580.731				
3	32	108.02272	4.11	3387.518				
4	42	112.97201	7.32	3092.009				
5	55	117.01983	2.59	1763.122				
6	61	124.00863	1.55	94844.3				
7	62	124.99235	1.60	17863.86				
8	65	129.01975	1.61	4777.24				
9	78	134.04765	3.51	1930.246				
10	83	137.02495	6.01	2832.21				
11	84	138.03104	4.78	2276.388				
12	85	138.05640	8.57	8136.385				
13	87	144.04659	9.10	714318.6				
14	88	145.06179	1.93	3425.172				
15	94	148.04020	5.25	6628.784				
16	95	148.04045	6.58	6301.42				
17	96	149.04756	5.83	14453.05				
18	97	151.02653	2.49	1847.15				
19	98	151.04016	5.11	2848.082				
20	99	153.03055	1.90	4565.71				
21	100	154.02602	1.68	7424.508				
22	101	155.00996	2.38	7427.12				
23	102	157.03721	1.59	21650.64				
24	103	158.06117	7.77	3002.797				

Sheet1 Sheet2 Sheet3

Ready 100%

Figure S5 Sample excel sheet containing peak detection information to be imported to MATLAB DataReduction workflow

CHAPTER IV

CAROTENOID QUANTITATION IN ALGAL SAMPLES BY LIQUID CHROMATOGRAPHY-HIGH RESOLUTION MASS SPECTROMETRY

4.1 Abstract

Carotenoids are organic pigments found in plants, algae and some other microorganisms with antioxidant properties and vitamin A activity. Stress conditions, including reduced nitrogen source and long culturing time can induce up-regulation of some carotenoids in microalgae. In this work, we used a new liquid chromatography-mass spectrometry (LC-MS) quantitation assay to determine changes in carotenoid content in three different algae species of *Haematococcus*, *Oocystis* and *Muriellopsis* under stress conditions. Carotenoid separation and subsequent analysis were carried out by HPLC coupled to a hybrid quadrupole time-of-flight mass spectrometer. Based on accurate mass measurements, four carotenoids including astaxanthin, canthaxanthin, β -carotene and lutein were quantified in control and stressed algal samples. Results show that the astaxanthin and canthaxanthin content was up-regulated by stress conditions while a decrease in lutein and β -carotene was revealed in *Haematococcus* and *Muriellopsis*. Decrease of all four studied carotenoids were observed in *Oocystis*.

4.2 Introduction

Carotenoids are a family of compounds consisting of approximately 700 known species. They are divided into two categories, 1) carotenes, consisting of eight isoprenoid units joined to form conjugated hydrocarbons (e.g. α -carotene, β -carotene, lycopene) and their hydroxylated derivatives, and 2) xanthophylls, which are oxygenated derivatives of carotenes (including lutein, zeaxanthin, and β -cryptoxanthin) (Granado *et al.* 2001; Oliveira and Watson 2001). This class of

pigments is responsible for the red color of tomato and orange color of carrots as well contributing to fall season colors after leaf chlorophyll is destroyed (Krinsky and Johnson 2005).

In addition to the aesthetic role of carotenoids in nature, they have a great biochemical role in health. Since they are exclusively synthesized in plants, algae and some fungal and bacterial species, they should be supplied in the diet of animals (Davies 1985). They have important biological roles for vitamin A biosynthesis (important for vision) as well as antioxidant activity, immunostimulant, yolk nourishment to embryos, photo-protection, limiting age-related macular degeneration of the eye (Johnson 2002) and can also be useful as therapeutic agents for treating cardiovascular disease and prostatic cancer (Fassett and Coombes 2011). The effect of dietary carotenoids on health are reviewed by several groups (Cooper *et al.* 1999; Hughes 2001; Young and Lowe 2001).

Green algae can produce all xanthophylls generated by higher plants (Jin *et al.* 2003) as well as some other xanthophylls which is specific to green algae such as luteoxanthin (Baroli and Niyogi 2000), astaxanthin (Grünewald *et al.* 2001) and canthaxanthin (Grünewald *et al.* 2001). Although synthetic carotenoids could be used, the natural pigment is preferred for two reasons. First, suspected role of synthetic food additives as promoters of carcinogenesis, besides claims of liver and renal toxicities (El-Baky *et al.* 2003) and secondly, natural carotenoids is a mixture of trans and cis (favorable) isomers which is very hard to obtain through chemical synthesis (Demmig-Adams and Adams 2002).

Interestingly, the carotenoid production rate is up-regulated in microalgae by unfavorable environmental conditions namely exposure to intense light, nitrogen starvation, excess acetate addition, salt stress, and the addition of specific cell division inhibitors (Ben-Amotz *et al.* 1989; Borowitzka 1992; Margalith 1999).

This method has been used in industry by several companies to produce astaxanthin from *H. pluvilis* algae (Olaizola 2000).

Liquid chromatography is the most frequently used technique for the separation of carotenoids in biological samples (Oliveira and Watson 2001), while various detection techniques have been used, including electrochemical detection (Finckh *et al.* 1995; Ferruzzi *et al.* 1998; van het Hof *et al.* 2000), fluorescence detection (Yap *et al.* 1999), UV detection (van het Hof *et al.* 2000) and mass spectrometry-based detection (van Breemen *et al.* 1998; Chu *et al.* 2011; Fu *et al.* 2012). High sensitivity and low detection limit of mass spectrometry provide high accuracy to measure low concentration of carotenoids in complex biological samples. By employing metabolomic analysis (targeted or untargeted), simultaneous identification and quantification of large number of compounds is feasible. Different ionization sources have been used for this purpose, including ESI, APCI (van Breemen *et al.* 1998; Taylor *et al.* 2006; Matsumoto *et al.* 2007) and matrix-assisted laser desorption ionization (MALDI)-TOF-MS for metabolite profiling of plant carotenoids in a high-throughput manner (Fraser *et al.* 2007).

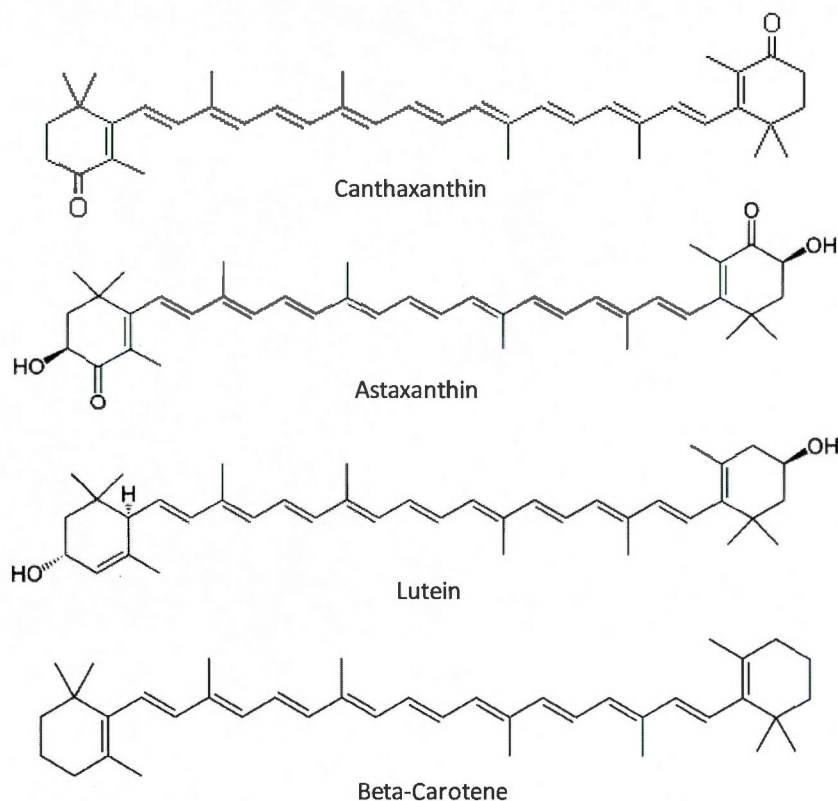


Figure 4.1 Four carotenoid compounds studied in this work

In this work, we used a mass spectrometry-based quantification assay to determine changes in carotenoid content under stress conditions in three different algae species, *Haematococcus*, *Oocystis*, and *Muriellopsis*. Carotenoid separation and subsequent analysis were done using a UHPLC instrument coupled to a hybrid quadrupole time-of-flight (QqTOF) mass spectrometer. An on-line UV detector was also used for further confirmation of studied compounds. Based on exact mass measurements, four carotenoids (Figure 4.1) were quantified in control and stressed algal samples.

4.3 Experimental

4.3.1 Materials

Methyl tert-butyl ether (MTBE, HPLC grade) was purchased from Caledon Laboratory Chemicals (Georgetown, ON, Canada). HPLC grade methanol, formic acid, menaquinone (vitamin K₂) and ultrapure ammonium acetate as well as β -carotene, β -apo-8'-carotenal, canthaxanthin, lutein and alpha-tocopherol (vitamin E) were obtained from Sigma-Aldrich (Oakville, ON, Canada). Ultrapure water was supplied by a Millipore Synergy UV purification system from Fisher Scientific (Mississauga, ON, Canada). Astaxanthin was purchased from Chromadex (Irvine, CA, USA). Filters (0.45 μ m, 25 mm nylon) were obtained from Millipore and glass beads (0.5 mm) purchased from Bertin Technologies (Montigny, France). Carotenol and carotenal compounds were synthesized in the laboratory of Prof. René Roy by Tze Chieh Shiao.

4.3.2 Algal samples

Algae were cultured at 23°C in 250 mL erlenmeyer flasks under continuous 50 μ mol photons $\text{m}^{-2}.\text{s}^{-1}$ illumination. To induce secondary carotenoid accumulation (orange/red algae), cultures were prepared with BBM medium without nitrate. Cultures in Bold's Basal Medium (BBM, a highly enriched culturing medium including several minerals and vitamins such as KH_2PO_4 , H_3BO_3 *etc.*) (Stein 1980) were used as control (green algae). Algal cultures were exposed for 1-3 months to this treatment. The cellular density in all cultures was determined by optical microscope counting. All cultures were grown in triplicate.

4.3.3 Carotenoid extraction and sample preparation

Each algal culture (0.1-1 ml) with known cell density was filtered using a 2.5 cm nylon filter, and then transferred to 2 ml screw-cap tubes for carotenoid extraction. Glass beads (0.5 mm, 200 mg) were added along with 500 μ l extraction solvent (33%MTBE, 66% MeOH) and 100 μ l vitamin K2 (menaquinone) solution (2 μ g/ml in MeOH) as internal standard. Carotenoid extraction was carried out using a Minilys bead beater (Bertin Technologies, Montigny-le-Bretonneux, France) at medium speed (4,000 rpm) for 1 min. After centrifuging at 14,000 rpm for 4 min, 350 μ l supernatant was transferred to a clean tube and bead beating was repeated once more with another 500 μ l of fresh extraction solvent. Combined supernatants (700 μ l) were then evaporated (Thermo Electron, SpeedVac Concentrator). Dried samples were reconstituted in 200 μ l MeOH and 20 μ l of extract was injected for LC-MS/MS analysis. The choices of internal standard and timing for the addition of IS to compounds are presented in the method development section of this chapter. The summary of the sample preparation procedure used in this study is presented in Figure 4.2.

