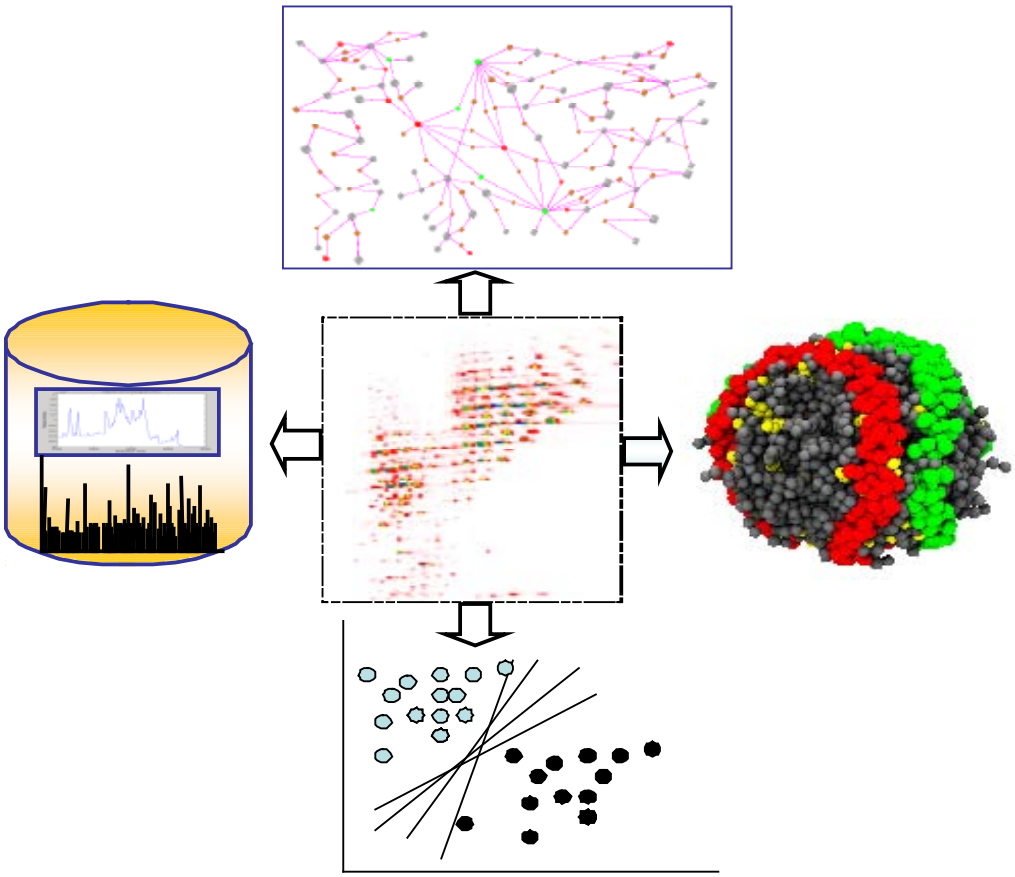


Laxmana Rao Yetukuri

Bioinformatics approaches for the analysis of lipidomics data

VTT PUBLICATIONS 741

Bioinformatics approaches for the analysis of lipidomics data

Laxmana Rao Yetukuri

Department of Biomedical Engineering and Computational Sciences (BECS)

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences, The Aalto University School of Science and Technology, for public examination and debate in Auditorium F239a at Aalto University (Otakaari 3, Espoo, Finland) on 11th of June, 2010 at 12 noon.

ISBN 978-951-38-7402-5 (soft back ed.)
ISSN 1235-0621 (soft back ed.)

ISBN 978-951-38-7403-2 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)

JULKAISIJA – UTGIVARE – PUBLISHER

VTT, Vuorimiehentie 5, PL 1000, 02044 VTT
puh. vaihde 020 722 111, faksi 020 722 4374

VTT, Bergsmansvägen 5, PB 1000, 02044 VTT
tel. växel 020 722 111, fax 020 722 4374

VTT Technical Research Centre of Finland, Vuorimiehentie 5, P.O. Box 1000, FI-02044 VTT, Finland
phone internat. +358 20 722 111, fax + 358 20 722 4374

Technical editing Maini Manninen

Edita Prima Oy, Helsinki 2010

Laxmana Rao Yetukuri. Bioinformatics approaches for the analysis of lipidomics data. Espoo 2010. VTT Publications 741. 75 p. + app. 106 p.

Keywords Lipids, Lipidomics, Bioinformatics, Lipid pathways, High density lipoproteins, k-nearest neighbours, Liquid chromatography/mass spectrometry, Principal component analysis, Partial least squares and discriminant analysis, Obesity, Support vector machines, LipidDB

Abstract

The potential impact of lipid research has been increasingly realised both in disease treatment and prevention. Recent advances in soft ionization mass spectrometry (MS) such as electrospray ionization (ESI) have permitted parallel monitoring of several hundreds of lipids in a single experiment and thus facilitated lipidomics level studies. These advances, however, pose a greater challenge for bioinformaticians to handle massive amounts of information-rich MS data from modern analytical instruments in order to understand complex functions of lipids**.** The main aims of this thesis were to 1) develop bioinformatics approaches for lipid identification based on ultra performance liquid chromatography coupled to mass spectrometry (UPLC/MS) data, 2) predict the functional annotations for unidentified lipids, 3) understand the *omics* data in the context of pathways and 4) apply existing chemometric methods for exploratory data analysis as well as biomarker discovery.

A bioinformatics strategy for the construction of lipid database for major classes of lipids is presented using simplified molecular input line entry system (SMILES) approach. The database was annotated with relevant information such as lipid names including short names, SMILES information, scores, molecular weight, monoisotopic mass, and isotope distribution. The database was tailored for UPLC/MS experiments by incorporating the information such as retention time range, adduct information and main fragments to screen for the potential lipids. This database information facilitated building experimental tandem mass spectrometry libraries for different biological tissues.

Non-targeted metabolomics screening is often get plagued by the presence of unknown peaks and thus present an additional challenge for data interpretation. Multiple supervised classification methods were employed and compared for the functional prediction of class labels for unidentified lipids to facilitate exploratory analysis further as well as ease the identification process. As lipidomics goes

beyond complete characterization of lipids, new strategies were developed to understand lipids in the context of pathways and thereby providing insights for the phenotype characterization. Chemometric methods such as principal component analysis (PCA) and partial least squares and discriminant analysis (PLS/DA) were utilised for exploratory analysis as well as biomarker discovery in the context of different disease phenotypes.

Preface

This thesis work has been carried out at Quantitative Biology and Bioinformatics (QBIX) group at VTT Technical Research Center of Finland (VTT). I derive great pleasure in thanking people who made this thesis a reality. Firstly, I would like to express my deep gratitude to my thesis advisor Research Professor Matej Orešič for providing this opportunity to pursue my PhD work in his fast growing QBIX group. The QBIX group gave me ample opportunity to work in the amazing field of lipidomics, which, in the beginning, was my unexplored territory of science. Prof. Orešič has been a constant source of encouragement, new ideas and inspiration throughout my PhD period. I would like to thank my co-advisor Dr. Jaakko Hollmén, Chief Research Scientist at Aalto University, for providing encouragement, scientific guidance and constructive criticism towards my thesis work. I wish to thank Prof. Kimmo Kaski who readily accepted me as his PhD student at Aalto University and followed the progress of my thesis. I greatly acknowledge my thesis pre-examiners Docent Pentti Somerharju and Prof. Seppo Auriola for their valuable comments which helped me present my thesis in a more meaningful way. I extend my sincere thanks to Technology Manager Dr. Richard Fagerström for his timely help in all walks of life at VTT. I am grateful to our former Vice President Prof. Juha Ahvenainen, Vice President Prof. Anu Kaukovirta-Norja and Research Professor Hans Söderlund for providing nice working facilities and environment for my research.

This thesis work would not have been possible without pleasant co-operation from my colleagues in our group. I thank the team leader of Biosystems Modeling, Dr. Marko Sysi-aho who encouraged and supported me to work on my thesis. I thank my close friend Mr. Gopalacharyulu Peddinti (Gopal) for his insightful scientific and non-scientific (heated) discussions. I like to thank Mr. Erno Lindfors and Mr. Mikko Katajamaa for developing software tools which I used for my research. I especially thank Erno, with whom I share my office place, for translating Finnish language documents for me.

I may be working in front of computer but data are coming from analytical lab. I greatly recognize the importance of all lab people who helped either directly or indirectly in generating the data. I thank the team leader of Metabolomics Dr. Tuulia Hyötyläinen for her continued support in this multi-disciplinary research. My special thanks go to Dr. Vidya Velagapudi and Dr. Tuulikki Seppänen-Laakso who generated most of the data on which this PhD is mainly based on. I thank Dr. Tapani Suortti and former team leader Dr. Kim Ekroos and Dr. Heli Nygren for broadening my analytical knowledge. I thank Dr. Mika Hilvo, Ms. Sandra Castillo, Mr. Pekka Savolahti, Mr. Matti Kankainen, Dr. Jing Tang, Dr. Catherine Bounsaythip, Dr. Perttu S. Niemelä and Mr. Han Zhao for their kind co-operation, friendly environment and scientific discussions. I wish to thank my previous supervisor at VTT, the team leader of Metabolic Engineering group, Dr. Laura Ruohonen for her guidance in yeast proteomics studies. I greatly acknowledge Dr. Marilyn Wiebe, Mrs. Eija Rintala and Dr. Laura Salusjärvi for their kind help during the proteomics related work. I also thank all my Indian friends who provided me social life outside of working place.

I express my gratitude to our collaborators who provided valuable samples for our research. In this regard, I thank Prof. Antonio Vidal-Puig, Prof. Marja-Riitta Taskinen, Prof. Matti Jauhiainen and Prof. Hannele Yki-Järvinen. I would not think of my PhD without their generous help of biological samples. I thank all my large list of coauthors for their kind co-operation in the publications.

Finally, I would like to take this opportunity to thank my family members. I like to thank my parents for their encouragement and blessings. I thank my wife, Krishnaveni Kannekanti, for her unconditional love and affection. I wish to express my warm greetings to my son, Bhuvan, for bringing new wave of happiness in my family.

List of publications

I L. Yetukuri, M. Katajamaa, G. Medina-Gomez, T. Seppänen-Laakso, A. Vidal-Puig, and M. Orešič (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis, *BMC Syst. Biol.* 1: 12.

II L. Yetukuri, J. Tikka, J. Hollmén, and M. Orešič (2010) Functional prediction of unidentified lipids using supervised classifiers, *Metabolomics 6: 18–26.*

III L. Yetukuri, S. Söderlund, A. Koivuniemi, T. Seppänen-Laakso, P. S. Niemelä, M. Hyvönen, M.-R. Taskinen, I. Vattulainen, M. Jauhiainen, and M. Orešič (2010) Composition and lipid spatial distribution of High Density Lipoprotein particles in subjects with low and high HDL-cholesterol, *J. Lipid Res*. In press.

IV G. Medina-Gomez, L. Yetukuri, V. Velagapudi, M. Campbell, M. Blount, M. Jimenez-Linan, M. Ros, M. Orešič, and A. Vidal-Puig (2009) Adaptation and failure of pancreatic beta cells in murine models with different degrees of metabolic syndrome, *Dis. Model. Mech.* 2: 582–92.

V G. Medina-Gomez, S. L. Gray, L. Yetukuri, K. Shimomura, S. Virtue, M. Campbell, R. K. Curtis, M. Jimenez-Linan, M. Blount, G. S. H. Yeo, M. Lopez, T. Seppänen-Laakso, F. M. Ashcroft, M. Orešič, and A. Vidal-Puig (2007) PPARgamma 2 prevents lipotoxicity by controlling adipose tissue expandability and peripheral lipid metabolism, *PLoS Genet.* 3: e64.

VI A. Kotronen, V. R. Velagapudi, L. Yetukuri, J. Westerbacka, R. Bergholm, K. Ekroos, J. Makkonen, M.-R. Taskinen, M. Orešič, and H. Yki-Järvinen (2009) Serum saturated fatty acids containing triacylglycerols are better markers of insulin resistance than total serum triacylglycerol concentrations, *Diabetologia* 52: 684–690.

These articles are referred to as Publication I–VI in the text.

Author's contribution

I In Publication **I**, the author of this thesis involved in the research plan together with Matej Orešič to build the basic informatics work for lipidomics. The author constructed lipid database computationally, developed informatics methodology for lipid pathways, performed data analyses and drafted the manuscript. Mikko Katajamaa developed method for processing of UPLC/MS lipidomics data. Gema Medina-Gomez performed the experiments with ob/ob and WT animals. Tuulikki Seppänen-Laakso performed the lipidomics analysis. Antonio Vidal-Puig coordinated the *in vivo* studies and contributed to manuscript drafting. Matej Orešič initiated the study, performed data analyses and contributed to draft the manuscript.

II In Publication **II**, the author planned the study together with Matej Orešič and Jaakko Hollmén. The author performed programming tasks, carried out computational work and interpreted the results. The author drafted the manuscript. Jarkko Tikka participated in study design. Matej Orešič initiated study, helped in the study progress and provided valuable comments on the manuscript. Jaakko Hollmén supervised computational tasks and provided expert comments on the manuscript.

III In Publication **III**, the author carried out data preprocessing using MZmine software and identification of lipids. The author performed data analysis related to lipidomics and contributed to the major part of the manuscript drafting. Sanni Söderlund performed biochemical experiments and participated in data analysis. Artturi Koivuniemi performed molecular dynamics simulation studies and contributed to drafting of the manuscript, Tuulikki Seppänen-Laakso performed lipidomics experiments, Perttu S. Niemelä participated in discussions and co-ordinated modeling studies with Artturi Koivuniemi. Marja-Riitta Taskinen, Matti Jauhiainen, and Matej Orešič initiated and supervised the progress of the study. Perttu S. Niemelä, Marja

Hyvönen, Marja-Riitta Taskinen, Ilpo Vattulainen, Matti Jauhiainen, and Matej Orešič participated in the drafting and discussions of the manuscript.

IV In Publication **IV**, the author preprocessed the lipidomic data using MZmine software for all the tissues, and identified lipids. The author performed data analysis related to lipidomics and contributed to the manuscript drafting in statistical methods. Gema Medina-Gomez performed biochemical experiments in mouse, participated in discussion and the study design and drafted the manuscript. Vidya Velagapudi performed lipidomics experiments, Mark Campbell, Margaret Blount, Mercedes Jimenez-Linan and Manuel Ros participated in the biochemical experiments. Matej Orešič supervised the lipidomics studies and Antonio Vidal-Puig supervised the in vivo studies and drafted the manuscript.

V In Publication **V**, the author preprocessed the data with MZmine software and contributed to lipid identification and data analysis with Matej Orešič. Gema Medina-Gomez, Sarah L. Gray, Frances M. Ashcrof, and Antonio Vidal-Puig conceived and designed the experiments. Gema Medina-Gomez, Sarah L. Gray, Frances M. Ashcrof, Kenju Shimomura, Sam Virtue, Mark Campbell and Miguel Lopez performed the experiments. Gema Medina-Gomez, Sarah L. Gray, Kenju Shimomura, Sam Virtue, Mark Campbell, R. Keira Curtis, Mercedes Jimenez-Linan, Tuulikki Seppänen-Laakso, and Matej Orešič analysed the data. Gema Medina-Gomez, Kenju Shimomura, Sam Virtue, Mark Campbell, R. Keira Curtis, Margaret Blount, and Giles S. H. Yeo contributed reagents, materials and analysis tools. Gema Medina-Gomez and Antonio Vidal-Puig wrote the paper.

VI In the Publication **VI**, the author preprocessed the lipidomic data for all lipoprotein fractions. The author compiled tandem mass spectrometry (MS/MS)-based spectral library of all lipoproteins fractions combining both positive and negative ion mode. Anna Kotronen was involved in the data analysis and drafted the manuscript. Vidya R. Velagapudi performed lipidomics experiments and participated in the data analysis. Jukka Westerbacka, Robert Bergholm, Janne Makkonen were involved in the experimental design and biochemical experiments. Kim Ekroos was involved in the discussions on the manuscript. Marja-Riitta Taskinen, Matej Orešič and Hannele Yki-Järvinen were involved research design, discussions and drafted the manuscript. Matej Orešič participated in the data analysis.

Contents

Publications II and VI are not included in the PDF version. Please order the printed version to get the complete publication (http://www.vtt.fi/publications/index.jsp).