4.3.4 Standard mixture

Standard mixtures consisting of four carotenoids; astaxanthin, canthaxanthin, lutein and β -carotene were made by serial dilution of stock standard solution. Final component concentrations were 0.00, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, and 20.0 μ g/ml. Sample preparation was performed for standard mixtures by mimicking the procedure of sample preparation for algal samples starting with bead beating extraction. Menaquinone was added to standard solutions as internal standard in extraction solution. Evaluation of the effect of sample preparation on carotenoid quantitation is presented as method development section.

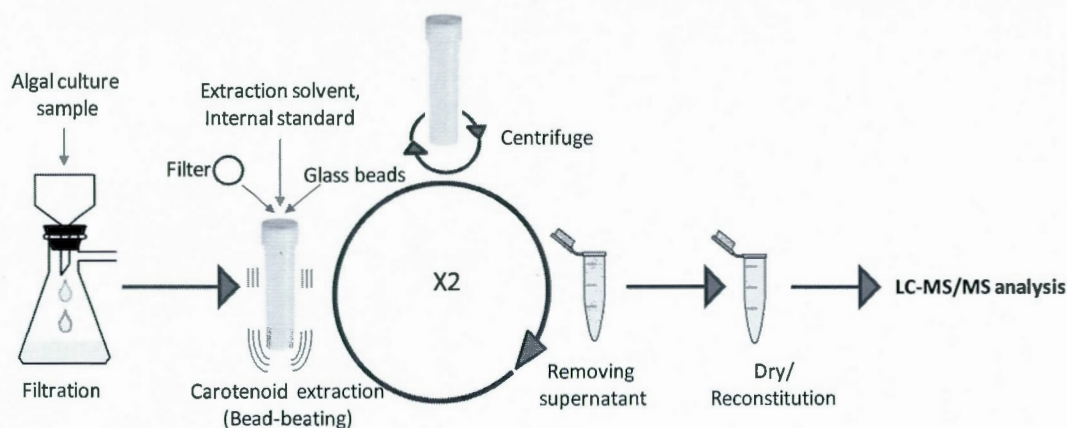


Figure 4.2 Sample preparation used for carotenoid quantification in algal samples. It starts with filtration of the algal culture followed by bead beating extraction of carotenoids. Less solvent consumption and faster sample preparation was achieved with this optimized extraction method

4.3.5 HPLC-UV-MS analysis

Extracted algal samples and standard solutions were injected into a Shimadzu Nexera HPLC system equipped with an online Shimadzu UV/Vis-detector ($\lambda=450$ nm). The LC-MS criteria used in this study was previously optimized in our group by colleagues (unpublished data). Separation was performed on a Phenomenex Gemini-NX reversed phase C18 column (150 x 2 mm, 3 μ m) at 40°C. Mobile phases A (90% MeOH/10% H₂O/0.1% FA) and B (85% MTBE/15% MeOH/0.1% FA) were used in a gradient elution (Figure 4.3) at 0.2 mL/min. LC was coupled to an AB Sciex Triple-TOF 5600 with positive electrospray ionization on a DuoSpray Ion Source. Source parameters were set as follows: TEM 400 °C, CUR 30 psi, source voltage 5 kV and GS1/2 50 psi. An in-house standard mix (m/z 119-966 in negative mode and m/z 121-922 in positive mode) was used for internal mass calibration. Peak assignments were based on high resolution m/z data and verified by coincident UV signal (with appropriate retention time shift from delay volume between UV and MS). These results were in

agreement with previous work in our group which carotenoid structure was confirmed with MS/MS matching (unpublished data).

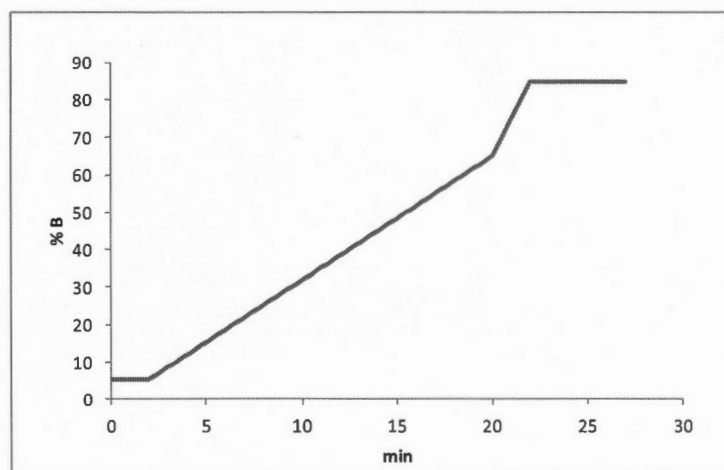


Figure 4.3 Gradient elution used for HPLC-UV-MS analysis. Mobile phase A was (90% MeOH/10% H₂O/0.1% FA) and B was (85% MTBE/15% MeOH/0.1% FA)

4.3.6 Data Processing

Raw LC-HRMS data was processed by PeakView® 1.2 software (AB SCIEX) for peak extraction and visualization of chromatograms. Multiquant 2.1 was used for quantitation purposes based on peak area integration for carotenoid compounds and internal standard. Calibration curves were produced using peak area ratio of compounds and internal standard versus their concentration ratio.

4.4 Method development

In this work, quantitation of changes in carotenoid content of algal samples grown under stress conditions was investigated by HPLC-UV-MS method. As any other quantitative analytical experiment, an internal standard is a crucial step that needs to be considered for method development.

In an internal calibration method, the relative intensity of instrument's response of an analyte of interest to the internal standard's signal is used for comparison calibrating standards. This relative comparison eliminates the effect of errors in sample preparation, introduction of the sample into mass spectrometer as well as possible fluctuation in ion source conditions and instability of the mass scale (Hoffmann and Stroobant 2007).

An appropriate internal standard should exhibit similar chemical and physical properties to the analytes of interest in order to give the optimal results with the analytical method. It must have three important qualities, including being pure, inert toward sample's mixture and be absent from the sample (Hoffmann and Stroobant 2007).

In previously published work on the analysis of carotenoids in algal samples (Chu *et al.* 2011) from our group, β -apo-8'-carotenal served as internal standard. However, due to the aldehyde reactive site of this molecule, a methylated form of this compound was also observed in the extracted ion chromatogram (Figure 4.4). Hence, the stability of this molecule was in doubt, leading us to change the internal standard for more accurate quantitation.

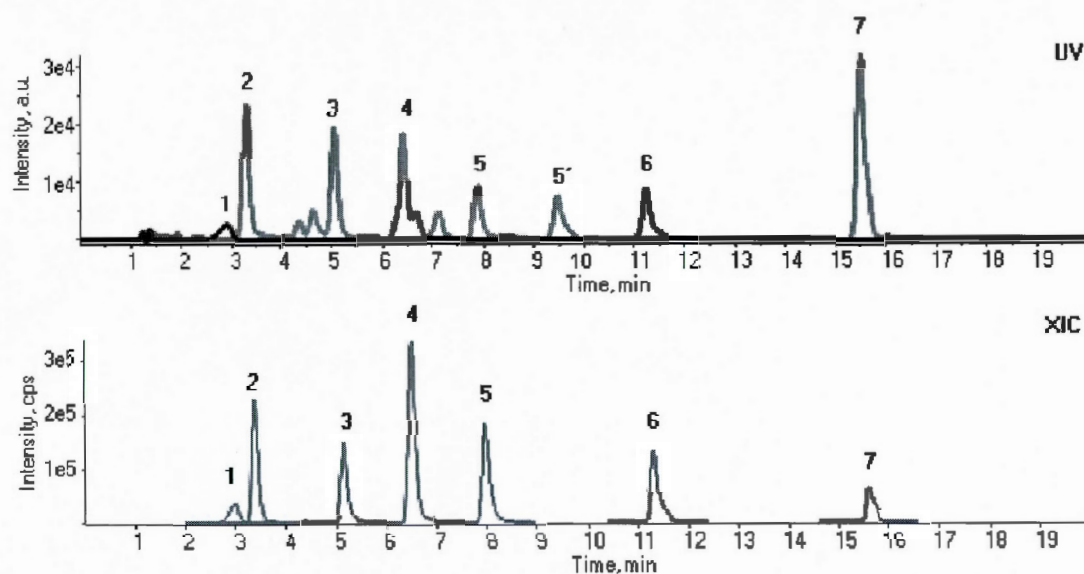


Figure 4.4 UV trace at 450 nm (top) and extracted ion chromatograms (bottom) of standards used in this study: 1) echinenone, 2) astaxanthin, 3) lutein, 4) canthaxanthin, 5) β -apo-8'-carotenal (IS), 5') methylated β -apo- β' -carotenal, 6) echinenone, 7) β -carotene (Meier 2012, unpublished data)

Three candidates were selected to be tested with the HPLC-MS method in order to replace previous internal standard (β -apo- β' -carotenal) (Figure 4.5). These were of vitamin E (alpha-tocopherol), vitamin K2 (Menaquinone) and β -apo-8'-carotenol. Selection criteria for IS candidates was based on having similarity with carotenoid molecular structure, and chromatographic elution profile by HPLC-MS.

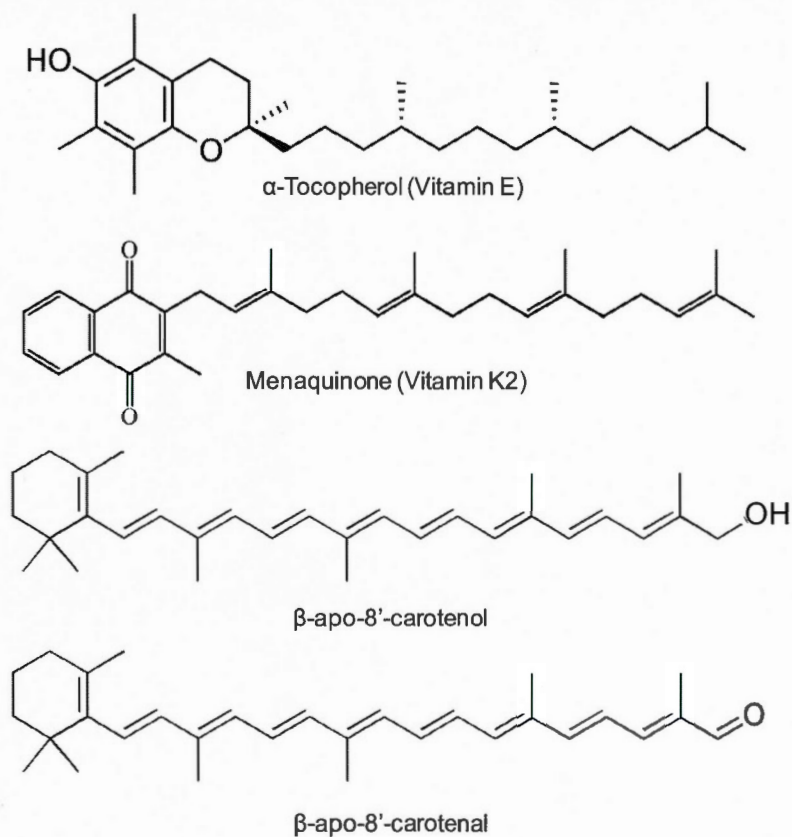


Figure 4.5 Candidate internal standard compounds tested for this study including α -Tocopherol (vitamin E), Menaquinone (vitamin K2) and β -apo-8'-carotenol. β -apo-8'-carotenol was the internal standard previously used for quantification of carotenoids in algal samples

The standard solutions of three candidates were injected onto the LC-UV-MS system and data acquisition was performed with the same method previously developed for quantitation of carotenoids (Chu *et al.* 2011). The HPLC-MS data (Figures 4.6 to 4.8) was investigated based on two criteria: purity of compounds and co-elution of observed ions.