List of abbreviations

ANOVA	Analysis of variance
APCI	Atmospheric pressure chemical ionization
Cer	Ceramide
ChoE	Cholesteryl ester
ESI	Electrospray ionization
FDR	False discovery rate
HCA	Hierarchical clustering analysis
HDL	High density lipoproteins
HDL-C	High density lipoprotein-cholesterol
HPLC	High-performance liquid chromatography
KEGG	Kyoto encyclopedia of genes and genomes
k-NN	k-nearest neighbours
LC/MS	Liquid chromatography / mass spectrometry
lysoPC	Lysophosphatidylcholine
m/z	Mass-to-charge ratio

PC	Phosphatidylcholine
PC(e)	Ether linked phosphatidylcholine
PCA	Principal component analysis
POKO	Ablation of PPARg2 in Ob/Ob background (PPARg2(-/-) Lep(ob)/Lep(ob))
PE	Phosphatidylethanolamine
PE(p)	Ethanolamine plasmalogen
PLS/DA	Partial least squares and discriminant analysis
Q-TOF	Quadrapule time of flight
RT	Retention time
SM	Sphingomyelin
SMILES	Simplified molecular input line entry system
SVM	Support vector machines
TG	Triacylglycerol
TLC	Thin layer chromatography
UPLC/MS	Ultra performance liquid chromatography coupled to mass spectrometry

1. INTRODUCTION

Lipids are an important class of compounds that have a wide variety of key cellular functions including compartmentalisation, energy storage, cell-signalling, protein trafficking and membrane anchoring (Orešič et al. 2008, van Meer 2005, Vance and Vance 2008). Emerging evidence suggests that abnormalities in these functions are either directly or indirectly linked to the pathogenesis of various diseases (Wenk 2005) including obesity (Shi and Burn 2004), Alzheimer's diseas (Cutler et al. 2004), cancer (Menendez and Lupu 2007) and atherosclerosis (Lusis 2000). The lipids are generally hydrophobic in nature and are soluble in organic solvents. They are defined as hydrophobic or amphipathic small molecules that may originate entirely or in part by carbanion based condensation of thioesters, and/or by carbocation based condensation of isoprene units (Fahy et al. 2005). These lipids vary widely ranging from simple fatty acids to complex glycolipids. Systematic cataloguing of all these lipid classes and their nomenclature is required for databases and bioinformatics needs. According to new classification system, lipids are classified into eight main categories: 1) Fatty acids, 2) Glycerolipids, 3) Glycerophospholipids, 4) Sphingolipids, 5) Sterols, 6) Prenol lipids, 7) Saccharolipids, and 8) Polyketides (Fahy et al. 2009).

The structural diversity of lipids stems mainly from various combinations of fatty acid chain lengths and possible head groups (*e.g.*, for glycerophospholipids) that are linked to glycerol backbone (Figure 1.1). The presence of ethylene-interrupted or less common methylene-spaced double bonds in fatty acid moeities introduces an additional diversity in the lipidome. The presence of various types of glycerol-alkyl chain linkages such as ester, ether and vinyl ether bonds provide further diversity. Ether and vinyl ether bonds are more common in sn-1 position and are found mainly in phosphatidylcholine (PC) and phosphatidylethanolamine (PE) type of lipid classes (Snyder 1999). This diversity and abundance of these lipid species varies from tissue to tissue. For

instance, ether-linked (plasmanyl) and vinyl-ether linked (plasmenyl) phospholipids are more abundant in tissues such as heart, kidney, and central nervous system (Druilhet et al. 1975, Panganamala et al. 1971). Similar diversity can be seen in most other classes of lipids and the theoretical number of possible lipids, when conservatively estimated, exceeds ~180,000 lipids (Yetukuri et al. 2008).

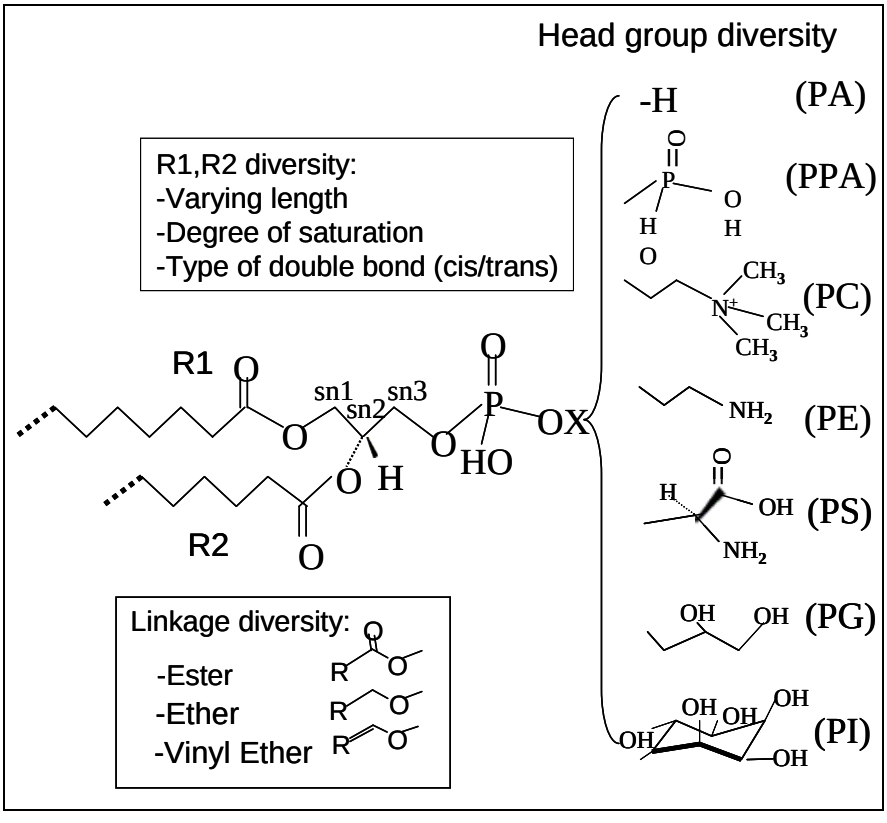


Figure 1.1. Glycerophospholipid structural diversity: R1 and R2 are fatty acids at position 1 and position 2, respectively. X represents head group moiety at position 3. Head groups legend: PA = Phosphate, PPA = Pyrophosphate, PE = Phosphoethanolamine, PC = Phosphocholine, PS = Phosphoserine, PG = Phosphoglycerol and PI = Phosphoinositol.

Recent burgeoning interest in lipid research illustrates the critical physiological importance of lipids. Moreover, recently emerged new lipidomics consortia such as the US-based LIPID MAPS (www.lipidmaps.org) and its affiliated sphinGOMAP (http://sphingolab.biology.gatech.edu/), and similar community-wide efforts in Japan (www.lipidbank.jp) and Europe (www.lipidomics.net) emphasis the growing need for indepth lipidomic research. Other related initiatives include The

Alliance for Cell Signaling (www.signaling-gateway.org/) and Lipid Profiles (www.lipidprofiles.com). As a result, lipid research is now beginning to appear as *omics* level science with the emerging precipitous developments. Lipidomics is defined as "the comprehensive understanding of the influence of all lipids on a biological system with respect to cell signaling, membrane architecture, transcriptional and translational modulation, cell-cell and cell-protein interactions, and response to environmental changes over time" (Watson 2006).

Any successful lipid analytical method should be a flawless combination of extraction, separation, detection and easy processing of the data. However, no single methodology or technique is yet in widespread use to screen all lipids. Recent advances in mass spectrometry-based studies have revolutionized lipidomics research at molecular level (Griffiths 2003, Navas-Iglesias et al. 2009, Ståhlman et al. 2009). Especially, the soft ionization MS-based analytical methods have gained popularity for their specificity, sensitivity and fast detection of different lipids from various biological matrices. The development of electrospray ionization techniques (Fernandis and Wenk 2009, Pulfer and Murphy 2003) have led to the study of lipids to a greater detail mainly in two ways: separation of lipid extracts using liquid chromatogram (LC) followed by on-line MS monitoring (Ogiso et al. 2008, Wang et al. 2005) and direct infusion of lipid extracts into a mass spectrometer where lipids are selectively detectable using techniques such as specific precursor ion scans (PIS) and neutral loss scans (NLS) (Ejsing et al. 2006, Ekroos et al. 2002, Han and Gross 2005a). As analytical technologies are becoming more mature and moving towards achieving the true quantitative or at least semi-quantitative characterization of molecular lipid species and lipid classes, it is becoming possible to study lipid pathways at the molecular level. It is evident that this knowledge will significantly advance our knowledge on thd roles of lipids in the context of cellular and organismal physiology.

Lipidomics, a branch of metabolomics, is the end point of *omics* cascade and bears the direct link to several disease phenotypes and therefore has recently become the target of post-genomics research. The advent of modern MS technologies has facilitated in the analysis of hundreds of lipid molecules from a given biological matrix. As a result, large scale data sets are being generated from the modern analytical methods, presenting new challenges for lipid informatics. The informatics approaches in co-ordination with improved analytical methods should resolve and identify individual lipids, unravel minute

systemic changes at molecular level and integrate the complex lipidome data with other enitities such as genes, proteins and other metabolites.

1.1 Aims of the thesis

The main aim of this thesis was to develop bioinformatics approaches for the non-targeted lipidomics data coming from UPLC/MS. The specific goals of my thesis (Figure 1.2) were the following:

- Development of computational spectral libraries and their customization for the UPLC/MS platform to facilitate lipid identification (Publication **I**).
- Building of tissue-specific experimental spectral libraries combining both positive and negative ion mode tandem mass spectrometry data (e.g., Publication **VI**).
- Functional class label annotation of unidentified lipids (Publication **II**).
- Mapping of lipidomics data in metabolic pathways (Publication **I**).
- Applications of developed informatics approaches followed by exploratory analysis (Publication **III, IV, V)**.

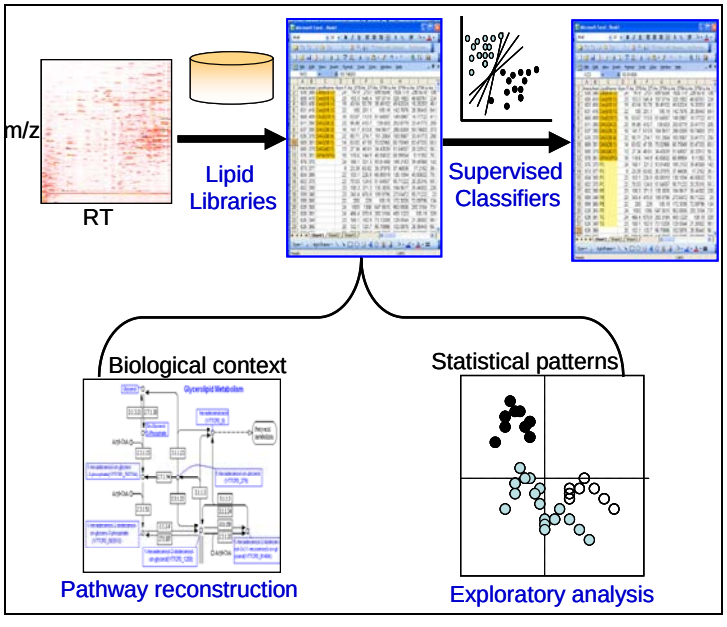


Figure 1.2. A schematic diagram showing the overview of this thesis in lipid bioinformatics approaches.

2. LITERATURE REVIEW

The lipid bioinformatics field is interdisciplinary in nature and topics related to this thesis are reviewed in this chapter. This chapter reviews the use of liquid chromatography and mass spectrometery in the lipid analysis, and advances in dataprocessing tools, databases and biochemical pathways.

2.1 Lipid analysis using LC/MS techniques

Biological matrix usually contains several lipids some of which have the same effective carbon number (ECN, a measure of non-polar characteristic of fatty acid chain or molecule) and therefore either co-elute partially or entirely from the liquid chromatography (LC) column. Thus simple LC alone is not enough to separate them. Favourably, mass spectrometry (MS) can act as second dimensional separation by discriminating compounds based on their mass fragments. Successful integration of LC with MS has played a significant role in the characterization of multiple compounds in a single sample.

2.1.1 Separation of lipids using liquid chromatography

Several studies have successfully demonstrated the usefulness of liquid chromatographic techniques in the analysis of complex lipid mixture (Bijlsma et al. 2005, Hermansson et al. 2005, Houjou et al. 2005). LC-based methods make use of properties such as differential solubility and partition between mobile and stationary phases. One of the earlier developments of LC is thin layer chromatography (TLC) which has been successfully used for the analysis of lipids (Bennet and Heftmann 1962, Michalec et al. 1962). TLC is developed from paper chromatography and comprises thin layer of stationary phase such as silica or cellulose on a flat support. Various combinations of aqueous stationary and organic mobile phases facilitate separation of several classes of lipids and

thus serve as a rapid screening tool prior to the development of more advanced and sensitive methods. Another notable methodology is solid phase extraction (SPE). This technology is useful in separating crude lipid mixtures into several classes including phospholipids, fatty acids, cholesterol esters, acyl glycerols and cholesterol (Kaluzny et al. 1985). High performance liquid chromatography (HPLC) has gained high popularity for its selectivity and efficiency. The separation in HPLC can be achieved using either normal-phase or reverse-phase columns. Normal-phase HPLC facilitates separation of lipids based on their polar head group (Lesnefsky et al. 2000) without major effect by the fatty acid substituents. In this case, each class co-elutes as single chromatographic peak with small retention time differences across the individual molecules in the class. In case of reverse phase column, the separation of lipids is based on their polarity, degree of alkyl chain saturation, and chain length. The reverse phase column thus enables separation of lipids with different fatty acid compositions (McHowat et al. 1997).

2.1.2 MS-based lipid analysis

Mass spectrometer, which has profound influence on modern analytical chemistry, can measure the mass of charge carrying molecules. The instrument measures both mass-to-charge ratio of molecule and its intensity and thus serves as an invaluable tool in structural elucidation (using tandem mass spectrometry) as well as quantification.

Electron ionization (EI) is useful in gas chromatography where the eluting gaseous molecules are bombarded with a beam of high-energy electrons and thus generating a specific fragmentation patterns. Other techniques such as chemical ionization (CI) use a reagent gas to ionize molecules which do not give molecular ion in EI and produce less fragmentation pattern than EI analysis. This ionization method was initially developed for gas chromatography and has been used as atmospheric pressure chemical ionization (APCI) for liquid samples. Laser-based soft ionisation technique called matrix-assisted laser desorption/ ionization (MALDI) is used for the analysis of large molecules and also can also be used for the analysis of lipids. The sample is mixed with a chemical matrix and then applied to sample holder as small spot. The matrix absorbs the energy from laser beam and thereby analyte receives the energy and results in ionization of molecules. ESI-based MS provides the most promising soft ionization technique and now has become the system of choice for both characterization

and quantification of lipids. This technology has been successfully applied in numerous lipidomic studies (Griffiths 2003, Han and Gross 2005a, Pulfer and Murphy 2003). This technique does not require any derivatisation and can directly be applied on solutions. The technique is moreover characterized by high-sensitivity with reasonable experimental complexity and high reproducibility. The ESI-based MS methodology has become the preferred method for the analysis of phospholipids (Brugger et al. 1997, Pulfer and Murphy 2003) and sphingolipids (Haynes et al. 2009). The methodology is also adopted for the analysis of non-polar lipids such as acylglyerols (Han and Gross 2001).

2.2 UPLC/MS-based non-targeted lipidomic analysis

Advances in MS-based analytical methods, in combination with ESI-based soft ionization technique, have spurred lipidomic research in recent years**.** The HPLC-based methods, however, are time-consuming and often present a bottleneck in the routine analysis. Shorter LC analysis times allow increasing sample throughput as well as lab productivity.

Recent advances in novel mass analysers such as orthogonal-accelerated time of flight (oa-TOF) have led to the development of mass spectrometer. A commonly used mass spectrometer with oa-TOF analyser contains a quadruple-time of flight (Q-TOF) configuration (Figure 2.1) and resembles as triple quadrupole MS in which third quadrupole has been replaced with oa-TOF. The TOF analyser permits full scan acquisitions with high resolution and mass accuracy. This powerful combination of single quadrupole and high performance of TOF enables both MS and tandem mass spectrometry (MS/MS) experiments. In the MS mode, the first quadrupole and collision cell merely guide the ions, while TOF seperates all the ions that are orthogonally accelerated. In MS/MS mode, the filter capabilities of quadrupole are exploited to transmit and to select only precursor ions. These precursor ions are accelerated due to potential difference before they get fragmented in the collision cell induced by collision with neutral gases such as argon or nitrogen. The resulting ions are analysed in TOF tube with high mass accuracy.

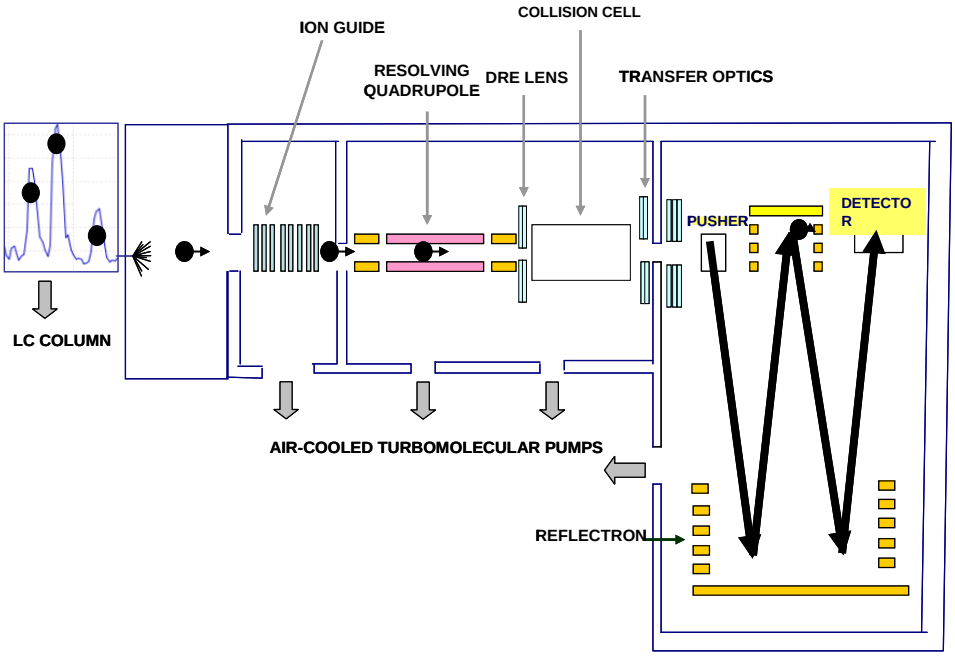


Figure 2.1. Schematic view showing the configuration of quadrupole time of flight (Q-TOF) mass analyzer.

Recent advances in reverse phase chromatographic columns (e.g., Bridged Ethane Hybrid (BEH) C18 columns packed with 1.7 µm particles) and mobile phase systems allowed HPLC system to operate at much higher back-pressures. The new column, called ultra performance liquid chromatography (UPLC), offers significant advantages in resolution, speed, and sensitivity as compared to conventional HPLC analysis (Apollonio et al. 2006, Churchwell et al. 2005, Leandro et al. 2006, Wilson et al. 2005). With improved speed and sensitivity, UPLC/MS platform thus provide a greater advantage in high-throughput sample analysis. Here, UPLC is coupled to high resolution quadrapole time of flight (Q-TOF) mass analyzer which enables accurate mass measurements of precursor and fragment ions. Non-targeted lipidomics approaches have the advantage of detecting greater number of lipid components and possibility for detecting novel compounds. It is also advantageous in picking up global changes thereby serving as a guide for designing targeted approaches.

2.3 Data processing tools for lipid analysis

The modern analytical instruments allowed lipidomic studies, generating huge amounts of data to process. The data processing step is critical, labor-intensive and becomes the rate-limiting step in metabolomics studies. This impetus has led to the development of many data processing tools both in open source and commercial environment (Katajamaa and Orešič 2007). Common data processing steps in lipidomics include peak detection, lipid identification, isotope correction, response correction and quantification. The typical data processing starts by inputting data in some vendor-specific format, which often becomes practical difficulty in applying same software for different data formats coming from different vendors. Recently introduced tools such as Lipid Qualitative/Quantitative Analysis (LipidQA) software platform (Song et al. 2007) can identify and quantitate the complex lipids in biological mixtures. The software can process the data coming from TSQ-7000 triple stage quadrupole and LTQ linear ion trap mass spectrometers from Thermo-Finnigan and Q-TOF hybrid quadrupole/time-of-flight instrument from Waters-Micromass. The algorithm can also handle data-dependent manner lipid identification based on MS/MS spectra of glycerophospholipid species. Fatty Acid Analysis Tool (FAAT) tool (Leavell and Leary 2006) is developed for the analysis of data coming from Fourier transform mass spectrometry and the tool is demonstrated using mycobacteria species data. The main functionalities of the software include identification of overlapping saturated and unsaturated lipids, assignment of known ions from a user-defined library and handling of isotopic shifts from stable isotope labeling experiments. The software tools such as SECD and LIMSA (Hermansson et al. 2005) are useful for the display of chromatograms and performing several data processing steps including peak picking, integration, isotope correction and internal standards-based quantification. Other tools such as Lipid Profiler (Ejsing et al. 2006) and LipidInspector (Schwudke et al. 2005) are compatible for the data acquisition with Applied Biosystems hybrid quadrupole/time-of-flight instruments that can perform multiple precursor ion scans in a single experiment. The novel lipid-mediator informatics developments such as cognoscitive-contrast-angle algorithm and database (COCAD) (Lu et al. 2006) enhance correct identifications of lipid-mediators by matching either known standard MS/MS spectra with chromatograms and UV spectra or virtual liquid chromatography-ultraviolet-tandem mass spectra.

MZmine software (Katajamaa et al. 2006, Katajamaa and Orešič 2005) is an open source Java-based data processing tool for LC/MS-based metabolomics experiments, with a particular focus on differential analysis of lipidomics data from UPLC/MS experimental setup. The software implements several key methods for data processing stage including spectral filtering, peak picking, deisotoping, alignment of samples and quantification. Moreover, recursive peak search algorithm and peak picking methods facilitate the improvement of already aligned data. Several data visualization options are available for the display of spectral data across multiple samples. MZmine2 (http://mzmine.sourceforge.net/) presents new features and improved modularity for better expandability. Some of new features include processing of high-resolution instrumental data, better visualisation (3D visualisation) and a new implementation of 2D visualiser. It allows storage of parameters for defining sample properties. The software supports importing of several data formats such as netCDF, Thermo RAW, mzML and mzXML and also the stored project-specific parameters defining the samples.

2.4 Lipid databases

Development of databases (Table 2.1) and related bioinformatics tools has become an essential part of functional genomics studies. Over the recent years, empowered by high-throughput technologies for *omics* fields, creation of databases devoted to certain entities such as lipids was undertaken. Consequently, lipid-centric databases were developed that enabled researchers to comfortably analyse expression patterns of lipid related genes and gene products. For example, a database of genomics of lipid-associated disorders, called GOLD, offers annotated pathways, curated data sets and possibility to study experimental data in the context of biological pathways (Hackl et al. 2004). The LIPID MAPS Proteome Database (LMPD) (Cotter et al. 2006) is a database of lipid-associated protein sequences and annotations. Presently, the database mainly comprises human and mouse related proteins of lipid metabolism. The protein database is enhanced with annotations from external databases.

Several lipid databases such as LIPID BANK (http://www.lipidbank.jp/), LIPIDAT (Caffrey and Hogan 1992) and LMSD (Sud et al. 2007) are publicly available offering wide-range of information including lipid structures. Notably, LMSD offers systematic structures of lipids as well as other related information according to the classification scheme recommended by LIPID MAPS consortium. Users can retrieve the data from LMSD using text- or structure-based queries.

Web tools such as LIPID MAPS online tools (Fahy et al. 2007) permit user friendly-queries from underlying databases. Lipid library (http://www.lipidlibrary.co.uk/) and CyberLipid (http://www.cyberlipid.org/) offer rich source of lipid information.

The proposed classification, nomenclature, and chemical representation system introduced by the the LIPID MAPS consortium has become standard reference for the construction of lipid databases and management of lipidomics data. One of the main goals of LIPID MAPS initiative includes building of lipid databases and related bioinformatics approaches. These databases need to be complemented with annotation and curation of lipid structures. The LIPID BANK aims at curation of lipid structures as well as annotation of the related literature. Other databases like PubChem (http://pubchem.ncbi.nlm.nih.gov/) offer huge repository of chemical compounds including lipids. The database also provides various physical/chemical properties and cross links to other databases. All these databases are handy in the analysis of lipids and may serve as tools for validation of results. Notably, LMSD also provide various tools for mass spectrometry data. However, given the diversity of lipids across different organisms, tissues, and cell types, it is unlikely any one database can become a reference for mass spectrometry data. Mainly for LC/MS-based analysis, the development of in-house databases that are customized for the instrumental settings are needed

2.5 Lipid pathway resources

The existing databases offer rich source of information on lipid pathways. Databases such as Kyoto encyclopedia of genes and genomes (KEGG) database (Kanehisa and Goto 2000, Kanehisa et al. 2004) serve as a valuable resource for analyzing cells, not only at genomic level but also for metabolic networks in different organisms. The database offers information on most metabolic pathways including lipid pathways. Additionally, KEGG provides generic pathways (i.e., species-independent pathways) as reference pathways for the reconstruction of context- or organism-specific pathways. Moreover, the KEGG Brite (http://www.genome.jp/kegg/brite.html) maintains a collection of hierarchical classifications of lipid species whose reactions and pathways can be viewed. Other more annotated databases such as MetaCyc (Krieger et al. 2004) and EcoCyc (Keseler et al. 2005) serve as a good starting point for the study of lipids. SphinGOMAP (http://sphingolab.biology.gatech.edu/) offers comprehensive pathway mapping of about 450 distinct sphingolipids and glycosphingolipids species. LIPID

MAPS biopathways workbench (http://www.biopathwaysworkbench.org/) provides a graphic tool that facilitates to display, edit and analyse biochemical pathways of lipids**.** In a recent study (Gupta et al. 2009), kinetic model was built from the lipidomics flux analysis using integrated network of eicosanoids metabolism and signaling pathways. The developed integrated model is based on the KEGG pathways and literature knowledge. Rate constants in the kinetic model are estimated and tuned using generalised constrained non-linear optimization. These quantitative models are quite useful for perturbation studies to gain mechanistic understanding about the underlying phenotype. These tools and databases allow reconstruction of integrated pathway models and thus open new avenues for building system level quantitative models.

Table 2.1. List of publicly available lipid resources and their characteristics.

Lipid database/source	**Description**
LIPID MAPS (www.lipidmaps.org)	Provides guidelines for classification system for lipids, hosts databases of lipids and lipid-associated protein data and develop tools for identifying lipids.
Lipidomics Expertise Platform (http://www.lipidomics.net/)	European level Initiative for lipid research. Provides databases for the registered users and serves as a source for establishing European level networks, industrial relations and exchange of standard materials.
CyberLipids (www.cyberlipid.org)	Offers huge collection of updated scientific knowledge on all aspects of lipids. Also facilitates establishing relationships among students, teachers, scientists and technicians and present and provides updated bibliography devoted to lipid biology.
LIPIDAT (ww.lipidat.chemystry.ohio-state.edu/home.stm)	Presents thermodynamic information on lipids including lipid phase transition temperatures and enthalpy changes for synthetic and biologically relevant complex polar lipids.
LIPID BANK (www.lipidbank.jp)	Provides chemical structures of lipids with names, chemical and physical properties, biological activities and metabolism. In addition, spectral information from various instruments such as ultraviolet, infrared spectrometry, nuclear magnetic resonance, mass spectrometry, liquid chromatography, and thin-layer chromatography can also be obtained.
KEGG lipids (http://www.genome.jp/kegg-in/get_htext?br08002.kegg)	Provides lipid pathway maps as well as associated information such as the name, formula, mass, structure, biochemical reactions and external links to other public databases
THE LIPID LIBRARY (http://www.lipidlibrary.co.uk/)	Portal for the study of many classes of lipids and their analysis both in mass spectrometry and chromatography.
sphinGOMAP (http://www.sphingomap.org/)	Offers a database for biochemical mapping of sphingo- and glycosphingo-lipids.

3. METHODS

3.1 Lipid analysis using UPLC/MS platform

3.1.1 Lipid standards and chemicals

Sample was extracted with chloroform / methanol (2:1, 100 µl) after addition of an aliquot (20 µl) containing internal standard mixture. After homogenization and vortexing, the sample was centrifuged (10000 rpm, 3 min) and the lower phase was collected. The lower lipid phase was mixed with another aliquot of labeled standard mixture. The labeled mixture containing 3 standards was added before analysis in order to control the extraction process (Pietiläinen et al. 2007). The labeled standards were PC(16:0/0:0-d3), PC(16:0/16:0-d6) and TG(16:0/16:0/16:0-^{13}C3) and were obtained from Larodan Fine Chemicals (Malmo, Sweden). The internal standards mixture comprised MG(17:0/0:0/0:0)[rac], DG(17:0/17:0/0:0)[rac] and TG(17:0/17:0/17:0) from Larodan Fine Chemicals (Malmo, Sweden) and PC(17:0/0:0), PC(17:0/17:0), PE(17:0/17:0), PG(17:0/17:0)[rac], Cer(d18:1/17:0), PS(17:0/17:0), PA(17:0/17:0) and D-erythro-Sphingosine-1-Phosphate (C17 Base) from Avanti Polar Lipids (Alabaster, AL).

3.1.2 Mass spectrometry conditions

The extracted lipid samples were analysed on quadrupole time-of-flight (Q-Tof Premier) mass spectrometer combined with an Acquity ultra performance liquid chormatogram (UPLC) (Waters Inc., Milford, MA). The column was an Acquity UPLC™ BEH C18 10×50 mm with particle size of 1.7 µm and was maintained at 50°C. The composition of the binary solvent system at the flow rate of 0.200 ml/min was A: water (1% 1 M NH_4Ac, 0.1% HCOOH) and B: LC/MS grade (Rathburn) acetonitrile/isopropanol (5 2, 1% 1 M NH_4Ac, 0.1% HCOOH). The

initial gradient of the solvent composition was 65% A/35% B. The gradient reached 100% B in 6 min and maintained there for the next 7 min. The total run time was 18 min which included a 5 min re-equilibration step. The sample organizer was set at 10°C.

The lipid profiling was carried out on Waters Q-Tof Premier mass spectrometer using electron spray ionization in either positive or negative ion mode. The data were collected usually in the mass range of m/z 300–1200 with scan duration of 0.2 sec. The temperature of source was maintained at 120°C and nitrogen was used as desolvation gas (800 L/h) at 250°C. The voltages of the sampling cone and the capillary were 39 V and 3.2 kV, respectively. Reserpine (50 µg/L) was used as the lock spray reference compound (5 µl/min; 10 sec scan frequency). The samples were analysed in a randomized order.

3.1.3 Data processing with MZMine

Lipid data from UPLC/MS experiments were first converted from raw data to netCDF file format using DataBridge utility of MassLynx 4.1 software (Waters, Inc.). The netCDF files were preprocessed using an in-house developed MZmine software version 0.60 (Katajamaa et al. 2006). Main functionalities of the software include peak picking, chromatographic alignment, spectral filtering, peak area calculations, visualisation (i.e. peak maps, curvilinear distance analysis and Sammon's mapping), gapfilling, normalisation, and data export. De-isotoping step was performed using in-house developed MATLAB scripts. Lipids were identified using an internal spectral library or alternatively with tandem mass spectrometry. Calibration (normalisation) was done based on multiple internal standards and was performed as follows: All monoacyl lipids (monoacylglycerols and lysophospholipids) were normalized with LysoPC(17:0/0:0), all diacyl lipids except phosphatidylethanolamines and ethanolamine plasmalogens were normalized with PC(17:0/17:0), the phosphatidylethanolamines and ethanolamine plasmalogens were normalized with PE(17:0/17:0), and the triacylglycerols and cholesterol esters with TG(17:0/17:0/17:0). Calibration of unidentified lipids, similar to method described earlier (Bijlsma et al. 2005), was done using three internal standards as follows: lysoPC(17:0/0:0) was used to normalize the peaks eluting with retention time (RT) < 300s, PC(17:0/17:0) for 300s < RT< 410s, and TG(17:0/17:0/17:0) for RT > 410s.

3.2 Lipid database construction

3.2.1 Lipid scaffold generation

Lipid database was constructed computationally using Simplified Molecular Input Line Entry System (SMILES) approach (Publication **I**) which is a widely used chemical notation to represent a chemical structure in terms of atoms and bonds governed by set of syntax rules (Weininger 1988). The database accommodated main classes of lipids including fatty acids, phospholipids, glycerolipids, cholesterol esters, and sphingolipids. The scaffolds of theoretically possible lipids were computed based on known lipid building blocks such as polar head groups and fatty acids in order to facilitate identification of lipids. In order to construct a particular lipid class (e.g., glycerophospholipid), SMILES template was constructed to represent the structure of the class. Possible lipids in the class along with their names were generated compuationally by varying alkyl moiety, nature of linkage and head group. Each lipid in the database was annotated with systematic name, SMILES, molecular formula and exact average mass and monoisotopic mass. A score value was assigned to each compound based on natural abundance of fatty acid(s). Common factors considered while assigning the score were natural abundance of the fatty acid, and odd or even number of carbon atoms present in the fatty acid chain. This kind of scoring scheme facilitated quick search of possibly abundant compounds.

3.2.2 Lipid nomenclature

Lipids were named according to recent nomenclature system for lipids introduced by LIPID MAPS consortium (Fahy et al. 2009). For example, lysophosphatidylcholine with 17:0 fatty acid chain at sn-1 position was named as 1-heptadecanoyl-sn-glycero-3-phosphocholine (short name: PC(17:0/0:0)). If the exact fatty acid composition was not determined, total number of carbons and double bonds was indicated. For example, a phosphatidylcholine species PC(18:0/20:4) is represented as PC(38:4). However, PC(38:4) may correspond to isobaric (e.g., PC(22:4/16:0)) or isomeric species (e.g., PC (20:4/18:0)).