The observation of only one peak in the total ion chromatogram (TIC) of vitamins K2 and E exhibit high purity of both tested compounds, while detecting two peaks for custom-synthesized carotenol is due to the impurity from starting material (5- β -

apo-8'-carotenol) in the solution that implicates incomplete reaction of producing carotenol from carotenal.

As it is shown in Figure 4.8, the LC-MS trace of vitamin K2 shows high purity and good shape as well as no in-source fragment hence, this compound was selected as internal standard for the quantitative study of carotenoids.

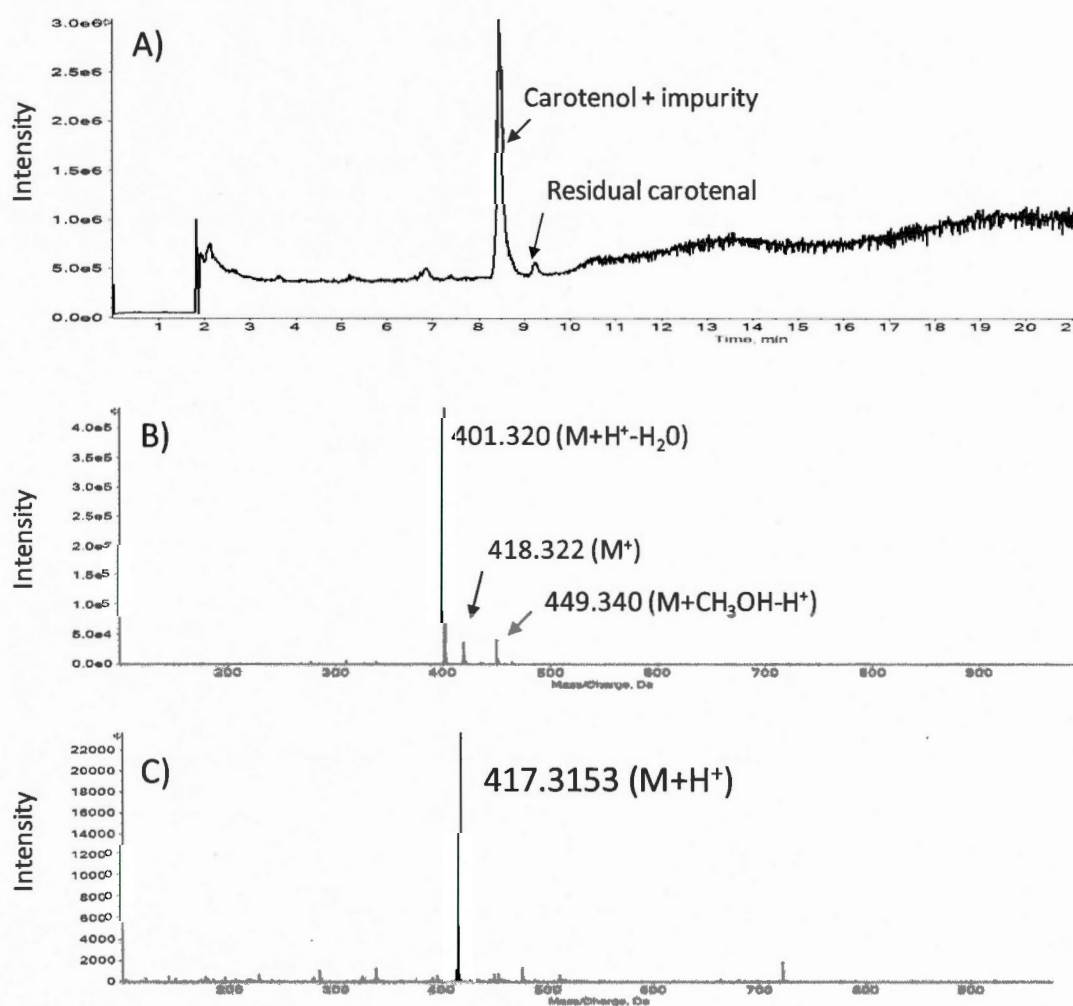


Figure 4.6 Evaluation of purity and coelution of β -apo-8'-carotenol (50 μ g/ml) to be used as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 8.3 to 8.6 min), C) Mass spectrum (from 9.2 to 9.3 min)

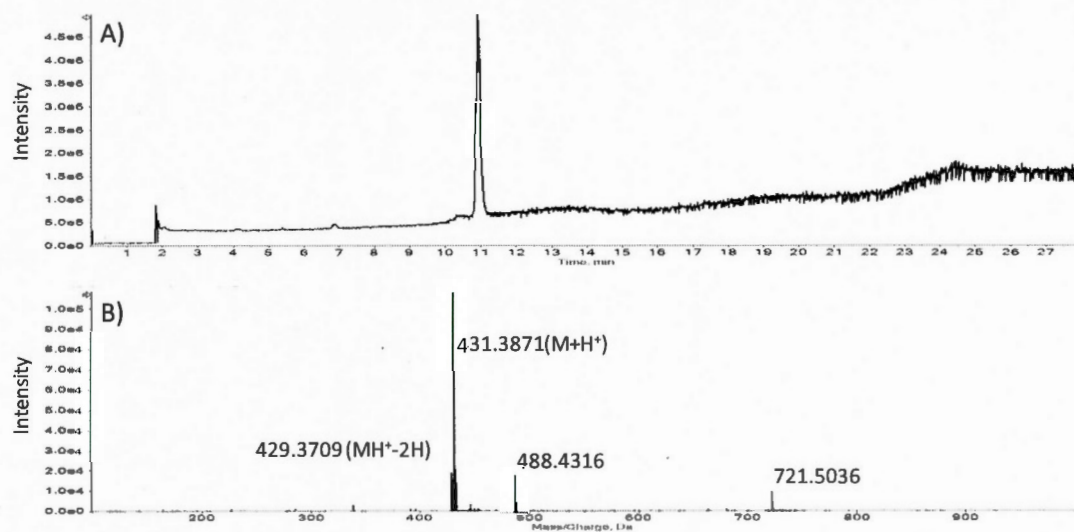


Figure 4.7 Evaluation of purity and co-elution of vitamin E (50 µg/ml) tested as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 10.8 to 11.0 min)

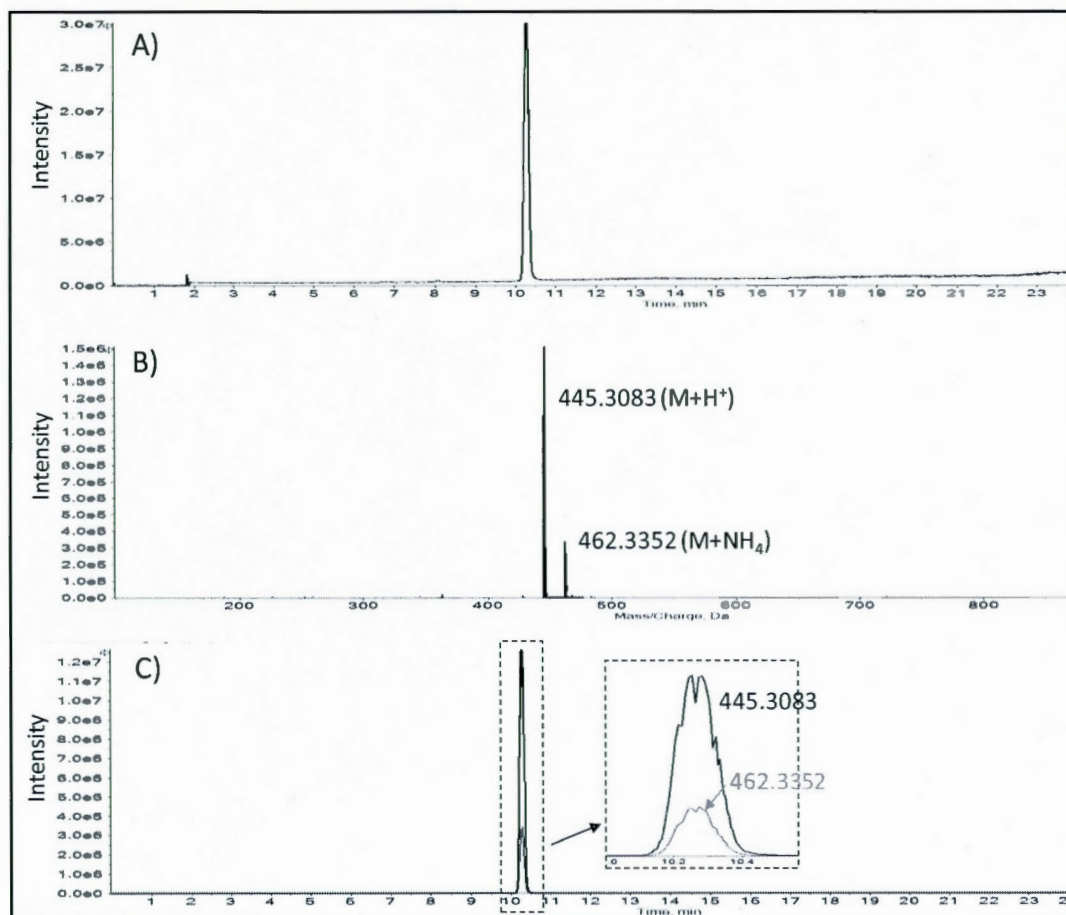


Figure 4.8 Evaluation of purity and coelution of vitamin K2 (50 µg/ml) to be used as internal standard. A) Total ion chromatogram (TIC), B) Mass spectrum (from 10.2 to 10.4 min), C) Extracted ion chromatograms (XICs) which shows perfect co-elution of two observed ions (protonated molecule and ammonium adduct)

Due to the impact of the sample preparation procedure on quantitation, the timing for the addition of internal standard to the algal sample was further investigated. Sample preparation for this study includes four steps: filtration, extraction, evaporation of solvent and reconstitution. Three sets of samples were prepared with the addition of the internal standard at different steps of preparation method. In one set of samples, menaquinone solution (IS) was added prior to filtration, while IS was added after the filtration but prior to extraction step for the second set of samples. IS was added to samples immediately before the injection for a third group

of samples. The result for three set of samples (Figure 4.9) shows different average peak areas for internal standard between experiments. It is demonstrated that peak areas corresponding to the addition of IS in filtration were smaller than those for the extraction. Furthermore, peak areas corresponding to the addition of IS before injection (without extraction) are even higher than two other ones. Loss of menaquinone and higher standard deviation for the average peak area were observed when IS was added prior to filtration, since it is soluble at this step compared to algal cells containing the carotenoid compounds. The addition of internal standard prior to extraction has the advantage of considering possible errors of extraction in comparison with the results from the addition of IS before injection (without extraction). Hence, addition of internal standard prior to the extraction step was chosen for quantitation of carotenoids in algal samples.

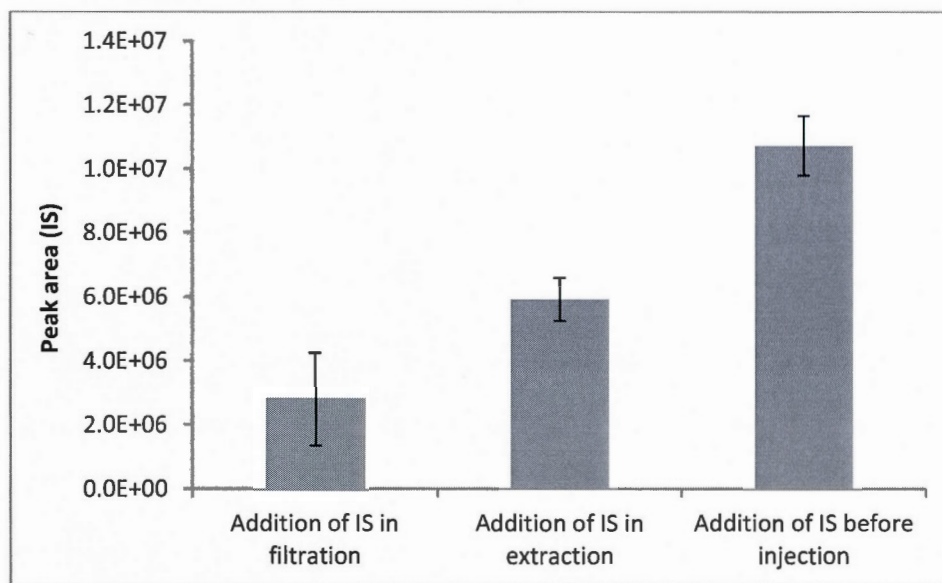


Figure 4.9 Peak areas detected for internal standard when it was added (to the sample in different stages of sample preparation, during filtration, during extraction and right before the injection to HPLC-HRMS system (n=9)

To consider the effects of sample preparation on the amount of quantified analytes (rather than internal standard), sample preparation was also performed for the

standard mixture. To determine the best scenario for sample treatment of standard mixture, an experiment was performed for comparing two sets of MS data from samples prepared differently. One set of samples were prepared like algal samples (filtration, extraction, reconstitution) while the second set of samples were prepared without filtration (extraction, reconstitution). Although the results of this experiment show loss of standard compounds for both data sets (Figure 4.10), the loss of compounds during the filtration step is much more significant than the real algal samples since algal samples used in this project are intact cells (containing carotenoids). Considering the size of algal cells to the pore size of filter, the filtration process has little influence on the biological content of algal medium, meanwhile a much greater effect is seen for chemical solutions such as standards. Hence, loss of compounds in standard solution is much greater than in algal cell. Extraction of carotenoid compounds seems to be the most important cause of possible errors during analytical sample preparation. Therefore, for the preparation of the calibration curve, bead-beating was also done for standard solutions with the same procedure used for algal samples (while filtration was omitted for standards). These results show the importance of applying a similar procedure of sample preparation for biological samples on standard mixtures.

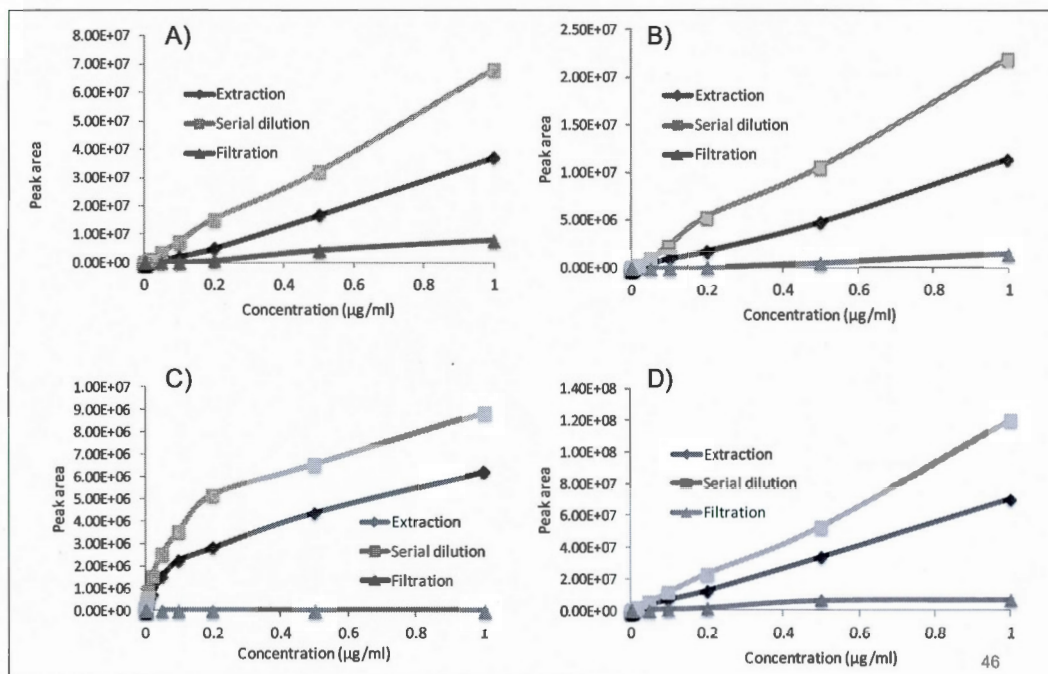


Figure 4.10 Evaluating the influence of sample preparation on peak area for four standard compounds, A) astaxanthin, B) lutein, C) β -Carotene, D) canthaxanthin. Three different sets of standard mixtures were prepared starting with i) filtration ii) extraction or iii) serial dilution of stock solution of compounds to obtain desired concentration (without filtration and extraction).

4.5 Results

The aim of this study was to quantify the changes in carotenoid content of three algal species *Haematococcus*, *Oocystis* and *Muriellopsis* exposed to a specific stress. A long culturing time course of 4 and 6 month (reduced nutrients) was respectively applied to *Haematococcus* and *Oocystis* species which results in reduced nutrients. *Muriellopsis* was cultured in a high salt medium to induce stress. Carotenoid standards, astaxanthin, canthaxanthin, lutein and β -carotene were baseline separated on a reverse phase C18 column (Figure 4.11). Peak assignments were made based on high-resolution accurate mass data and the verification of retention time was done by correlating with UV signal. This result was in

agreement with previous work from our group, where carotenoid structure was confirmed with MS/MS matching (unpublished data).

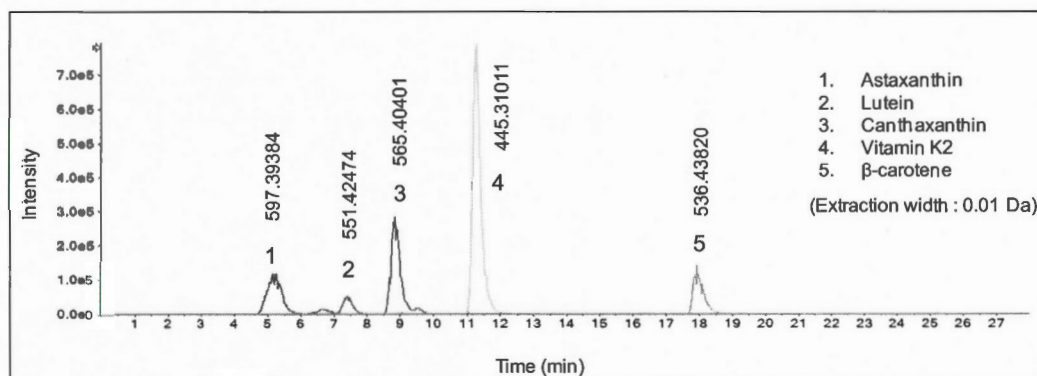


Figure 4.11 Extracted ion chromatograms of standards used in this study: 1-astaxanthin, 2-lutein, 3-canthaxanthin, 4-Vitamin K2 (IS), 5- β -carotene. Extracted m/z are also presented for each peak.

Accurate masses and retention times of carotenoids observed in algae extracts by HPLC-UV-ESI-MS/MS are presented in Table 4.1. All compounds were observed as protonated ions ($[M+H]^+$), except for lutein and β -carotene which were detected as the in-source water loss fragment of the protonated molecule $[M+H-H_2O]^+$ and radical molecular ion $[M^{+\bullet}]$, respectively. As it is shown, all extracted m/z has less than 5 ppm error relative to their corresponding theoretical exact masses.

Table 4.1 Accurate masses and retention times of carotenoids observed in algae extracts by HPLC-UV-ESI-MS/MS

Component name	Formula	Neutral Mass (Da)	Ion species	Extraction mass (m/z)	Error (ppm)
Astaxanthin	C ₄₀ H ₅₂ O ₄	596.38656	M+H ⁺	597.39384	1
Lutein	C ₄₀ H ₅₆ O ₂	568.42803	M-(H ₂ O)+H ⁺	551.42474	0.4
Canthaxanthin	C ₄₀ H ₅₂ O ₂	564.39673	M+H ⁺	565.40401	3.5
β-carotene	C ₄₀ H ₅₆	536.43820	M ⁺ ⊖	536.43820	1.9
Menaquinone (IS)	C ₃₁ H ₄₀ O ₂	444.30283	M+H ⁺	445.31011	3.3

Quantitation of four carotenoids of interest was performed by using a calibration curve for each standard compound with data normalization of data using the menaquinone peak (IS). It was observed that although the protonated ion is more intense for β-carotene, the calibration curve has much better linearity and dynamic range when the molecular ion is quantified. Thus, the molecular ion was used for quantification of β-carotene, while the protonated molecule was utilized for astaxanthin and canthaxanthin. As it was described earlier, water loss in-source fragment $[M+H-H_2O]^+$ of lutein was used as the ion of interest for quantitation. The calibration curve obtained for each compound is presented at the end of this chapter (Figures 4.16 to 4.19).

Induced stress on algal samples results in a change in their color; Green algae were converted into orange (*Murriellopsis* and *Oocystis*) or into red (*Haematococcus*) and this change is a result of change in the carotenoid content of the cells. The results of quantitation followed by t-test calculation show that lutein and β-carotene were down-regulated in *Haematococcus* and *Murriellopsis*, while astaxanthin and cantaxanthin were up-regulated in these samples (p-value <0.05) (Figures 4.12 and 4.13). On the other hand, the concentration of the four quantified carotenoid

compounds were decreased in *Oocystis* algae in response to the stress (p-value<0.05) (figure 4.14). However, considering the orange color of these stressed cultures, other carotenoid compounds may be up-regulated (Figure 4.15).