3.2.3 Database design

Information on each lipid entry was converted into XML document and the resulting documents were loaded to the database using mass-loading tool of Tamino server. The database is a native XML database implemented in Tamino XML Server (Software AG). Each entry in the database was annotated with an internal identifier, scoring information, class, canonical SMILES, molecular formula, molecular weight and isotopic distribution. All the relevant scripts were implemented in the Perl language. When implementing the database, we used XMLSPY software (Altova, Inc.) and Tamino Schema Editor Software (Software AG) for the construction and validation of logical and physical schemas, respectively.

3.3 Lipid pathways

3.3.1 Lipid pathways and extensions

System level characterization by integrating genes, proteins, lipids and other molecules provide better insights in an organism (Joyce and Palsson 2006). The existing databases provide different levels of information. The databases such as KEGG database (Kanehisa and Goto 2000, Kanehisa et al. 2004) serve as a valuable resource for analyzing cells not only at genomic level but also for metabolic networks in different organisms. The KEGG is a database of biological systems that integrates genomic, chemical and network information (http://www.genome.jp/kegg/). The KEGG hosts a collection of manually drawn pathway maps based on the current knowledge on the molecular interaction and reaction networks. The KEGG PATHWAY database offers information on most metabolic pathways including lipid pathways include fatty acid biosynthesis, fatty acid elongation in mitochondria, fatty acid metabolism, synthesis and degradation of ketone bodies, steroid biosynthesis, primary and secondary bile acid biosynthesis, C21-Steroid hormone metabolism, androgen and estrogen metabolism, glycerolipid metabolism, glycerophospholipid metabolism, ether lipid metabolism, sphingolipid metabolism, arachidonic acid metabolism, linoleic acid metabolism, alpha-linolenic acid metabolism and biosynthesis of unsaturated fatty acids. Additionally, KEGG also provides generic pathways (i.e., species-independent pathways) to serve as reference pathways for the reconstruction of context- or organism-specific pathways.

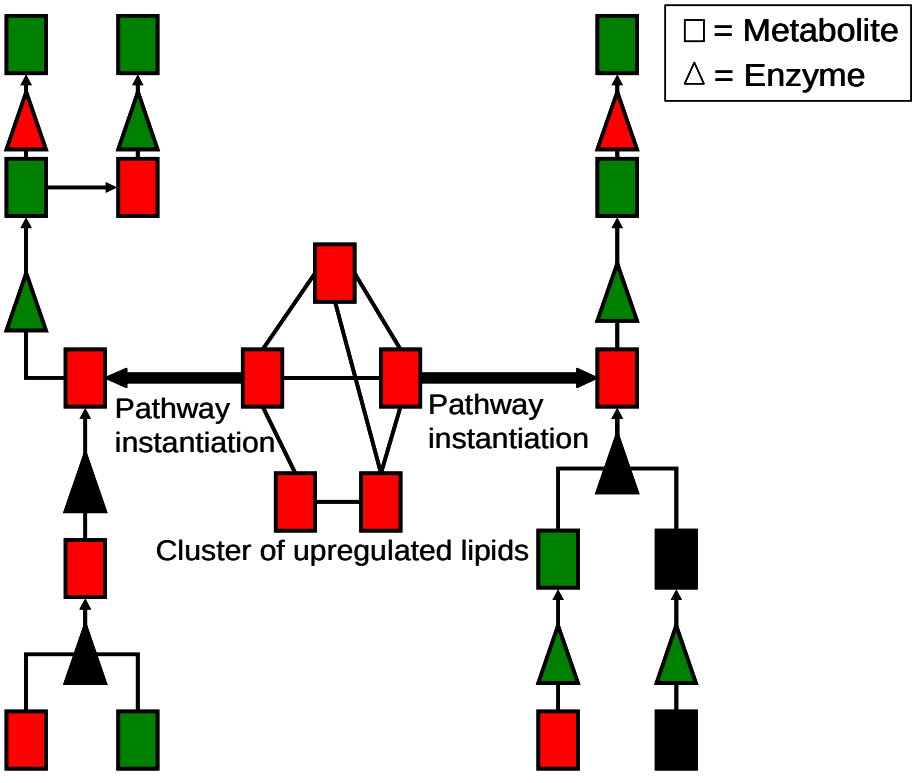


Figure 3.1. An illustrative example of lipid pathway reconstruction methodology via pathway instantiation. Interesting lipids from co-regulation network are linked to molecular instance pathways. Red/green colour coding refers to up/down regulation of enzymes and lipids.

Modern MS-based analytical techniques provide lipid species level information, whereas present lipid pathway information is mostly limited to the generic lipid class level. At the generic level, most lipid entries on pathways may contain one or more fatty acids and/or head groups. Due to enormous diversity in fatty acids and head groups, large number of specific lipids that are measurable now can be substitutable for a particular entry on pathway. As a result, lipid pathway reconstruction may easily end up in combinatorial explosion with varying complexity from pathway to pathway. To avoid this combinatorial problem, generic pathway templates are utilized to create molecular instance pathways for molecular species selected based upon multivariate and co-regulation analyses (Figure 3.1). Pathway instantiation is a method of converting generic names of lipids on biochemical lipid pathways to corresponding specific names of interest. Instance pathways allows mapping of lipids observed in mass spectrometric experiments.

3.3.2 Visualization tools for lipid pathways

Managing and organising lipid-related pathways into useful, interactive pathways and networks present a greater challenge for lipid bioinformatics. In-house developed megNET software tool enables retrieval and visualisation of biological relationships across heterogeneous data sources from an integrated database (Gopalacharyulu et al. 2005). Other open source visualization tools such as VANTED (Junker et al. 2006) enable importing and customisation of KEGG lipid-specific pathways.

3.4 Modeling with supervised and unsupervised methods

There are two main categories of mathematical modeling approaches: supervised and unsupervised. Supervised modeling (Mitchell 1997) is an approach that uses pairs of input objects (usually in the form of matrix, **X**) and desired outputs (usually in the form of matrix, **Y**). The output of the function can either be continuous, as in the case of regression, or categorical, as in the case of classification. Information in matrix **Y** is used to guide the construction of the model for **X** and hence the name supervised. On the other hand, unsupervised modeling only utilizes the observed data in X and the model tries to learn the statistical patterns or trends available in X (Duda et al. 2001). Conventional methods for unsupervised learning such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) are generally employed in exploratory analysis.

3.4.1 Preprocessing of multivariate data

Preprocessing of multivariate data is advocated to extract relevant information from a given data matrix. One of the most commonly employed procedures is mean-centering (also called column centering) where the goal is to model the actual variation in the data. In mean-centering, the mean of each measured variable (column mean) is substracted from each value of the respective variables in the data set so that resulting data matrix contains columns with zero mean. Other routinely used preprocessing step, especially in chemometrics, is unit variance (UV) scaling where each variable (column vector in the data matrix) is divided by the respective standard deviation. This scaling alleviates the effect of differences in magnitude of variables i.e., the higher magnitude

variables have the greater influence on the results. Combination of mean-centering and UV-scaling can also be employed (i.e., auto scaling).

3.4.2 PCA

Principal Component Analysis (PCA) is a latent variables-based unsupervised method for exploratory analysis (Hotelling 1933). PCA uncovers simpler patterns from the complex inter-correlated variables. The PCA can also be seen as a dimensionality reduction strategy while retaining as much information as possible. This is achieved by creating new set of variables which are linear combinations of original variables to produce principal components. These principal components are orthogonal to each other and are uncorrelated. The first principal component is in the direction of the greatest variance in the data and subsequent components are constructed orthogonal (independent) to the previous ones in the direction of largest remaining variance. Usually, the first few latent components account for the most of the variation in the data matrix (X [nxm]). The data matrix, X, can be decomposed into two matrices: scores matrix (T) and loadings matrix (P)

$$X = TP^t = \sum_{i=1}^{m} t_i p_i^t \tag{4.1}$$

where t denotes a transpose operation on matrix. The loading matrix contains information on the variables while the scoring matrix contains information about the objects. When the data are projected into a lower dimensional space spanned by few principal components corresponding to maximum variation, the data matrix X can be written as in equation (4.2).

$$X = \sum_{i=1}^{k} t_i p_i^t + E \tag{4.2}$$

where E is unexplained variation in data matrix X and k (k << m) is the number of first principal components.

3.4.3 HCA

General clustering methods are based on the distance between the samples whose observed parameters are co-ordinates in the multi-dimensional space.

Similarity or dissimilarity of the samples is based on whether they are close or not in the high dimensional space. Hierarchical cluster analysis (HCA) (Jolliffe 1986) can broadly be divided into two methods: agglomerative methods and divisive methods. The divisive methods start with all of the observations in one cluster and then proceed to split (partition) them into smaller clusters. On the other hand, the agglomerative methods initially treat each observation as a separate cluster and then proceed to fuse pair of clusters with smallest distance. The fusion will continue until all observations belong to one cluster. Popular agglomerative methods are single linkage (nearest neighbor approach), average linkage, complete linkage (furthest neighbor) and Ward's method.

Average linkage clustering uses the average similarity of observations between two groups as the measure between the two groups. Complete linkage clustering uses the furthest pair of observations between two groups to determine the similarity of the two groups. Single linkage clustering, on the other hand, computes the similarity between two groups as the similarity of the closest pair of observations between the two groups. Ward's linkage is distinct from all the other methods in that it uses an analysis of variance approach to evaluate the distances between clusters.

The outcome of HCA is a hierarchy or tree-like structure (dendogram) showing the relations among the entities. Dendogram can be interpreted based on the length of branches which are proportional to distance between various clusters.

3.4.4 PLS-DA

PLS-DA is a widely used supervised classification algorithm when dimensionality reduction is needed and discrimination is sought in multivariate analysis (Matthew Barker 2003). In the mass spectrometry data, it is typical to observe that the number variables are more than the number of samples. Moreover, many variables are correlated. The partial least squares (PLS) method permits investigation of complex problem of collinearity (i.e., X-variables). The PLS-DA model establishes the relation between predictor variables (i.e., X matrix) and response variables (i.e., Y matrix) by finding latent variables in such a way that the covariance between the two variables is maximum. The obtained latent variables are linear combinations of old X-variables. Often optimum number of latent variables is needed and can reliably be computed from cross validation procedures.

3.4.5 k-NN

k-NN is a supervised learning algorithm that works without any prior assumptions about the distribution from which training samples are drawn. The training data are vectors in the high-dimensional feature space which is partitioned into regions by class label. The algorithm involves storing of feature vectors and class labels in the learning phase. In order to estimate the class label for the test sample, k-NN computes distances from test sample to all other samples from training set and k nearest neighbours are selected. In order to compute the distances, distance metrics such as Euclidean distance can be employed. Test sample is assigned to most frequent class label among k nearest neighbours in high dimensional feature space. If the tie scenario arises, the ties are broken at random or closer neighbours are given priority. It is common to select larger k value to help reduce noisy effects in data and odd k value to break ties. The optimal choice of k is important and can be selected by cross-validation approach (Duda et al. 2001).

3.4.6 SVM

Support vector machines are a group of supervised methods introduced earlier by Vapnik (Vapnik 1995). The SVMs have gained popularity and have been successfully applied to number of applications including protein structural classification, image recognition, text classification, microarray gene expression data analysis and protein fold recognition (Brown et al. 2000, Cai et al. 2001, Joachims 1998). SVMs are primarily designed for binary classification problems where the training data with two classes are transformed into a high dimensional space by kernel functions. These classifiers rely on hyperplanes corresponding to decision functions. SVM model achieves its objective of classification by constructing optimal hyperplane, i.e., the hyperplane that maximises separation between the two classes. The solution for the classification lies in the support vectors that determine the maximum margin hyperplane. The margin of a linear classifier is the minimal distance of any training point to the hyperplane. Multi-class problem of SVMs can be regarded as multiple binary class problems. One way to solve multi-class classification is using ‘one-versus-one’ approach where the model constructs a binary classifier for every pair of classes, resulting in k (k-1)/2 SVM models for k-class classification problem.

3.4.7 Naive Bayes

Naive Bayes is a probability based classifier and is obtained by assuming conditional independence of the predictor variables given the class label in Bayes theorem (Hand et al. 2001). As a result, likelihood term of Bayes rules can be decomposed into product terms. The classifier basically ignores the potential inter-dependencies such as correlations among the inputs and reduces a complex multivariate problem to a group of simple univariate problems. For a given set of predictor variables, X, the model constructs posterior probability for any event C_j among the set of categorical levels of C. Labeling of new predictor variable to a particular class is based on the highest posterior probability. Naive Bayes methodology simplifies a classification task by allowing the computation of class conditional densities for each variable separately.

3.5 Statistical hypothesis testing

3.5.1 Student's t-test

Two sample t-test (Fisher Box 1987, Snedecor and Cochran 1989) is commonly employed to investigate whether the means of two groups of samples are significantly different from each other. The t-test compares difference in the two means in relation to existing variation in the data. The t-test, as shown in equation (4.3), is a ratio with numerator representing the differences between the means and denominator denoting the measure of variability (standard error of differences) in the data.

The t-test investigates the following hypothesis:

For null hypothesis (H_o): $\mu_1 = \mu_2$

For alternative hypothesis (H_a): $\mu_1 \neq \mu_2$

t-statistic is given by the equation (4.3)

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.3}$$

where $\bar{y}_1$ and $\bar{y}_2$ are the means of the two samples, s_1 and s_2 are the standard deviations of the two samples, and n_1 and n_2 are sample sizes.

3.5.2 Wilcoxon test

Wilcoxon test is a non-parametric (Mann and Whitney 1947, Wilcoxon 1945) equivalent of parametric t-test. The non-parametric test assumes that samples are randomly taken from population with symmetric frequency distribution and does not require that data to follow normal distribution. The test investigates hypothesis on median and can be applied on single sample or two samples (paired or unpaired samples). In single sample case, the test investigates whether the median of sample is different from the hypothesised median of the population. In two samples case, Wilcoxon test investigates median of one sample is different from the second one. Two common non-parametric tests are: Wilcoxon signed-rank test for paired data and the Mann-Whitney U test (also known as Mann-Whitney-Wilcoxon test, the Wilcoxon T test, the Wilcoxon two-sample test, or the Wilcoxon W test) for unpaired data. These tests are based on ranking of the data and looking at the ranks rather than the actual values of the observations.

3.5.3 Analysis of variance

Analysis variance (ANOVA) is used to compare the means of two or more groups using F-statistic under the assumption that sampled population are normally distributed (Snedecor and Cochran 1989). One-way ANOVA allows determining whether one given factor (factor is an independent variable whose values are controlled and varied, for example, in experiments) has significant effect in mean values of any groups in the data.

F-test statistic computes ratio of two sources of variability as below:

F = between group variability / within group variability

ANOVA tests the following hypothesis:

Null hypothesis (H_o): $\mu_1 = \mu_2 = \mu_3 ... = \mu_k$

Alternative hypothesis (H_a): Means of all groups are not equal.

The significant *p* value means that there is at least one group whose mean is different from the rest of groups. One-way ANOVA, however, does not provide information on which group is different from the rest. Post-hoc tests are needed to find which specific group(s) is different from the rest.

3.5.4 Multiple hypothesis testing

The probability of making Type I errors increases rapidly along with the number of hypotheses tested simultaneously. This is called multiple hypothesis testing problem. This problem has become routine in *omics* studies (Farcomeni 2008) where large numbers of statistical tests are performed in the same data set independently on a number of variables. This is the condition where one needs to account for the multiple tests performed. For instance, when employing a t-test for comparison of means across two groups at 5% significant level, the test is willing to accept 5% error i.e., if 100 hypothesis tests are performed, it is expected to see five significantly different variables by chance alone even if there are no actual differences. Many solutions (e.g., Anisimova and Yang 2007) are suggested to account for multiple hypotheses testing including Boneferroni and false discovery rate approaches. Boneferroni correction (Miller 1981), which controls family wise error rate, is the simplest and more conservative correction to account for the multiple hypothesis testing. This correction obtains acceptable significant level by taking into account number of hypothesis tests performed. This is achieved by dividing the p-value of the test by the number of tests performed. Other notable multiple hypothesis testing correction is false discovery rate (FDR) (Benjamini and Hochberg 1995) which is the expected proportion of Type I errors among the rejected hypotheses. It is less conservative approach as compared to family wise error rate correction, which is the probability of making at least one Type 1 error over all hypothesis tests.

3.5.5 Correlations

Correlation describes the degree of relationship between two variables (X and Y) and is measured using correlation coefficient. The value of correlation coefficient varies from 0 (no relationship between X and Y) to 1 (perfect linear relation ship) or -1 (perfect negative linear relationship). A positive value for the correlation implies a positive association (large values of X tend to be associated with large values of Y and small values of X tend to be associated with small values of Y). A negative value for the correlation implies a negative or inverse association (large values of X tend to be associated with small values of Y and vice versa).