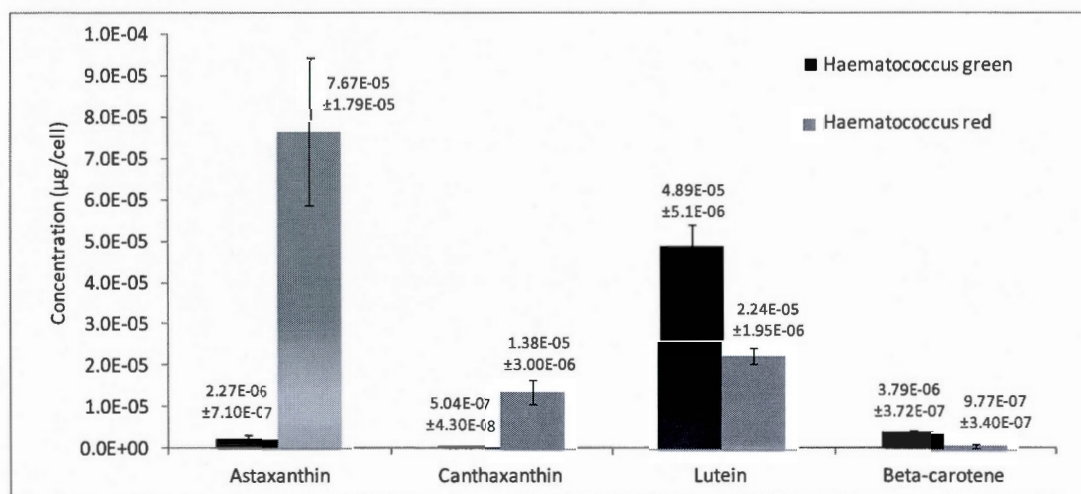


Figure 4.12 The change in carotenoid concentration in *Haematococcus* algae induced by stress condition (culturing time). The quantity of beta-carotene and lutein exist in *Haematococcus* green are higher than the limit of detection of our method for these compounds. The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration (µg/cell) and standard deviation are also shown in the figure (n=3)

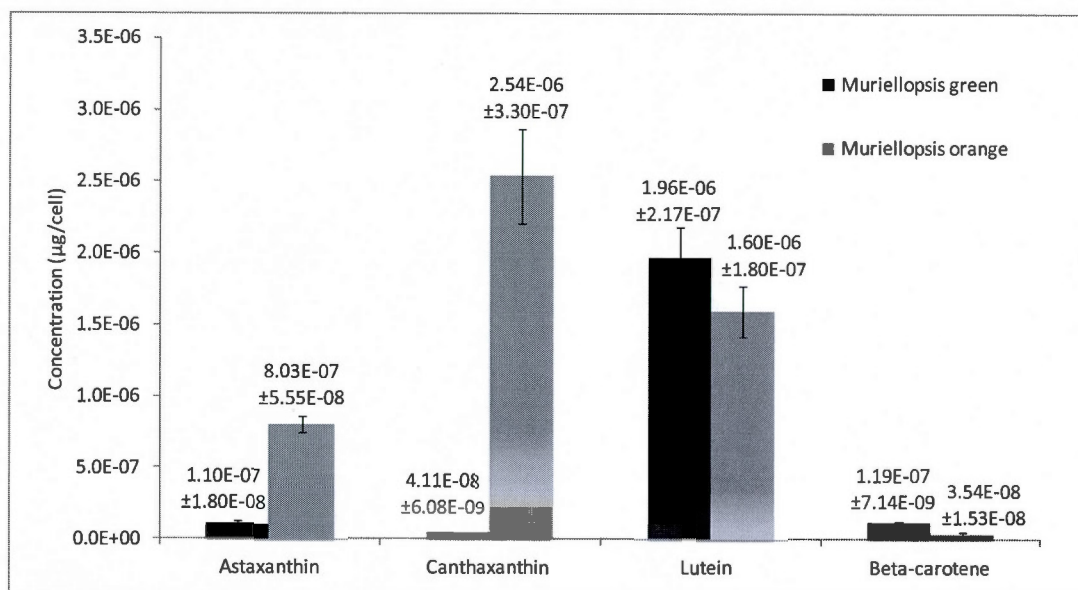


Figure 4.13 Change in carotenoid concentration in *Muriellopsis* algae induced by stress condition (culturing time). The quantity of cantaxanthin exist in *Muriellopsis* orange and also β -carotene exist in *Muriellopsis* green are higher than the limit of detection The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration ($\mu\text{g}/\text{cell}$) and standard deviation are also shown in the figure ($n=3$)

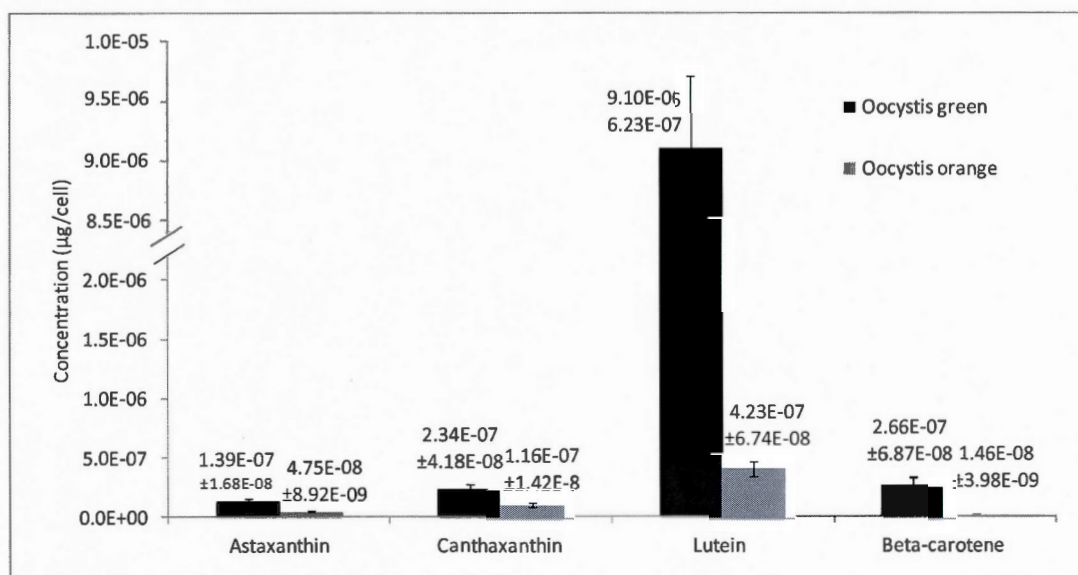


Figure 4.14 The change in carotenoid concentration in *Oocystis* algae induced by stress condition (culturing time). The quantity of β -carotene, cantaxanthin and lutein in *Oocystis* orange, and also β -carotene and lutein in *Oocystis* green is higher than the limit of detection. The cellular density in all cultures was initially determined by optical microscope counting and analytical calculation was performed accordingly. Average concentration ($\mu\text{g}/\text{cell}$) and standard deviation are also shown in the figure ($n=3$)

Table 4.2 Summarized changes in carotenoid content from control sample to stressed algae samples. Based on our experimental results (fold change and t-test) *Muriellopsis* and *Haematococcus* showed an up-regulation (Up) of astaxanthin and canthaxanthin while β -carotene and lutein were down-regulated (Down). All four studied carotenoids were down-regulated in *Oocystis* (fold changes are shown in the table and p-values were all below 0.05)

Algae	Astaxanthin	Fold change	Canthaxanthin	Fold change	β -Carotene	Fold change	Lutein	Fold change
Muriellopsis	Up	7.3	Up	61.9	Down	3.6	Down	1.2
Haematococcus	Up	33.7	Up	27.3	Down	3.9	Down	2.1
Oocystis	Down	3.0	Down	2.0	Down	18.2	Down	7.3

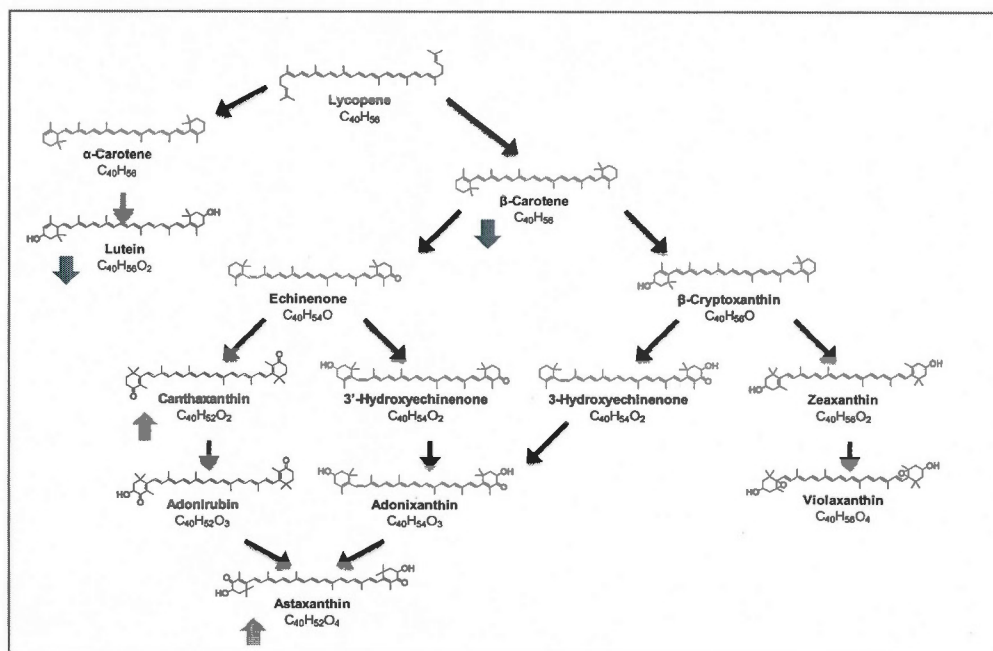


Figure 4.15 Carotenogenesis pathway. It was observed that in *Haematococcus* and *Muriellopsis* algae species, astaxanthin and cantaxanthin were up-regulated (tick up arrows) and decreases of lutein and β -carotene levels were revealed under stress treatment (tick down arrows) (adapted representation based on (Álvarez *et al.* 2006))

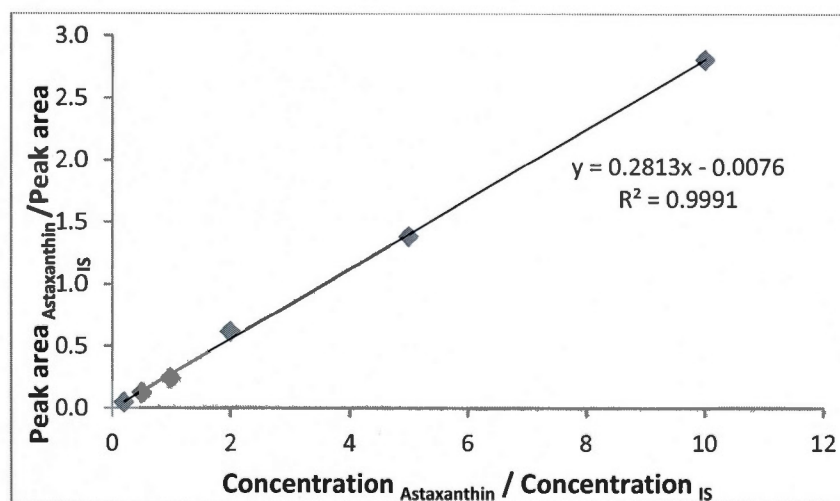


Figure 4.16 Calibration curve obtained for Astaxanthin using linear least squares regression analysis. It covers 0.2 to 10 $\mu\text{g/ml}$ of Astaxanthin in standard solutions (IS concentration was 1 $\mu\text{g/ml}$)