The most common measure of correlation is Pearson correlation (Pearson 1896) which is computed using equation (4.4)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \tag{4.4}$$

where r_{xy} is correlation coefficient between X, Y variables, x_i, y_i are series of measurements on X and Y respectively, S_x and S_Y are standard deviations of X and Y respectively and $\bar{x}$ and $\bar{y}$ are sample means of X and Y respectively.

The non-paramteric version of measuring correlations is Spearman's rank correlation (Spearman 1904) which is computed as below:

$$\rho = 1 - \frac{6\sum d_i^{\,2}}{n(n^2-1)} \tag{4.5}$$

where ρ is Spearman's rank correlation coefficient, n is the number of values in each data set and d_i is the difference between the ranks of corresponding values X_i and Y_i.

4. RESULTS AND DISCUSSION

This chapter covers main results related to informatics framework for non-targeted screening of lipids, computational aspects of functional class label prediction for unidentified lipids, lipid pathways instantiation strategy and exploratory analysis of lipidomics data. More detailed information can be found in the original publications **I–VI**.

4.1 Lipid identification

One of the main challenges in lipidomics is to characterize the complete lipid inventory present in a given biological matrix. This challenge motivated the present work to compute spectral libraries for the screening of potential lipid species. This was important especially as there were no publicly available inter-laboratory transferable lipid libraries for the LC/MS platforms.

4.1.1 Database content and basic search

Lipid database (LipidDB) was constructed computationally using SMILES approach (Publication **I**). The LipidDB comprised main classes of lipids such as glycerophospholipids, sphingolipids, glycerolipids, and sterol esters. Specific contents of each lipid class are as shown in the Table 4.1. The enormous structural diversity found in these classes of lipids is due to the differences in length and degree of unsaturation in alkyl chains. Structural rules of specific class (i.e., glycerophospholipids) typically follow a common template which allowed incorporating the structural diversity computationally. The computational framework was based on the construction of “seed” fatty acids most likely to occur in living systems. Each lipid entry was assigned a scoring value based on the seed fatty acid composition to facilitate the searches of experimental results against LipidDB. The scoring value aided when search results were associated

with multiple hits due to isobaric and/or isomeric matches in mass as well as limitations in the analytical approach. Such a heuristic scoring scheme can be modified and the scheme may be different for different cell or tissue types.

Table 4.1 List of different lipid classes and their specific contents in the LipidDB.

Class	**Description**
Fatty Acyls	Alcohols/aldehydes/carboxylic acids and CoAs
Glycerolipids	Mono acyl/alkyl glycerols Diacyl/alkyl glycerols Triacylglycerols
Glycerophospholipids	Glycerophosphocholines, glycerophosphoethanolamines, glycerophosphoserines, glycerophosphates, glyceropyrophosphates and glycerophosphoglycerols.
Sphingolipids	Sphingoid bases, various ceramides including ceramide phosphoinositols, ceramide phosphocholines, ceramide phosphoethanolamines, N-acylsphingosines, N-acylsphinganines, ceramide 1-phosphates and sulfatides.
Sterols	Cholesteryl esters
Plasmalogens (glycerophospholipids)	Phospholipids with vinyl ether bonds at sn-1 position

LipidDB was stored in a native XML database implemented in Tamino XML Server (Software AG). Each lipid entry in LipidDB was described by an internal identifier, scoring information, class, canonical SMILES, molecular formula, molecular weight and isotopic distribution. XML schema for LipidDB is shown in the Figure 4.1. This database was extensively used for the extraction of crucial information using basic search interface as shown in Figure 4.2. The basic search allows queries on lipids species based on their molecular masses, adduct information, lipid classes, fatty acid chain positions, head group information and number of bonds.

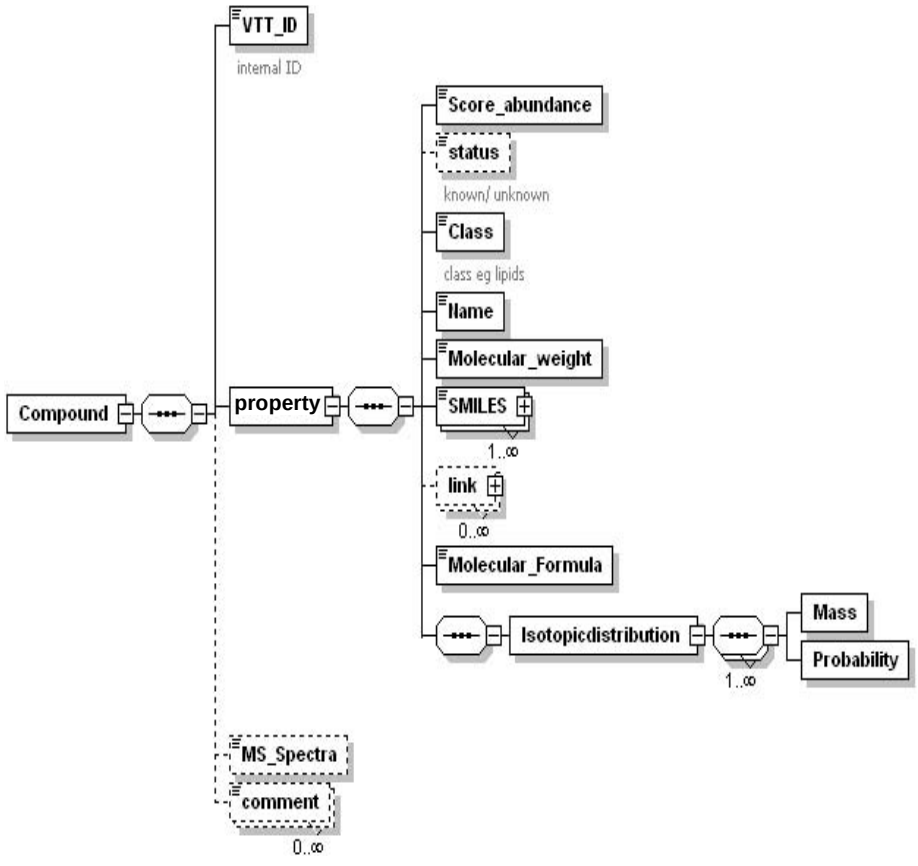


Figure 4.1. XML database schema for LipidDB stored in the native Tamino XML database.

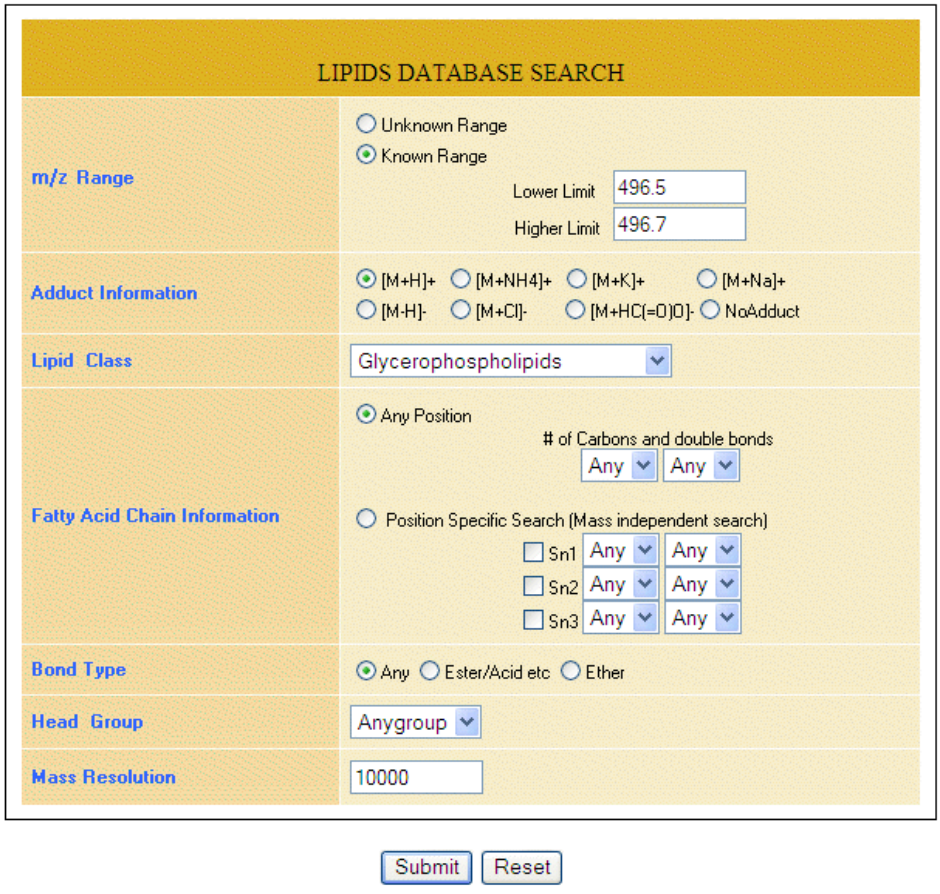


Figure 4.2. Basic search interface for the extraction of lipid information from the underlying LipidDB.

4.1.2 Customisation of LipidDB for UPLC/MS platform

Customisation of experimental information such as retention time, adducts and MS/MS fragmentation in *in silico* database is useful for the screening of potential lipids (Figure 4.3).

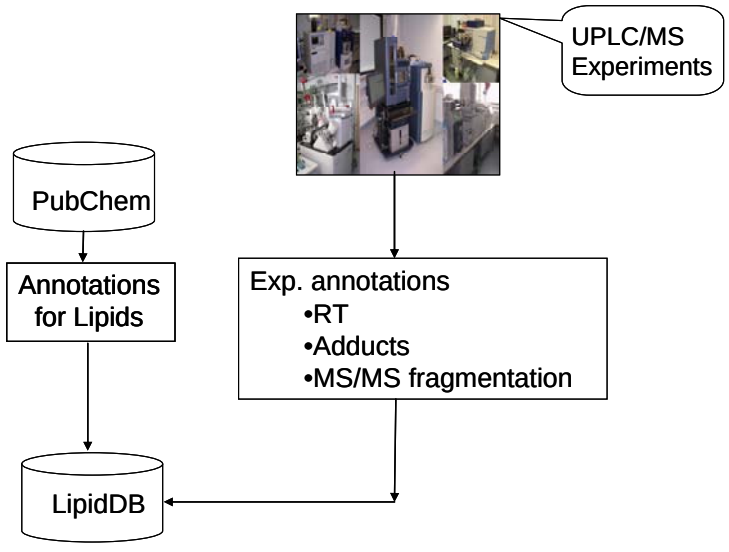


Figure 4.3. LipidDB was updated with external annotations from public databases as well as experimental information from UPLC/MS platform.

Annotations from external databases: Publicly available databases such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) provide valuable information on small molecules including lipids. LipidDB was annotated with physical and chemical properties as well as external IDs of different databases available in the PubChem database.

Addition of retention time ranges and adducts information: In UPLC/MS-based global lipidomics screening, retention time range information is useful parameter and can serve as a coarse filter to avoid false positive hits. Detailed tandem mass spectrometry (MS/MS)-based characterization of all peaks revealed information about retention time ranges for several classes of lipids. The information was obtained from the two dimensional m/z and retention time plot (Figure 5, Publication **II**) generated using MZmine software version 0.60. Characterization of class specific regions and their adduct information was based on MS/MS spectra.

Addition of MS/MS fragmental information: In order to facilitate the identification of lipid species, the main fragmental peaks of acyl glycerols, phospholipids, cholesteryl esters and sphingolipids were included in *in silico* LipidDB. The computational library had greatly facilitated lipid identifications

while comparing with experimental MS/MS spectra. The main fragmentation information of different classes of lipids is summarised as below:

Glycerophosphocholine (PC): In the positive ion mode, PC molecular species form either protonated [M+H]+ or sodiated [M+Na]+ adduct ions as well as dominant peak at m/z 184, representing choline head group. Sodiated PC molecular species produce characteristic fragments at m/z [M+Na-59]+, [M+Na-205]+ and [M+Na-183]+. To facilitate the assignment of fatty acid moieties at sn-1 and sn-2 positions, negative ion mode ESI-MS/MS analysis was performed. In general, negative ion spectra of all phospholipids classes yield four series of ions (Pulfer and Murphy 2003) corresponding to (1) loss of fatty acyl substituents as free fatty acids (2) loss of fatty acyl substituents as ketenes (3) fatty acyl carboxylate anions and (4) head group specific ions. In the negative ion mode, PC molecular species yield ions characteristic for formate adduct [M+HCOO]- and demethylated [M-CH3]- species as well as ions characteristic of the fatty-acyl group esterified at the sn-1 and sn-2 positions. Product ion spectra of plasmanyl (alkyl ether linkage at sn-1 position) / plasmenyl (vinyl ether linkage at sn-1 position) molecular species contains information predominantly related to fatty acyl chain at sn-2 position as well as ion reflecting the loss of sn-2 fatty acyl chain and hence are distinguished from ester-linked phosphatidyl species (Khaselev and Murphy 2000, Zemski Berry and Murphy 2004).

Glycerophosphoethanolamine (PE): Phosphatidylethanolamine, being zwitterionic, can be detected both in positive and negative ion mode mass spectra. Characterization of molecular species as their protonated species [M+H]+ and subsequent yielding of major fragment at m/z [M+H -141]+ (due to loss of head group) is used for the identification of PE molecular species in the positive ion mode. Ethanolamine plasmalogens are detected based on two fragment ions characteristic of sn-1 and sn-2 positions (Khaselev and Murphy 2000, Zemski Berry and Murphy 2004). In the negative ion mode, PE molecular species form deprotonated ([M-H]-) ion which undergoes cleavage of fatty acyls substituents mainly as ketenes. Head groups specific ions are observed at m/z 140 (phosphoethanolamine ion) and 196 (i.e., loss of fatty acyl groups in PE).

Glycerophosphatidylserine (PS): In the positive ion mode, PS is detected as protonated ion. Structural characterization of PS species is mainly done using negative ion mode ESI-MS/MS. In this mode, PS species form [M-H]- ions and [M-H-87]- ions, arising from the loss of serine group upon fragmentation as well as ions corresponding to loss of fatty acyl substituents as ketenes.

Glycerophosphoglycerol (PG): These are less abundant ions of phospholipids. In the negative ion mode, these molecular species yield [M-H]- ions as well as characteristic peaks of lysophospholipids-like fragments due to loss of fatty acyl ketenes/acids and carboxylate anions. Head group specific fragments are detected at m/z 227 and 171.

Glycerophosphoric Acid (PA): PA is the simplest phospholipid and preferentially studied in the negative ion mode in which PA yields deprotonated ion ([M-H]-). Like in other phospholipids, PA forms ions corresponding to neutral loss of acids, neutral loss as ketenes and carboxylate anions. Head group specific ion is detected at m/z 153, a characteristic ion arising from loss of fatty acyls groups from PA species.

Glycerophosphoinositol (PI): Negative mode ESI-MS yields [M-H]- ions of PI. The major fragmentation pathways involves neutral loss of fatty acid, neutral loss as ketenes and loss of the inositol head group (m/z 162, inositol − H2O). A prominent characteristic ion at m/z 241 represents a dehydrated product of inositol phosphate.

Triacylglycerol (TG): ESI-MS of TG species yields ammonium adduct ions which are fragmented in MS/MS to diacylglycerol ([DG]+) like fragments and are similar to those of [DG]+ species in phospholipids due to loss of head groups in phospholipids. These [DG]+ species are informative in identifying TG species. However, in the analysis of mass spectra with co-eluting TG species, it is difficult to assign [DG]+ fragments to its parent TG molecular species correctly.

Cholesteryl Ester (ChoE): ESI-MS platform is not well suited for the analysis of free cholesterol. ChoEs, however, form ammonium adducts in the positive ion mode and generate a fragment ion at m/z 369 upon collision-induced fragmentation.

Sphingolipids: In the positive mode, ESI-MS analysis of sphingomyelin (SM) yields a characteristic protonated phosphocholine peak at m/z 184. PC and SM species are distinguished based on their characteristic m/z value (PC species occur at even at m/z and SM species at odd m/z). Similar to PC, sphingomyelin yields either [M+H]+ or [M+Na]+ ions in the positive mode, while in negative ion mode are [M-CH3]- and [M+ HCOO]- ions.

In the positive ion mode, ceramides form unstable protonated molecular species which undergo dehydration to form [M+H-H_2O]+ ion. Molecular ions in negative ionisation conditions are very informative in identifying the fatty acyl and long chain base substituents of ceramide. Ceramide species yields [M-H]- and [M-H-30]$^-$ (due to loss of HCHO group) ions in negative ionisation conditions. While positive mode analysis of long chain bases such as

sphingosine (d18:1), sphinganine (d18:0) and 4-D-hydroxysphinganine (t18:0) undergo dehydration to form fragments at m/z 282/264, 284/266 and 300/282, negative ion mode analysis of sphingosine and phytosphingosine moieties are characterised by the fragments at m/z 237/263 and 225/255/267 respectively (Merrill et al. 2005, Myoung Hee Lee and Jong 2003).

4.1.3 Building of tissue-specific lipid libraries

Comprehensive mass spectrometry studies allow building of MS/MS fragment libraries for different biological tissues. These libraries are useful for intra-laboratory use in rapidly assigning lipid species coming from new experiments under similar conditions. Tissue-specific libraries were built using the fragmentation characteristics of lipids described in the section 4.1.2. Such MS/MS spectral libraries have limitations due to laboratory-dependent ion source conditions. Moreover, ESI usually produce little structural information and single set of conditions are not applicable for broad spectrum of lipids, or metabolites in general.

Lipoprotein lipidomics in the context of insulin resistance and abdominal obesity

Background: In order to investigate the relation between serum lipid and lipoprotein abnormalities with insulin resistance, sixteen non-diabetic subjects between 18 and 60 years of age were recruited based on a healthy clinical background and modest alcohol consumption. Relevant clinical parameters of all subjects were measured using standard protocols as described in the original publication (Publication **VI**). The lipoprotein fractions such as VLDL, intermediate density lipoprotein (IDL), low density lipoprotein (LDL) and high density lipoprotein (HDL) were separated by sequential flotation in an ultracentrifuge (Taskinen et al. 1988).

In contrast to the traditional measurement of total protein, phospholipid, cholesteryl esters, and TG content in a given biological sample (Vance and Vance 2008), modern MS-based techniques allow analysis at the molecular species level. The MS-based methods have become mainstay of lipidomic research mainly with two strategies: global and targeted approaches. Global (or non-targeted) approaches are directed towards identification and quantification of several hundreds of lipids in a high-throughput basis. In this direction, multiple shotgun-based MS approaches (Ejsing et al. 2009, Han & Gross 2005b) have been developed for wider coverage of different lipid classes. Our recent analysis using UPLC/MS-based platform allowed the analysis of multiple

abundant lipid classes (e.g., triacylglycerols, cholesterol esters, sphingomyelins, phosphatidylcholines) as well as bioactive lipid species (e.g., ceramides, plasmalogens, and lysophosphatidylcholines) simultaneously (Pietiläinen et al. 2007). This global approach does not assume any prior knowledge on type of lipids to be screened and thus providing greater possibility for discovering new classes of lipids. The quantification of lipids in the presented methodology requires further optimization given that lipids cover wide concentration range of compounds. The targeted approaches have been developed for screening of one or few classes of lipids and are more quantitative. Like any other lipidomics methods, the UPLC/MS-based method has limitations in detecting the position and configuration of double bonds within the fatty acid moieties of lipid species. Recent developments on ozone-induced dissociation (OzID) of double bonds (Thomas et al. 2007) can offer a potential solution.

Lipid characterization: Non-targeted profiling was performed to determine the individual species in each lipoprotein fraction. Tandem mass spectrometry was performed both in positive and negative ion modes. Negative mode analysis was mainly utilized to determine fatty acid composition of phospholipids. This extensive study, typically involved examining each individual product ion spectra and compiling spectral information, allowed building of spectral libraries for lipoprotein fractions.

The comprehensive profiling allowed us to detect the compositional details of bioactive lipid species in different lipoproteins. Ceramides were found only in VLDL and LDL, whereas ethanolamine plasmalogens (PE(p)) were found only in LDL and HDL2. Lysophosphatidylcholines (lysoPC) and ether linked phosphatidylcholines (PC(e)) were present in all lipoproteins with the greatest abundance in HDL2, HDL3, and LDL (Figure 2, Publication **VI**). Identification of individual TGs in major lipoprotein particles allowed us to elucidate how changes in different TGs and fatty acids related to features of insulin resistance and abdominal obesity (Figure 1, Publication **VI**).

As a summary, the SMILES-based approach allowed the construction of lipid database for identification of lipids in mass spectrometry analysis. The direct application of databases as described in section 2.4 for mass spectrometry is limited due to the nature of LC/MS-based analysis. The LC/MS-based spectral libraries are very much dependent on the type of scanning mode and instrumental settings and hence it is very unlikely any single database can act as a standard reference for all types of lipidomic analysis. This is also partly due to

large diversity of lipids across different organisms, tissues, and cell types. Here, the developed database was customized to global screening methods by addition of RT information, adducts, mass fragments and other annotations from external databases. The database was extensively used in the construction of tissue specific spectral libraries.

4.2 Functional class label prediction of unidentified peaks

Mass spectrometry-based metabolomics experiments often results in unidentified peaks which can hamper in the interpretation of results. The problem is even more challenging in the non-targeted metabolome screening experiments since modern MS instruments have ability to detect several hundreds of peaks in a given sample. Even when the identification is relatively easier for lipids if MS/MS spectra are obtained, non-targeted UPLC/MS approach often results in unidentified peaks due to analytical limitations resulting from small peaks, co-fragmentation, ambiguous spectra, as well as complex spectra probably coming from modified and uncommon lipids. In fact, data analysis (univariate or multivariate analysis) often lead to interesting peaks which are often unidentified and may serve as potential biomarkers. Strategies are therefore needed to interpret the data when exact identifications are not available. The challenge was addressed computationally by predicting functional class labels for unidentified peaks (Publication **II**). This methodology serves as a helpful intermediate step in data analysis as well as a guide towards the further steps to identify the compounds.

Computational methodology was demonstrated using the lipidomics data from our earlier twin pair study (Pietiläinen et al. 2007). The lipidomic data were preprocessed using an MZmine software version 0.60 (Katajamaa and Orešič, 2005, Katajamaa et al., 2006). Based on identifications, lipids were assigned to one of the following classes: glycerophosphocholines (PC) glycerophosphoethanolamines (PE), sphingomyelins (SM) and triacylglycerols (TG). Cross validation strategy was employed to assess the generalisation performance of the classifiers. Here, computational work involved single cross validation method for models with no meta-parameter estimation (i.e., random model and Naive Bayes) and double cross validation for models requiring metaparameter estimation (i.e., k-NN, SVM and PLS-DA). Details of single and double cross validation methods are available in original publication (Publication **II**). Main results of employed supervised classifiers (PLS/DA, SVM, Naive Bayes and k-NN) are summarised below.

4.2.1 PLS/DA

PLS-DA is a latent variable-based supervised classifier and was investigated to evaluate the performance in predicting the functional labels for unlabeled peaks. Repeated double cross validation setting was employed to estimate the optimal number of latent variables, which was found to be mainly between 5 and 20 components. The optimum number of components from inner loop of double cross validation was used to evaluate the performance of classifier on corresponding test sets in the outer loop. The errors made in predicting test set labels were computed and overall accuracy of the PLS-DA classifier was found to be 63.3%. In all, the classifier performed slightly better than both Naive Bayes and random models and underperformed as compared to k-NN and SVM models (Table 1, Publication **II**)

4.2.2 SVM

SVM classifier was used to solve multiclass-classification problem using 'one-against-one' approach which trains L(L-1)/2 (L = number of levels) binary classifiers. The appropriate class was found by the majority voting scheme. Linear kernel function was utilised and the performance of classifier was evaluated using four-fold cross validation study with double cross validation being repeated for 25 times. Double cross validation was used to select an optimal regularisation parameter (C). The parameter corresponding to minimum cross validation error was varied mostly between 100 and 1000. The cross validation accuracy of SVM classifier was found to be 92.83 %. Prediction accuracies of SVM model on lipid classes were found to be better than those predicted by k-NN and Naive Bayes models (Table 1, Publication **II**).

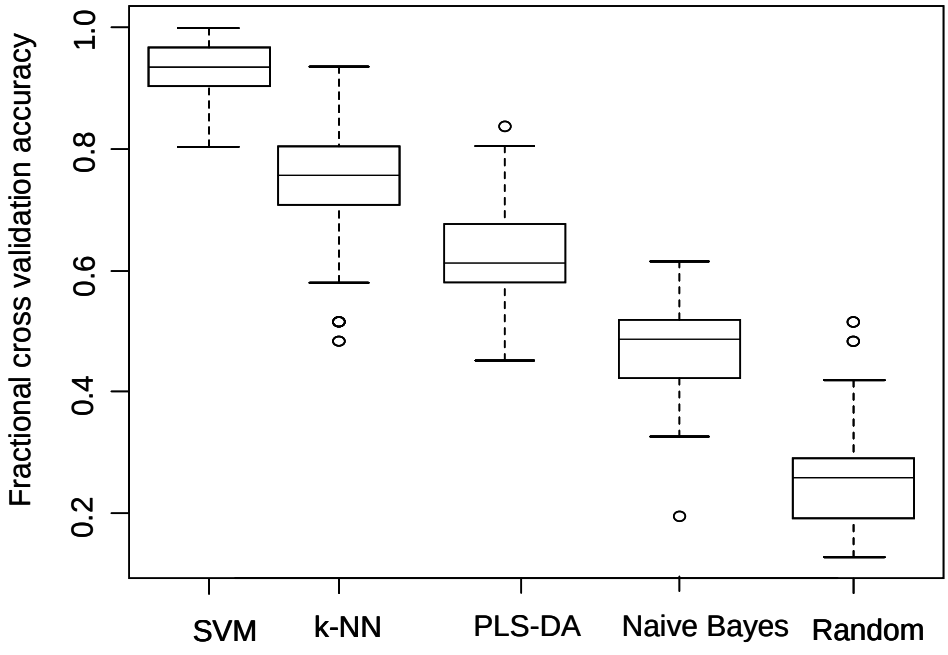


Figure 4.4. Comparative cross validation performances of SVM, k-NN, PLS-DA, Naive Bayes and random classifiers.

4.2.3 Naive Bayes

Naive Bayes classifiers are probabilistic version of classifiers. It was also trained using four-fold cross validation repeated 25 times. Cross validation accuracy in class label prediction on test set as well as individual classes are shown in Table 1 of Publication **II**. The performance based on cross validation accuracy of Naive Bayes model was found to be poor as compared to k-NN and SVM models.

4.2.4 k-NN

k-NN classifier requires estimation of parameter k to decide appropriate class label for a given test sample. The value of k represents the number of nearest training samples in the feature space considered when deciding class label for the test sample. Repeated four-fold cross validation procedure was employed to select optimal k and performance evaluation of the k-NN classifier. The average cross validation accuracy of k-NN classifier for the whole test set as well as for

each lipid class was summarized in Table 1 of Publication **II**. Cross validation accuracy of k-NN model was 75.97 %. k-NN classifier outperformed random assignment model both in terms of class-specific as well as over all cross validation performance. Naive Bayes model showed poor cross validation performance as compared to k-NN model on complete test set as well as TG and SM classes. Better prediction accuracy of lipid classes (TG, PC and SM) demonstrated predictive ability of k-NN classifier in lipid data.

Here we attempted to make use of unidentified peaks in the lipidomic analysis by predicting the functional labels using supervised classifiers. The k-NN and SVM classifiers outperformed Naive Bayes and PLS-DA classifiers. More robust prediction could be achieved by utilizing consensus predictions from both SVM and k-NN classifiers. The poor performance of Naive Bayes classifiers could be attributed to class-specific co-regulations of lipids. This computational framework complements the existing identification methodologies with predictions of class labels to facilitate exploratory analysis. These kinds of appproaches are more useful in the global screening approaches in the metabolomics where the identification of metabolites is a bottleneck.

4.3 Reconstruction of lipid pathways

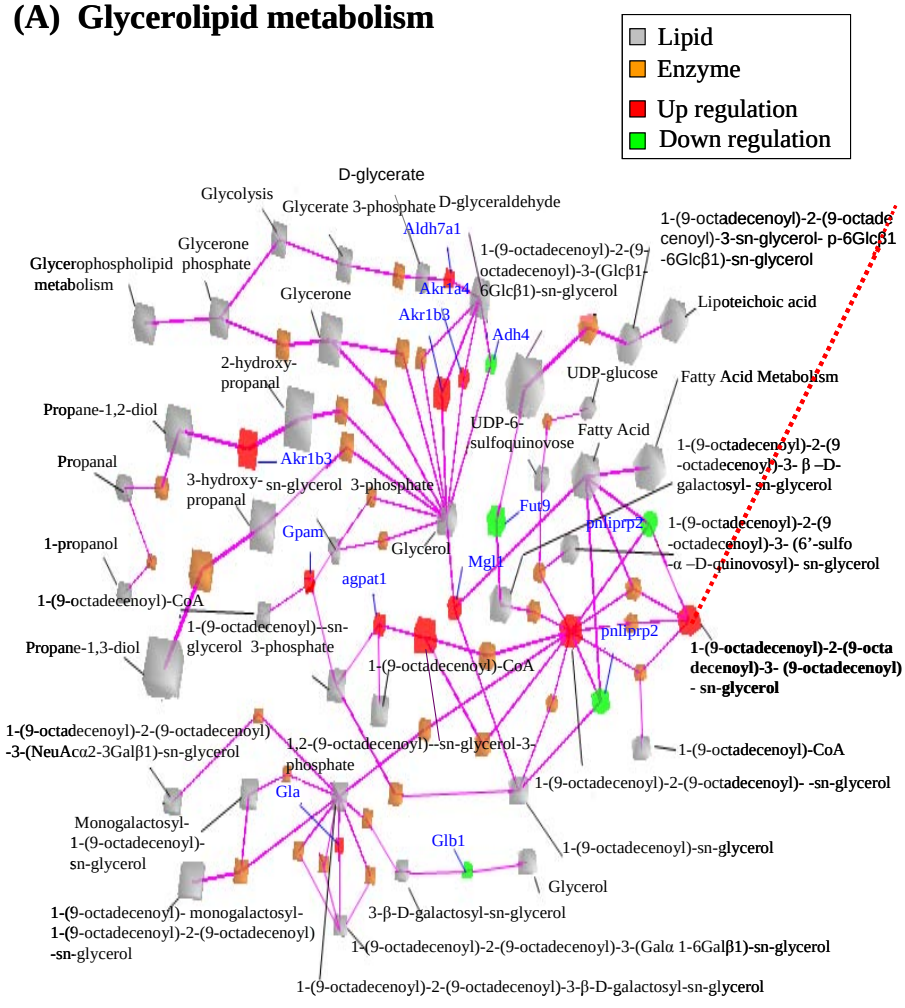
Biochemical pathways are rich sources of information and may help to gain mechanistic links behind underlying phenotype. Properly organized and curated databases are required to view or construct biochemical pathways reliably. Databases such as KEGG PATHWAY database provide information on available lipid pathways besides other biochemical pathways. The database also provides generic, organism-independent biochemical pathways that serve as reference pathways for constructing organism-specific pathways. The available databases therefore serve as a good starting point for pathway analysis.

4.3.1 Pathway instantiation

Biochemical research is empowered with modern analytical techniques which can provide plenty of detailed molecular information. As a result, available pathway databases need to accommodate these changes to the resolution of available information. This problem is obvious particularly in the case of lipids. The essential building blocks for the molecular pathway instantiation were explained in the original publication (Publication **I**).

Each node in Figure 4.5 is either a lipid metabolite or enzyme or other interconnecting metabolic pathway. Grey color represents metabolites/other metabolic pathways and brown represents enzymes. Up- and down-regulation of corresponding nodes are denoted by red and green colors, respectively. Enzyme names are shown only if they are differentially regulated (~ 1.5 fold change). Glycerolipid pathway instantiation was demonstrated with TG(18:1/18:1/18:1) (Figure 4.5A) lipid species where experimental measurements could be mapped directly on pathway unlike in generic pathways and thus bridging the gap between MS data and existing lipid pathways. From the sphingolipid pathway map (Figure 4.5B), two enzymes linked to the ceramide via metabolic reactions, one is SGPP1 (Sphingosine-1-phosphate phosphatase 1, UniProt ID Q9JI99), the other GALC (galactosylceramidase, UniProt ID P54818) were upregulated in ob/ob. SGPP1 is involved in *de novo* ceramide synthesis, while GALC hydrolyses galactosylceramide to form ceramide. Interestingly, sphingomyelin SM(d18:1/18:0) a precursor of ceramide via the sphingomyelinase reaction is downregulated, while the sphingomyelinase level is maintained. Therefore, these results indicate that both glycolipids and free fatty acids may act as a source of the elevated ceramides in the ob/ob fatty liver.