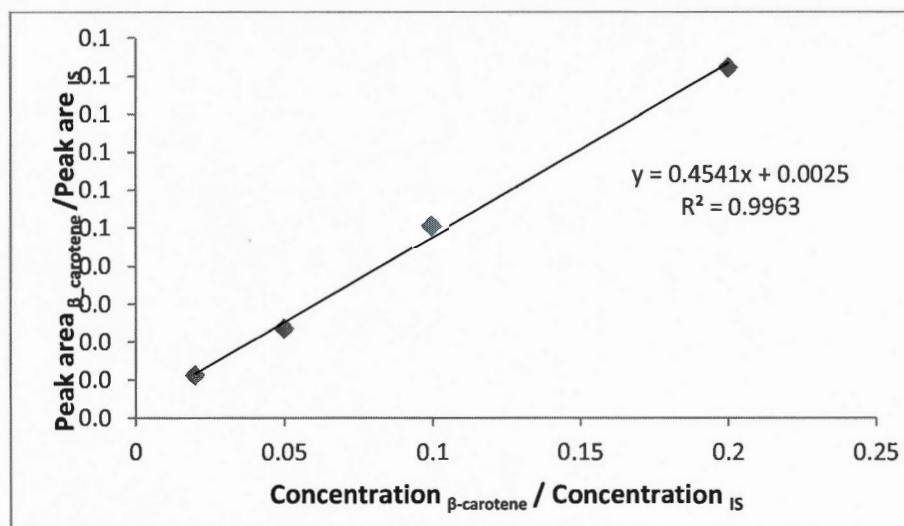


Figure 4.17 Calibration curve obtained for β -Carotene using linear least squares regression analysis. It covers 0.02 to 0.2 $\mu\text{g/ml}$ of β -Carotene in standard solutions (IS concentration of 1 $\mu\text{g/ml}$)

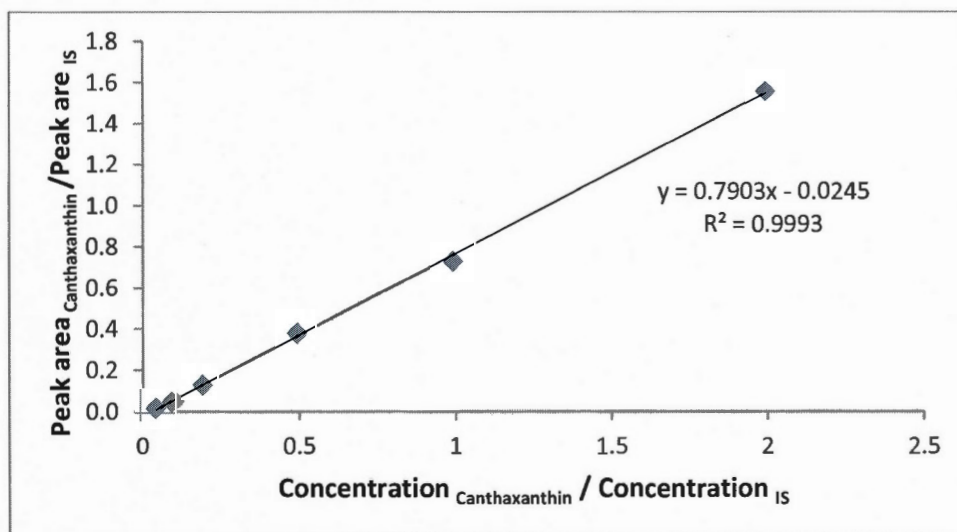


Figure 4.18 Calibration curve obtained for Canthaxanthin using linear least squares regression analysis. It covers 0.05 to 2 $\mu\text{g/ml}$ of Canthaxanthin in standard solutions (IS concentration was 1 $\mu\text{g/ml}$)

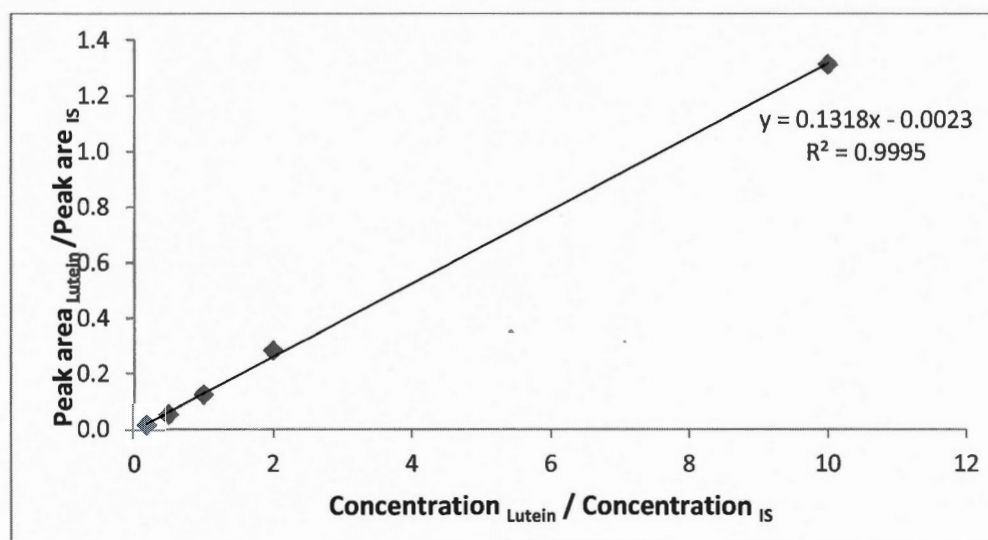


Figure 4.19 Calibration curve obtained for Lutein using linear least squares regression analysis. It covers 0.2 to 10 $\mu\text{g/ml}$ of Lutein in standard solutions (IS concentration was 1 $\mu\text{g/ml}$)

4.6 Conclusions

In this study, a filtration based sample preparation method was developed for quantitation of carotenoids in algal samples. Method development was performed in three steps including a) choice of internal standard, b) timing point for addition of internal standard to the mixtures and c) the sample preparation procedure used for the standard mixtures. From three tested compounds, vitamin k2 was selected as IS due to high purity, good peak shape and having no in-source fragment in corresponding MS spectra. Also, our results showed addition of internal standard prior to the extraction step was more appropriate for quantitation of carotenoids in algal samples due to considering possible errors of extraction. Furthermore, for the preparation of the calibration curve, bead-beating was also done for standard solutions with the same procedure used for algal samples (while filtration was omitted for standards).

Employing the sample preparation developed in this study and a previously developed LC-UV-MS method, four carotenoid compounds were successfully quantified in three algal samples in normal and stress-induced conditions. Based on t-test results, down-regulation of lutein and β -carotene biosynthesis as well as up-regulation of astaxanthin and canthaxanthin were revealed in response to stress conditions in *Haematococcus* and *Muriellopsis* species. Nitrogen deficiency stress induced orange color in *Oocystis* algae although all four surveyed carotenoids were down-regulated. The carotenoid source responsible for this orange colour will be investigated in future experiments using the sample preparation procedure developed in this study.

CHAPTER V

CONCLUSION

Data analysis is an important step in metabolomics studies, starting with peak detection on raw LC-MS data. As there are different peak picking software and algorithms available, the effect of employing different workflows was investigated (Chapter 2). Four peak picking workflows studied in this research include: MarkerView, MetabolitePilot, PeakView and XCMS online. The data analysis was performed on two types of biological samples (bile and urine) as well as a standard mixture of 84 compounds. A custom-made MATLAB code ("VennPro") was used to find the overlaps between the results of peak picking workflows. Interestingly, only a small fraction of detected peaks (7.7 % in average) were found by all workflows. It was shown that none of the studied workflows are perfect and each has advantages and disadvantages. However, MarkerView showed better performance in having bigger overlaps with other workflows. The results obtained in this study exemplify the importance of selecting appropriate peak picking workflows for metabolomics data analysis.

In a parallel research, a simple ion annotation method was developed to identify peaks correspond to isotopic peaks, radical ions, adducts and in-source-fragments and remove them. This MATLAB-based "*DataReduction*" workflow imports data from peak detection workflows (in excel format, sample sheet in supplementary data at the end of this chapter), performs several filtering steps and removes peaks corresponding to isotopic peaks, radical ions, adducts and in-source fragments. The performance of this MATLAB code was evaluated by comparison of its results to the isotopic peak filtering option of MarkerView and there was 85% agreement between them. Furthermore, four peak picking workflows mentioned previously were also evaluated in terms of automated ion annotation. It was observed that isotopic peaks are removed automatically in PeakView and MarkerView although it

was not included in processing options. On the other hand, MarkerView and XCMS can assign isotopic peaks for further removal by the user. The results of this project demonstrate the importance of data processing for LC-MS metabolomics as it deals with huge amounts of data. It seems the informatics tools for metabolomics are still in the developing stages since this area of research is relatively new.

In the fourth chapter of this dissertation, a targeted metabolomics approach was presented for quantifying the change in carotenoid content of algal samples introduced by stress conditions. An extraction-based method was developed followed by absolute quantification of four carotenoid compounds in three algal species, *Haematococcus*, *Oocystis* and *Muriellopsis*. It was observed that lutein and β -carotene were down-regulated in *Haematococcus* and *Muriellopsis* while astaxanthin and canthaxanthin were increased in this species in response to stress conditions. Nitrogen deficiency induced an orange color in *Oocystis* algae although all four surveyed carotenoids were found to be down-regulated. In future research, a more comprehensive study on the biochemical pathway governing carotenoid changes will be useful to declare the mechanism of biochemical response to stress conditions.

REFERENCES

- ACD/IntelliXtract, A. C. D., Inc. www.acdlabs.com/intellixtract, 2007. (2007). Advanced Chemistry Development, Inc.
- Álvarez, V., M. Rodríguez-Sáiz, et al. (2006). "The crtS gene of *Xanthophyllomyces dendrorhous* encodes a novel cytochrome-P450 hydroxylase involved in the conversion of β -carotene into astaxanthin and other xanthophylls." *Fungal Genetics and Biology* **43**(4): 261-272.
- Bakhtiar, R., L. Ramos, et al. (2002). "High-throughput mass spectrometric analysis of xenobiotics in biological fluids." *Journal of liquid chromatography & related technologies* **25**(4): 507-540.
- Bamba, T., N. Shimonishi, et al. (2008). "High throughput and exhaustive analysis of diverse lipids by using supercritical fluid chromatography-mass spectrometry for metabolomics." *Journal of bioscience and bioengineering* **105**(5): 460-469.
- Baroli, I. and K. K. Niyogi (2000). "Molecular genetics of xanthophyll-dependent photoprotection in green algae and plants." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **355**(1402): 1385-1394.
- Bauer, C., R. Cramer, et al. (2011). Evaluation of peak-picking algorithms for protein mass spectrometry. *Data Mining in Proteomics*, Springer: 341-352.
- Beger, R. D., J. Sun, et al. (2010). "Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity." *Toxicology and applied pharmacology* **243**(2): 154-166.
- Ben-Amotz, A., A. Shaish, et al. (1989). "Mode of action of the massively accumulated β -carotene of *Dunaliella bardawil* in protecting the alga against damage by excess irradiation." *Plant Physiology* **91**(3): 1040-1043.
- Borowitzka, M. (1992). "Comparing carotenogenesis in *Dunaliella* and *Haematococcus*: implications for commercial production strategies." *Profiles on Biotechnology. Servicio de Publicacions*: 301-310.
- Botros, L., D. Sakkas, et al. (2008). "Metabolomics and its application for non-invasive embryo assessment in IVF." *Molecular human reproduction* **14**(12): 679-690.
- Bouatra, S., F. Aziat, et al. (2013). "The human urine metabolome." *PloS one* **8**(9): e73076.
- Brooks, C., E. C. Horning, et al. (1968). "Characterization of sterols by gas chromatography-mass spectrometry of the trimethylsilyl ethers." *Lipids* **3**(5): 391-402.
- Brown, M., W. B. Dunn, et al. (2009). "Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics." *Analyst* **134**(7): 1322-1332.
- Brown, M., D. C. Wedge, et al. (2011). "Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets." *Bioinformatics* **27**(8): 1108-1112.