(A) Glycerolipid metabolism



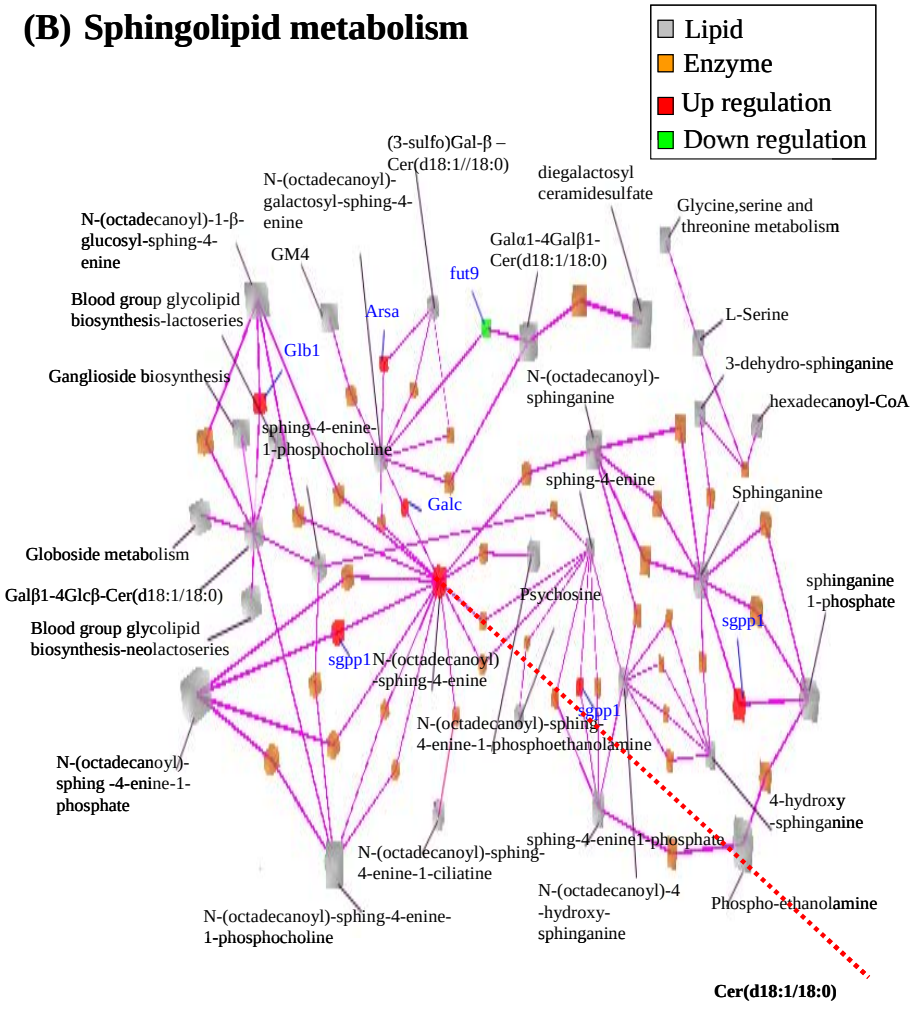


Figure 4.5A, B. Instantiated pathways for Cer(d18:1/18:0) and TG(18:1/18:1/18:1). A) Instantiation of TG(18:1/18:1/18:1) as part of glycerolipid metabolism. (B) Instantiation of Cer(d18:1/18:0) as part of sphingolipid metabolism.

4.3.2 Tissue- and context -specificity in lipid pathways

In order to facilitate any modifications in the existing pathways, the megNET visualisation tool is extended to construct pathways for a given list of EC numbers and lipid data. The developed frame work is now being targeted towards building context-sensitive and tissue-specific pathways. As a part of new developments, a new database was compiled with all available lipid-related

proteins with their pathway information such as KEGG pathways and context information such as gene ontology terms. Knowledge about the tissue-specific enzymes was incorporated at transcriptomic level. The presence or absence level information of mRNA in tissues is expected to facilitate the development of tissue-specific pathways.

4.4 Lipid profiling applications and data analysis

High-throughput molecular profiling technologies provide an opportunity to measure lipids on an unprecedented scale. The large amount of data presents a major challenge for statistical methods to handle and assemble proper knowledge for biomarkers discovery efforts. These analyses typically start with unsupervised methods where the main aim is to get accurate knowledge on samples whether they really cluster or provide trends in the data. These techniques also serve as good visualisation tools in data analysis. Unsupervised methods such as principal component analysis (PCA) are useful to capture the trends mainly when there are correlated variables in the data (e.g., lipid data). Such PCA-assisted analysis of projecting samples into lower dimensional space from high dimensional space not only allow one to confirm the expected patterns in the data based on the group membership of samples but also facilitates in detecting outliers in samples. Supervised methods such as PLS/DA can also be employed to describe complex data with few latent components as well as a obtain set of most distinguishing variables (lipids) among the groups of samples.

4.4.1 Lipidomic profiling of multiple tissues of the POKO mice

Increased obesity is the one of the risk factors for type 2 diabetes. The relationship of how obesity causes the diabetes still remains unknown. It is hypothesised that when adipose tissue reaches its full capacity, the excess fat spills over to other metabolically active organs such as liver, pancreas and skeletal muscle. This condition leads to insulin resistance and diabetes (Gray and Vidal-Puig 2007). Earlier studies indicate that peroxisome proliferator activated receptor gamma (PPARγ) plays a key role in adipogenesis and insulin sensitivity (Koutnikova et al. 2003, Rosen et al. 1999, Spiegelman 1998). However, the importance of PPARγ2, which is nutritionally regulated isoform of PPARγ, is still not clear.

In order to investigate the physiological importance of PPARγ2 under positive energy balance conditions in ob/ob mice (PPARγ2þ/þ Lepob/Lepob) (Publication **V**), lipidomic analyses were performed on relevant tissues. Profiling of adipose tissue, pancreatic islets, liver, and skeletal muscle samples revealed distinct differences in four genotypes: WT (PPARγ2þ/þ Lepþ/Lepþ), PPARγ2 KO (PPARγ2_/_ Lepþ/Lepþ), ob/ob (PPARγ2þ/þ Lepob/Lepob), and POKO (PPARγ2_/_ Lepob/Lepob). The adipose tissue from POKO mice was characterised by decreased triacylglycerols (TGs) and increased diacylglycerols (DGs). These changes were associated with increased levels of ceramides. Lipid profiling revealed decreased TG and DG levels and increased ceramides (Cer) levels in POKO islets, indicating the possible role of PPARγ2 in promoting TGs levels to increase the lipid-buffering capacity of β-cells thereby preventing lipotoxicity. Liver and skeletal muscle lipidomics revealed decreased TGs and increased formation of bioactive lipid species such as ceramides and lysophosphatidylcholines in POKO mice compared to ob/ob mice. In all, lipidomic profiling of four tissues showed similar pattern of changes.

The study on 16-week-old mice revealed increased levels of ceramide in POKO islets as compared to *ob/ob* islets. The study was later extended to investigate whether perturbed lipid metabolism is already present in islets at 4–5 weeks of age (Publication **IV**). There were no statistically significant changes in lipid composition among the four genotypes of mice (Figure 4.6) unlike in the 16-week-old mice. The statistical significance was based on one way analysis of variance (1-way ANOVA) and associated p value was adjusted for multiple hypothesis testing. Analysis in other metabolically active tissues such as serum, liver, adipose tissue and muscle from WT, PPARγ2 KO, *ob/ob* and POKO mice, however, showed significant lipid compositional changes in 4-week-old mice**.** In serum, POKO mice had higher levels of TGs and high levels of short- and medium-chain PC species as compared to mice with other genotypes. Interestingly, the levels of long-chain TGs were lower in ob*/ob* mice than in WT and PPARγ2 KO mice. Lipidomic characterization of liver tissue revealed increased TG levels (mainly short- and medium chain TGs) at 4 weeks of age in both POKO and *ob/ob* mice as compared to WT mice. Ob/ob and POKO mice livers had increased levels of medium chain PCs when compared with WT and PPARγ2 KO mice. Unsaturated long-chain TGs were enriched in POKO and *ob/ob* livers. Lipidomic profiling in adipose tissue revealed similar levels of short-, medium- and long-chain TGs in POKO and *ob/ob* mice. Polyunsaturated long-chain TGs were enriched in POKO and *ob/ob* mice as compared to WT and

PPARγ2 KO mice. Lipidomic data from skeletal muscle showed that 4-week-old ob/ob mice had increased short- and medium-chain TGs as compared to other genotypes. Muscle from POKO and PPARγ2 KO mice contained more long-chain TGs than WT and *ob/ob* muscles. Interestingly, the levels of ceramides and lysoPCs increased similarly in both *ob/ob* and POKO mice**.** Taken together, lipdomic analysis of relevant tissues from 4-week-old mice suggested abnormal accumulation of TGs and the resulting lipotoxicity may contribute to the severity of the metabolic syndrome in 16-week-old POKO mice**.** These studies and more specifically, the differences observed between the liver, serum and muscle, may also indicate the possibility of a hierarchical order of organs with respect to fat deposition and lipid-induced toxicity.

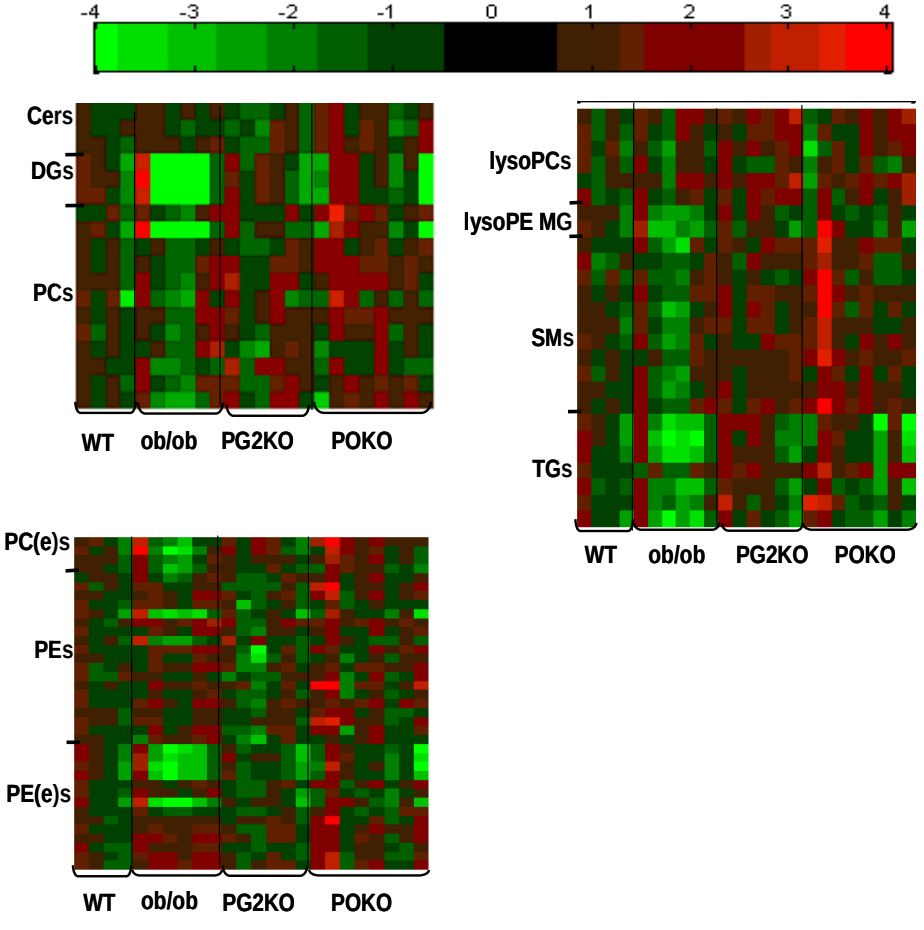


Figure 4.6. Lipidomic profiling of 4-wk-old mice islets from wild type, PPARγ2 KO, ob/ob and POKO mice (n = 5–8). No significant lipids with ANOVA p-values < 0.05 were found.

We further investigated the 4-wk-old lipidomics data using chemometric approaches to find which specific lipid variables are responsible for the separation of the four genotypes. Serum lipid profiles from four genotypes were used to demonstrate the usefulness of exploratory analysis. PCA analysis was performed to detect outlier samples and check whether samples of same genotypes are clustered together. X-block was autoscaled prior to PCA analysis. Separation of four genotypes was not very clear (Figure 4.7A) and more fine clusters were obtained using PLS/DA analysis. Outliers detected in PCA were removed prior to PLS/DA. The data were preprocessed by autoscaling X-block data and mean-centering Y-block data. Cross validation (contiguous block cross validation method) and Q^2 scores were used to optimise the PLS/DA model.

The variable importance in the projection (VIP) values (Wold et al. 1987) were computed to identify most important lipid species contributing to separation of four genotypes in the PLS/DA model. Top scoring VIP lipid variables responsible for separation each group were found. For instance, VIP plot of serum lipidomics for wild type is shown in Figure 4.8. Top VIP scoring lipids from the plot are SM(d18:1/22:0), SM(d18:1/24:1), TG(58:9), PE(36:2) and PE(38:6e). Similar VIP analysis was also performed on other genotypes to find their respective variables responsible for the separation among the genotypes.

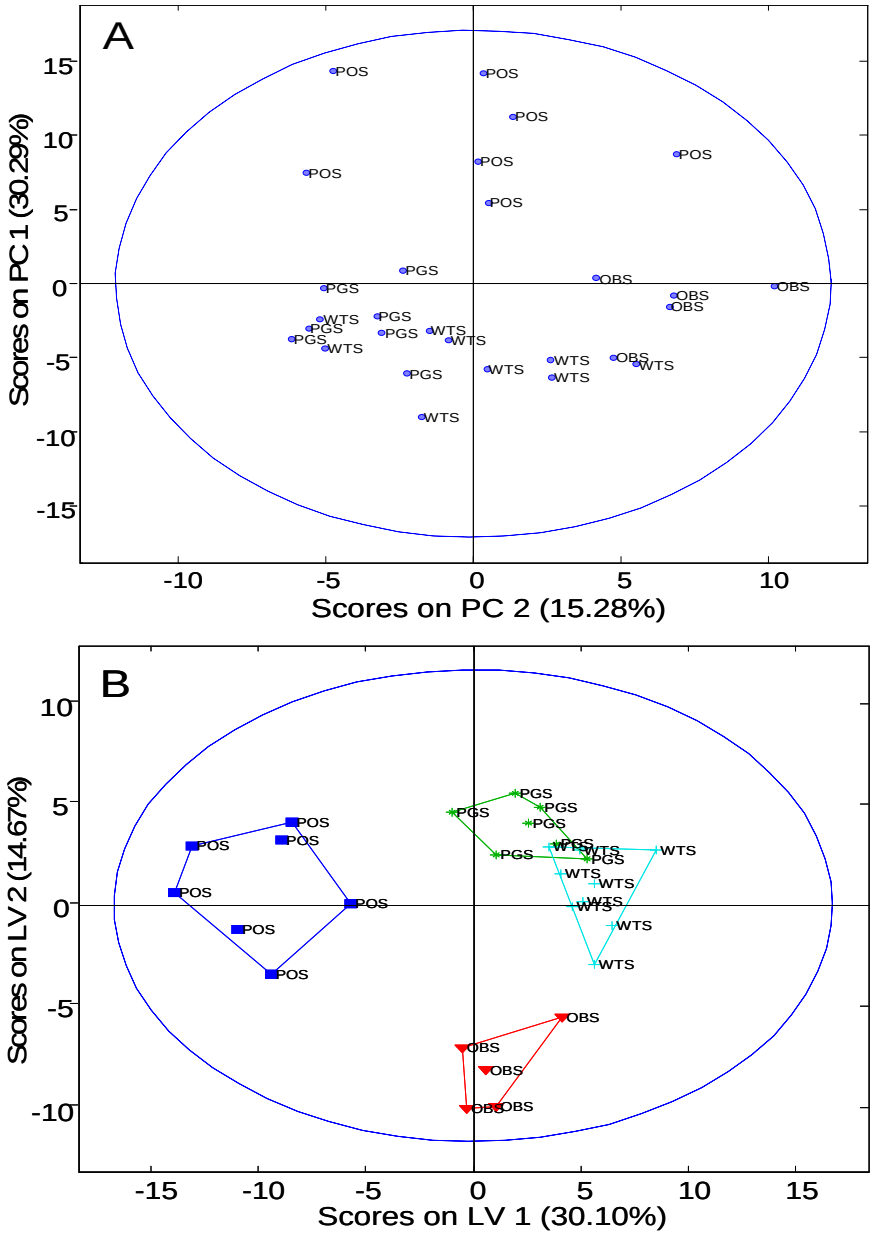


Figure 4.7A, B. Exploratory analysis of serum lipidomic data in the ob/ob mice model for four genotypes. Legends WTS, PGS, OBS and POS represent wild type, PPARγ2 KO, ob/ob and POKO mice respectively.