- Bueschl, C., R. Krska, et al. (2013). "Isotopic labeling-assisted metabolomics using LC-MS." Analytical and bioanalytical chemistry **405**(1): 27-33.
- Castillo, S., P. Gopalacharyulu, et al. (2011). "Algorithms and tools for the preprocessing of LC-MS metabolomics data." Chemometrics and Intelligent Laboratory Systems **108**(1): 23-32.
- Chambers, M. C., B. Maclean, et al. (2012). "A cross-platform toolkit for mass spectrometry and proteomics." Nature biotechnology **30**(10): 918-920.
- Chernushevich, I. V., A. V. Loboda, et al. (2001). "An introduction to quadrupole-time-of-flight mass spectrometry." Journal of Mass Spectrometry **36**(8): 849-865.
- Chu, F. L., L. Pirastru, et al. (2011). "Carotenogenesis Up-regulation in *Scenedesmus* sp. Using a Targeted Metabolomics Approach by Liquid Chromatography-High-Resolution Mass Spectrometry." Journal of agricultural and food chemistry **59**(7): 3004-3013.
- Cooper, D. A., A. L. Eldridge, et al. (1999). "Dietary carotenoids and lung cancer: a review of recent research." Nutrition reviews **57**(5): 133-145.
- Creek, D. J., A. Jankevics, et al. (2012). "IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data." Bioinformatics **28**(7): 1048-1049.
- Davies, B. H. (1985). "Carotenoid metabolism in animals: a biochemist's view." Pure and Applied Chemistry **57**(5): 679-684.
- Demmig-Adams, B. and W. W. Adams (2002). "Antioxidants in photosynthesis and human nutrition." Science **298**(5601): 2149-2153.
- Dettmer, K., P. A. Aronov, et al. (2007). "Mass spectrometry-based metabolomics." Mass spectrometry reviews **26**(1): 51-78.
- Dettmer, K. and B. D. Hammock (2004). "Metabolomics--a new exciting field within the "omics" sciences." Environmental health perspectives **112**(7): A396.
- Dixon, R. A., D. R. Gang, et al. (2006). "Applications of metabolomics in agriculture." Journal of agricultural and food chemistry **54**(24): 8984-8994.
- Dunn, W. B. and D. I. Ellis (2005). "Metabolomics: current analytical platforms and methodologies." TrAC Trends in Analytical Chemistry **24**(4): 285-294.
- Eilers, P. H. and H. F. Boelens (2005). "Baseline correction with asymmetric least squares smoothing." Leiden University Medical Centre Report.
- El-Baky, H. H. A., F. K. El Baz, et al. (2003). "Spirulina species as a source of carotenoids and a-tocopherol and its anticarcinoma factors."
- Elevated, C., O. To, et al. (1997). "Plasma homocysteine as a risk factor for vascular disease." JAMA **277**: 1775-1781.
- Fassett, R. G. and J. S. Coombes (2011). "Astaxanthin: a potential therapeutic agent in cardiovascular disease." Marine drugs **9**(3): 447-465.
- Feist, A. M., C. S. Henry, et al. (2007). "A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information." Molecular systems biology **3**(1).

- Ferruzzi, M. G., L. C. Sander, et al. (1998). "Carotenoid determination in biological microsamples using liquid chromatography with a coulometric electrochemical array detector." *Analytical biochemistry* **256**(1): 74-81.
- Fiehn, O. (2002). "Metabolomics—the link between genotypes and phenotypes." *Plant molecular biology* **48**(1-2): 155-171.
- Finckh, B., A. Kontush, et al. (1995). "Monitoring of ubiquinol-10, ubiquinone-10, carotenoids, and tocopherols in neonatal plasma microsamples using high-performance liquid chromatography with coulometric electrochemical detection." *Analytical biochemistry* **232**(2): 210-216.
- Fraser, P. D., E. Enfissi, et al. (2007). "Metabolite profiling of plant carotenoids using the matrix-assisted laser desorption ionization time-of-flight mass spectrometry." *The Plant Journal* **49**(3): 552-564.
- Fu, W., M. Magnúsdóttir, et al. (2012). "UPLC-UV-MSE analysis for quantification and identification of major carotenoid and chlorophyll species in algae." *Analytical and bioanalytical chemistry* **404**(10): 3145-3154.
- Gowda, G. N., S. Zhang, et al. (2008). "Metabolomics-based methods for early disease diagnostics."
- Granado, F., B. Olmedilla, et al. (2001). "A fast, reliable and low-cost saponification protocol for analysis of carotenoids in vegetables." *Journal of Food Composition and Analysis* **14**(5): 479-489.
- Griffiths, W. J., T. Koal, et al. (2010). "Targeted metabolomics for biomarker discovery." *Angewandte Chemie International Edition* **49**(32): 5426-5445.
- Group, N. D. D. (1979). "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance." *Diabetes* **28**(12): 1039-1057.
- Grünewald, K., J. Hirschberg, et al. (2001). "Ketocarotenoid biosynthesis outside of plastids in the unicellular green alga *Haematococcus pluvialis*." *Journal of Biological Chemistry* **276**(8): 6023-6029.
- Haimi, P., A. Uphoff, et al. (2006). "Software tools for analysis of mass spectrometric lipidome data." *Analytical chemistry* **78**(24): 8324-8331.
- Hall, R. D. (2011). "Plant metabolomics in a nutshell: potential and future challenges." *Annual Plant Reviews, Biology of Plant Metabolomics* **43**: 1.
- Harrigan, G. G., R. H. LaPlante, et al. (2004). "Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity." *Toxicology Letters* **146**(3): 197-205.
- Herbert, R. B. (1989). *The biosynthesis of secondary metabolites*, Springer.
- Herrgård, M. J., N. Swainston, et al. (2008). "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology." *Nature biotechnology* **26**(10): 1155-1160.
- Hilario, M., A. Kalousis, et al. (2006). "Processing and classification of protein mass spectra." *Mass spectrometry reviews* **25**(3): 409-449.
- Hoffmann, E. and V. Stroobant (2007). *Mass spectrometry: principles and applications 2007*, John Wiley and sons.

- Huang, E. C., T. Wachs, et al. (1990). "Atmospheric pressure ionization mass spectrometry." *Analytical chemistry* **62**(13): 713A-725A.
- Huang, N., M. M. Siegel, et al. (1999). "Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data." *Journal of the American Society for Mass Spectrometry* **10**(11): 1166-1173.
- Hughes, D. A. (2001). "Dietary carotenoids and human immune function." *Nutrition* **17**(10): 823-827.
- Jaitly, N., A. Mayampurath, et al. (2009). "Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data." *BMC bioinformatics* **10**(1): 87.
- Jin, E., J. E. Polle, et al. (2003). "Xanthophylls in microalgae: From biosynthesis to biotechnological mass production and application." *Journal of microbiology and biotechnology* **13**(2): 165-174.
- Johnson, E. J. (2002). "The role of carotenoids in human health." *Nutrition in Clinical Care* **5**(2): 56-65.
- Kanehisa, M., S. Goto, et al. (2004). "The KEGG resource for deciphering the genome." *Nucleic acids research* **32**(suppl 1): D277-D280.
- Karp, P. D., C. A. Ouzounis, et al. (2005). "Expansion of the BioCyc collection of pathway/genome databases to 160 genomes." *Nucleic acids research* **33**(19): 6083-6089.
- Katajamaa, M., J. Miettinen, et al. (2006). "MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data." *Bioinformatics* **22**(5): 634-636.
- Katajamaa, M. and M. Orešič (2005). "Processing methods for differential analysis of LC/MS profile data." *BMC bioinformatics* **6**(1): 179.
- Katajamaa, M. and M. Orešič (2007). "Data processing for mass spectrometry-based metabolomics." *Journal of Chromatography A* **1158**(1): 318-328.
- Keller, B. O., J. Sui, et al. (2008). "Interferences and contaminants encountered in modern mass spectrometry." *analytica chimica acta* **627**(1): 71-81.
- Krinsky, N. I. and E. J. Johnson (2005). "Carotenoid actions and their relation to health and disease." *Molecular aspects of medicine* **26**(6): 459-516.
- Kuhl, C., R. Tautenhahn, et al. (2011). "CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets." *Analytical chemistry* **84**(1): 283-289.
- Lawton, K. A., A. Berger, et al. (2008). "Analysis of the adult human plasma metabolome."
- Leavens, W. J., S. J. Lane, et al. (2002). "Derivatization for liquid chromatography/electrospray mass spectrometry: synthesis of tris (trimethoxyphenyl) phosphonium compounds and their derivatives of amine and carboxylic acids." *Rapid Communications in Mass Spectrometry* **16**(5): 433-441.

- Li, X.-j., C. Y. Eugene, et al. (2005). "A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry." Molecular & cellular proteomics **4**(9): 1328-1340.
- Lim, C., J. Kim, et al. (2010). "Identification, fermentation, and bioactivity against *Xanthomonas oryzae* of antimicrobial metabolites isolated from *Phomopsis longicolla* S1B4." Journal of microbiology and biotechnology **20**(3): 494-500.
- Lindon, J. C., J. K. Nicholson, et al. (2011). The handbook of metabonomics and metabolomics, Elsevier.
- Lindon, J. C., J. K. Nicholson, et al. (2003). "Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project." Toxicology and applied pharmacology **187**(3): 137-146.
- Lommen, A. (2009). "MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing." Analytical chemistry **81**(8): 3079-3086.
- Mandal, R., A. C. Guo, et al. (2012). "Multi-platform characterization of the human cerebrospinal fluid metabolome: a comprehensive and quantitative update." Genome medicine **4**(4): 1-11.
- Mapelli, V., L. Olsson, et al. (2008). "Metabolic footprinting in microbiology: methods and applications in functional genomics and biotechnology." TRENDS in Biotechnology **26**(9): 490-497.
- Margalith, P. (1999). "Production of ketocarotenoids by microalgae." Applied microbiology and biotechnology **51**(4): 431-438.
- Matsumoto, H., Y. Ikoma, et al. (2007). "Quantification of carotenoids in citrus fruit by LC-MS and comparison of patterns of seasonal changes for carotenoids among citrus varieties." Journal of agricultural and food chemistry **55**(6): 2356-2368.
- McNaught, A. D. and A. Wilkinson (2000). IUPAC Compendium of chemical terminology, International Union of Pure and Applied Chemistry.
- Metz, T. O., Q. Zhang, et al. (2007). "Future of liquid chromatography-mass spectrometry in metabolic profiling and metabolomic studies for biomarker discovery."
- Nam, H., B. C. Chung, et al. (2009). "Combining tissue transcriptomics and urine metabolomics for breast cancer biomarker identification." Bioinformatics **25**(23): 3151-3157.
- Nicholson, J. K., J. C. Lindon, et al. (1999). "'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data." Xenobiotica **29**(11): 1181-1189.
- Olaizola, M. (2000). "Commercial production of astaxanthin from *Haematococcus pluvialis* using 25,000-liter outdoor photobioreactors." Journal of Applied Phycology **12**(3-5): 499-506.

- Oliveira, E. and D. Watson (2001). "Chromatographic techniques for the determination of putative dietary anticancer compounds in biological fluids." Journal of Chromatography B: Biomedical Sciences and Applications **764**(1): 3-25.
- Oliver, S. G., M. K. Winson, et al. (1998). "Systematic functional analysis of the yeast genome." TRENDS in Biotechnology **16**(9): 373-378.
- Orešič, M., C. B. Clish, et al. (2004). "Phenotype characterisation using integrated gene transcript, protein and metabolite profiling." Applied bioinformatics **3**(4): 205-217.
- Pauling, L., A. B. Robinson, et al. (1971). "Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography." Proceedings of the National Academy of Sciences **68**(10): 2374-2376.
- Plumb, R., J. Castro-Perez, et al. (2004). "Ultra-performance liquid chromatography coupled to quadrupole-orthogonal time-of-flight mass spectrometry." Rapid Communications in Mass Spectrometry **18**(19): 2331-2337.
- Podwojski, K., A. Fritsch, et al. (2009). "Retention time alignment algorithms for LC/MS data must consider nonlinear shifts." Bioinformatics: btp052.
- Preet, A., T. M. Karve, et al. (2012). "Metabolomics: approaches and applications to diabetes research." J Diabetes Metab **6**(001).
- Psychogios, N., D. D. Hau, et al. (2011). "The human serum metabolome." PLoS one **6**(2): e16957.
- Radulovic, D., S. Jelveh, et al. (2004). "Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry." Molecular & cellular proteomics **3**(10): 984-997.
- Rafiei, A. and L. Sleno (2015). "Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis." Rapid Commun. Mass Spectrom **29**: 1-9.
- Ramirez, T., M. Daneshian, et al. (2013). "Metabolomics in toxicology and preclinical research." Altex **30**(2): 209.
- Roscini, L., L. Corte, et al. (2010). "Influence of cell geometry and number of replicas in the reproducibility of whole cell FTIR analysis." Analyst **135**(8): 2099-2105.
- Rossi, D. T. and M. Sinz (2001). Mass spectrometry in drug discovery, CRC Press.
- Rousu, T., O. Pelkonen, et al. (2009). "Rapid detection and characterization of reactive drug metabolites in vitro using several isotope-labeled trapping agents and ultra-performance liquid chromatography/time-of-flight mass spectrometry." Rapid Communications in Mass Spectrometry **23**(6): 843-855.
- Roux, A., D. Lison, et al. (2011). "Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review." Clinical biochemistry **44**(1): 119-135.
- Ryan, D. and K. Robards (2006). "Metabolomics: the greatest omics of them all?" Analytical chemistry **78**(23): 7954-7958.

- Scalbert, A., L. Brennan, et al. (2009). "Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research." Metabolomics **5**(4): 435-458.
- Scholz, M., S. Gatzek, et al. (2004). "Metabolite fingerprinting: detecting biological features by independent component analysis." Bioinformatics **20**(15): 2447-2454.
- Seagle, C., M. A. Christie, et al. (2008). "High-throughput nuclear magnetic resonance metabolomic footprinting for tissue engineering." Tissue Engineering Part C: Methods **14**(2): 107-118.
- Shargel, L., B. Andrew, et al. (2005). Applied biopharmaceutics & pharmacokinetics, McGraw-Hill New York:.
- Smith, C. A., G. O'Maille, et al. (2005). "METLIN: a metabolite mass spectral database." Therapeutic drug monitoring **27**(6): 747-751.
- Smith, C. A., E. J. Want, et al. (2006). "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification." Analytical chemistry **78**(3): 779-787.
- Spratlin, J. L., N. J. Serkova, et al. (2009). "Clinical applications of metabolomics in oncology: a review." Clinical Cancer Research **15**(2): 431-440.
- Stein, J. R. (1980). Handbook of physiological methods: culture methods and growth measurements, CUP Archive.
- Sturm, M., A. Bertsch, et al. (2008). "OpenMS—an open-source software framework for mass spectrometry." BMC bioinformatics **9**(1): 163.
- Sud, M., E. Fahy, et al. (2007). "Lmsd: lipid maps structure database." Nucleic acids research **35**(suppl 1): D527-D532.
- Sumner, L. W., P. Mendes, et al. (2003). "Plant metabolomics: large-scale phytochemistry in the functional genomics era." Phytochemistry **62**(6): 817-836.
- Takeda, I., C. Stretch, et al. (2009). "Understanding the human salivary metabolome." NMR in Biomedicine **22**(6): 577-584.
- Tautenhahn, R., C. Böttcher, et al. (2008). "Highly sensitive feature detection for high resolution LC/MS." BMC bioinformatics **9**(1): 504.
- Taylor, K. L., A. E. Brackenridge, et al. (2006). "High-performance liquid chromatography profiling of the major carotenoids in *Arabidopsis thaliana* leaf tissue." Journal of Chromatography A **1121**(1): 83-91.
- Theodoridis, G., H. G. Gika, et al. (2008). "LC-MS-based methodology for global metabolite profiling in metabonomics/metabolomics." TrAC Trends in Analytical Chemistry **27**(3): 251-260.
- Theodoridis, G. A., H. G. Gika, et al. (2012). "Liquid chromatography-mass spectrometry based global metabolite profiling: a review." analytica chimica acta **711**: 7-16.
- Trethewey, R. N. (2004). "Metabolite profiling as an aid to metabolic engineering in plants." Current opinion in plant biology **7**(2): 196-201.
- Tugizimana, F., L. Piater, et al. (2013). "Plant metabolomics: A new frontier in phytochemical analysis." South African Journal of Science **109**(5-6): 01-11.

- Turner, W. S., C. Seagle, et al. (2008). "Nuclear Magnetic Resonance Metabolomic Footprinting of Human Hepatic Stem Cells and Hepatoblasts Cultured in Hyaluronan-Matrix Hydrogels." *Stem Cells* **26**(6): 1547-1555.
- van Breemen, R. B., D. Nikolic, et al. (1998). "Development of a method for quantitation of retinol and retinyl palmitate in human serum using high-performance liquid chromatography-atmospheric pressure chemical ionization-mass spectrometry." *Journal of Chromatography A* **794**(1): 245-251.
- van het Hof, K. H., B. C. de Boer, et al. (2000). "Carotenoid bioavailability in humans from tomatoes processed in different ways determined from the carotenoid response in the triglyceride-rich lipoprotein fraction of plasma after a single consumption and in plasma after four days of consumption." *The Journal of nutrition* **130**(5): 1189-1196.
- Vandenbogaert, M., S. Li-Thiao-Té, et al. (2008). "Alignment of LC-MS images, with applications to biomarker discovery and protein identification." *Proteomics* **8**(4): 650-672.
- Varghese, R. S., B. Zhou, et al. (2012). "Ion annotation-assisted analysis of LC-MS based metabolomic experiment." *Proteome Sci* **10**(Suppl 1): S8.
- Varghese, R. S., B. Zhou, et al. (2012). "Ion annotation-assisted analysis of LC-MS based metabolomic experiment." *Proteome science* **10**(Suppl 1): S8.
- Villas-Bôas, S. G., S. Rasmussen, et al. (2005). "Metabolomics or metabolite profiles?" *TRENDS in Biotechnology* **23**(8): 385-386.
- Wang, W., H. Zhou, et al. (2003). "Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards." *Analytical chemistry* **75**(18): 4818-4826.
- Wang, Y., J. Xiao, et al. (2009). "PubChem: a public information system for analyzing bioactivities of small molecules." *Nucleic acids research* **37**(suppl 2): W623-W633.
- Want, E. J., A. Nordström, et al. (2007). "From exogenous to endogenous: the inevitable imprint of mass spectrometry in metabolomics." *Journal of proteome research* **6**(2): 459-468.
- Watkins, S. M. and J. B. German (2002). "Toward the implementation of metabolomic assessments of human health and nutrition." *Current opinion in biotechnology* **13**(5): 512-516.
- Weljie, A. M., J. Newton, et al. (2006). "Targeted profiling: quantitative analysis of ¹H NMR metabolomics data." *Analytical chemistry* **78**(13): 4430-4442.
- Wishart, D. S. (2008). "Applications of metabolomics in drug discovery and development." *Drugs in R & D* **9**(5): 307-322.
- Wishart, D. S. (2008). "DrugBank and its relevance to pharmacogenomics."
- Wishart, D. S. (2008). "Metabolomics: applications to food science and nutrition research." *Trends in Food Science & Technology* **19**(9): 482-493.
- Wishart, D. S., T. Jewison, et al. (2012). "HMDB 3.0—the human metabolome database in 2013." *Nucleic acids research*: gks1065.

- Wishart, D. S., M. J. Lewis, et al. (2008). "The human cerebrospinal fluid metabolome." Journal of Chromatography B **871**(2): 164-173.
- Wishart, D. S., D. Tzur, et al. (2007). "HMDB: the human metabolome database." Nucleic acids research **35**(suppl 1): D521-D526.
- Xia, J., R. Mandal, et al. (2012). "MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis." Nucleic acids research **40**(W1): W127-W133.
- Xia, J., N. Psychogios, et al. (2009). "MetaboAnalyst: a web server for metabolomic data analysis and interpretation." Nucleic acids research **37**(suppl 2): W652-W660.
- Yang, C., Z. He, et al. (2009). "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis." BMC bioinformatics **10**(1): 4.
- Yap, S. P., T. Julianto, et al. (1999). "Simple high-performance liquid chromatographic method for the determination of tocotrienols in human plasma." Journal of Chromatography B: Biomedical Sciences and Applications **735**(2): 279-283.
- Young, A. J. and G. M. Lowe (2001). "Antioxidant and prooxidant properties of carotenoids." Archives of Biochemistry and Biophysics **385**(1): 20-27.
- Zhang, J., E. Gonzalez, et al. (2009). "Review of peak detection algorithms in liquid-chromatography-mass spectrometry." Current genomics **10**(6): 388.
- Zhang, S., T. J. DeGraba, et al. (2009). "A novel peak detection approach with chemical noise removal using short-time FFT for pTOF MS data." Proteomics **9**(15): 3833-3842.
- Zhang, W., M. W. Saif, et al. (2010). "Identification of chemicals and their metabolites from PHY906, a Chinese medicine formulation, in the plasma of a patient treated with irinotecan and PHY906 using liquid chromatography/tandem mass spectrometry (LC/MS/MS)." Journal of Chromatography A **1217**(37): 5785-5793.
- Zhou, B., J. F. Xiao, et al. (2012). "LC-MS-based metabolomics." Molecular BioSystems **8**(2): 470-481.