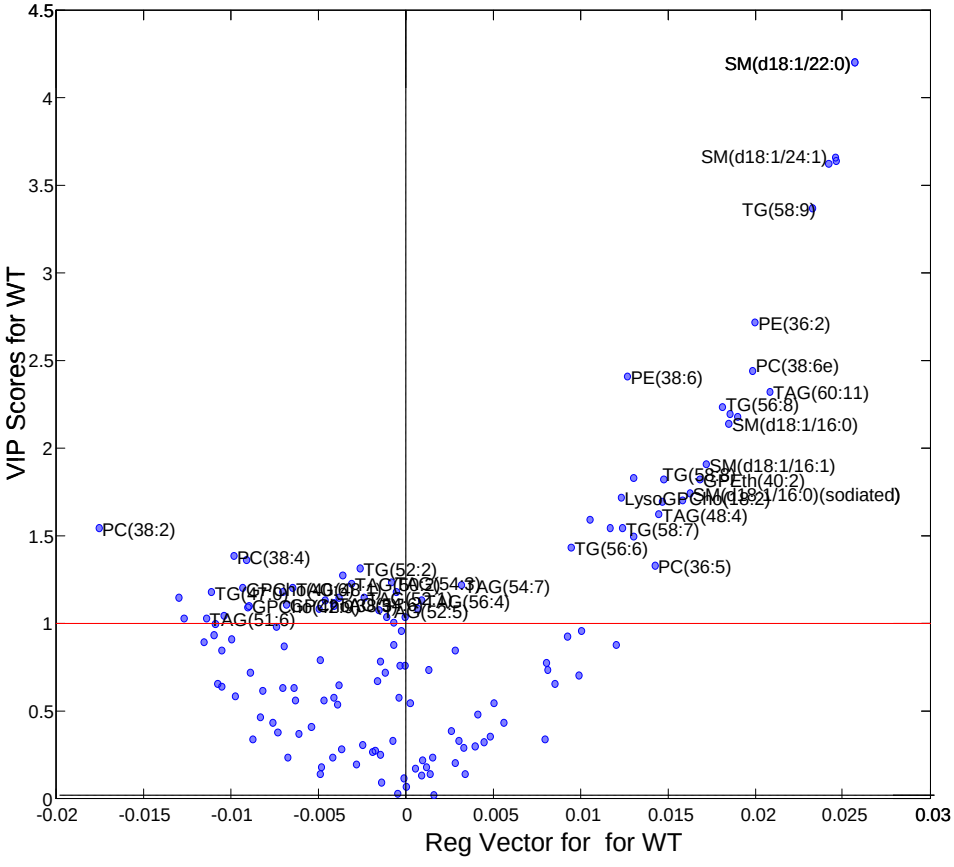


Figure 4.8. VIP analysis of serum lipidomic data for wild type group.

4.4.2 Lipidomic characterization of low and high HDL-C subjects

Low levels of high density lipoprotein cholesterol (HDL-C) is a recognised risk factor for heart disease. However, the mechanisms of how low HDL-C is contributed to cardiovascular diseases are still unknown. Here, lipidomics was utilised to study HDL derived from well characterized high and low HDL-C subjects**.** The study comprised 47 subjects: 24 low-HDL subjects and 23 high-HDL subjects who were participants of the Health 2000 Health Examination Survey. The subjects represented the extreme ends of HDL-C levels ($\leq 10^{th}$ and $\geq 90^{th}$ percentiles) and the HDL-C limits were as follows: for low-HDL-C men $\leq$ 1.03 mmol/l, low-HDL-C women $\leq$ 1.23 mmol/l, high-HDL-C men $\geq$ 1.79 mmol/l, and high-HDL-C women $\geq$ 2.24 mmol/l. Subjects with diabetes, alcohol abuse,

or malignancy were excluded. Alcohol abuse was defined as >160 grams of alcohol / week for women and >310 grams of alcohol / week for men.

Non-targeted lipidomic analysis was performed on low and high HDL-C subjects (Publication **III**) to uncover the differences in lipid composition and the resulting information was used to reconstitute HDL particles computationally.

Univariate analysis from both clinical measurements and lipidomic profiles was performed in order to find the parameters characterising low and high HDL-C subjects. Clinical and biochemical characteristics of low and high HDL-C subjects are summarised in (Table I, Publication **III**) and lipid species selected based on *p*-values ($p < 0.0001$) from student *t*-test between high and low HDL-C subjects are summarised in Table S1 of Publication **III**. Box plots for the most abundant lipids from lysoPC, SM, ChoE, ethanolamine plasmalogen (PEp) and TG classes are shown in Figure 4.10.

Supervised classification model was built for clustering and discrimination using partial least squares discriminant analysis (PLS/DA). The random subsets cross validation method and Q^2 scores were used to optimise the models. Based on the cross-validation, the model with two latent variables and $Q^2 = 0.51$ was selected. PLS/DA scores plot revealed clear separation between the two HDL-C groups (Figure 4.9A). Additionally, VIP analysis was performed to discover lipid variables responsible for the observed separation. Identified lipids with VIP value greater than two were further explored using fold changes and hierarchical clustering analysis. Heat map of fold changes with both samples and lipid variables ordered by hierarchical clustering is shown in Figure 4.9B. The clustering was based on Euclidean distance measure. We found that top VIP lipid variables from PLS/DA model formed two predominant clusters: mostly low HDL-C subjects and mostly high HDL-C subjects. Few mis-clustered samples may partly be attributed to individual variability. The fold changes in heat map were reflecting lipid profile changes relative to the average intensity of lipids within low HDL-C subjects. Bar plot in Figure 4.9B shows the mean fold change value of top VIP lipids within HDL-C subjects as compared to low HDL-C subjects.

Linear association of top VIP lipid variables with measured clinical variables were investigated using correlation analysis. Pearson correlations were computed between clinical and lipid variables. Both clinical and lipid variables were clustered using hierarchical clustering (Figure S3, Publication **III**). Interesting correlations were observed between lipid variables and HDL-C clinical parameters. Concentrations of TG molecular species were negatively

correlated with HDL-C concentration, while the SM and lysoPC lipid species were positively correlated with HDL-C parameter. However, PCs did not show any general trend in correlation with HDL-C parameter. We further investigated HDL-C concentration with some top lipid variables in both low and high HDL-C subjects (Figure S4 and S5, Publication **III**). Positive correlations between SM(d18:1/16:0) and HDL-C were similar in low HDL-C subjects ($r = 0.71$, $p = 0.0001$) and high HDL-C subjects ($r = 0.71$, $p = 0.0001$). Interestingly, positive correlation between lysoPC(18:0) and HDL-C in low HDL-C subjects ($r = 0.54$, $p = 0.006$) disappeared in high HDL-C subjects ($r = -0.06$, $p = 0.78$). We observed no correlation of TG(16:0/18:1/20:1) with HDL-C in low HDL-C subjects ($r = 0.06$, $p = 0.75$) and negative correlation in high HDL-C subjects ($r = -0.43$, $p = 0.04$). Additionally, linear association was also investigated for ChoEs with HDL-C. No significant correlations were found with concentrations for ChoE(18:1) and HDL-C variable either in low HDL-C subjects ($r = 0.33$, $p = 0.12$) or high HDL-C subjects ($r = 0.02$, $p = 0.92$).

The lipidomic level studies enabled to have a closer look at the molecular level details which are used in the simulation studies. While TLC or HPLC methods may serve as a valuable tool to analyse class-specific changes, lipidomics analysis not only guided to reconstitute HDL particle using simulation studies but also enabled us to investigate lipid molecular composition (Publication III). All previous simulations in the field have been based on either a single-component lipid particle composed of phosphatidylcholines lipids, or a two-component mixture of phosphatidylcholines and cholesteryl esters.

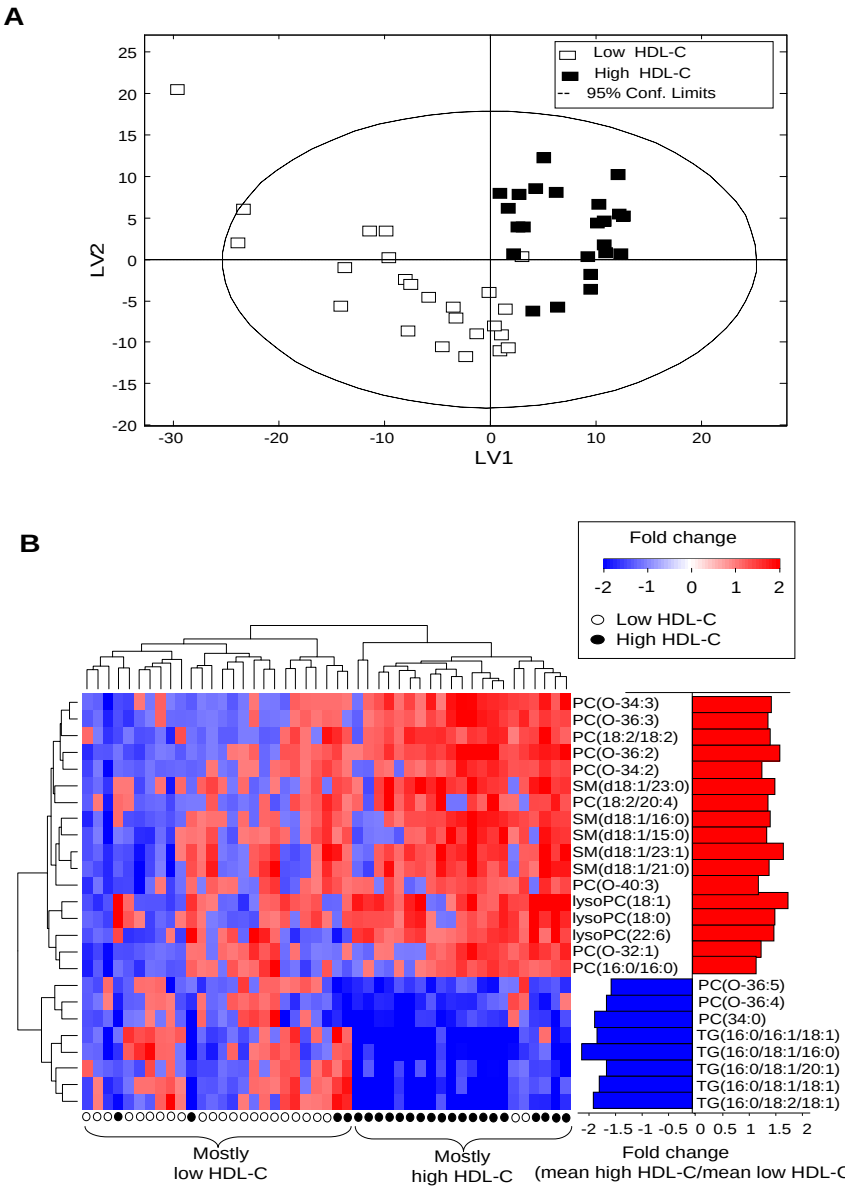


Figure 4.9 A. Partial least squares discriminant analysis (PLS/DA) of lipidomic profiles for low HDL-C and high HDL-C subjects. PLS/DA scores plot with two different HDL-C groups are indicated. B. Hierarchical clustering on most important VIP variables and samples in the heat map reflecting fold changes of lipids relative to mean intensity within low HDL-C group.

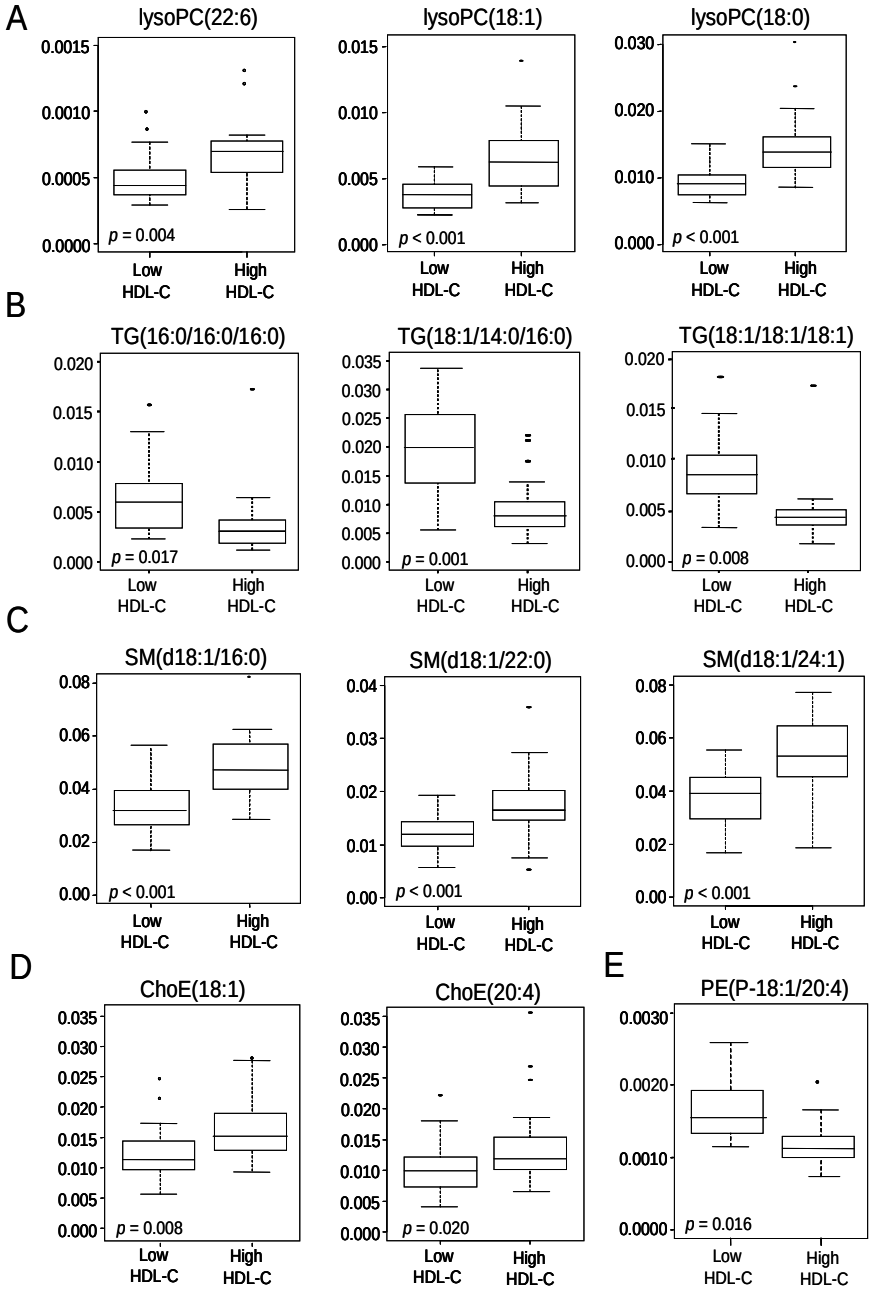


Figure 4.10. Box plots of the most abundant lipids within the TG, lysoPC, SM, ChoE and ethanolamine plasmalogen (PEp) classes. Concentrations are shown in mmol/l [lipid] / mg/dl [apoA-I].

5. SUMMARY AND CONCLUSIONS

Modern MS-based analytical technologies have generated wealth of information in lipidomics studies. The information presented many challenges for bioinformaticians due to its complexity. The goal of this thesis was to address such challenges in the context of non-targeted lipidomics studies based on UPLC/MS experimental methodologies.

A bioinformatics approach for the construction of lipid database for major classes of lipids is presented. Each lipid entry in the database was assigned with relevant information such as lipid names including short names, SMILES, scores, molecular weight, monoisotopic mass, isotope distribution (Publication **I**). The database was customised for UPLC/MS experiments by incorporating the information such as retention time range, adducts and main fragments to screen for potential lipids. This database information facilitated greatly building of experimental tandem MS libraries for different biological tissues. An example of such spectral libraries was built for different lipoprotein fractions (Publication **VI**).

Non-targeted metabolomic studies are often challenged by the presence of unknowns and hence present an additional challenge for the interpretation of experimental data. Frequently used supervised classification methods were employed for the functional prediction of class labels for unknown lipids to facilitate exploratory analysis as well as simplify the identification process (Publication **II**). As lipidomics goes beyond detecting the complete inventory of lipids, a new strategy called pathways *instantiation* is proposed to understand lipids in the context of pathways and thereby providing insights for the phenotype characterization (Publication **I**).

Lipid profiling was successfully applied to study mouse models in the context of POKO mice (Publication **IV** and **V**) and humans in the context of high and low HDL-C subjects (Publication **III**). Chemometric methods such as principal component analysis (PCA) and partial least squares and discriminant analysis

(PLS/DA) were employed for exploratory analysis as well as biomarker discovery in the context of different phenotypes such as characterization of high and low HDL-C subjects (Publication **III**) and POKO mice (Publication **IV**).

5.1 Future perspectives

While bioinformatics strategies presented herein facilitate lipidomic studies, the existing lipid informatics methods need to be extened to include more complex lipids and their mass spectra to databases to facilitate screening lipids such as steroids and glycolipids. The computational methodologies can be developed for the classification of unidentified metabolites for other analytical platforms. The novel tools are expected to integrate lipidome data with other omics level information in a context-dependent manner and thereby establishing lipid networks for underlying phenotypes. The complexity of lipidomes and their regulation at multiple levels makes their study a challenge for bioinformaticians and computational biologists.

REFERENCES

Anisimova, M., and Z. Yang. 2007. Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. *Mol. Biol. Evol.* 24(5): 1219–1228.

Apollonio, L. G., D. J. Pianca, I. R. Whittall, W. A. Maher, and J. M. Kyd. 2006. A demonstration of the use of ultra-performance liquid chromatography-mass spectrometry [UPLC/MS] in the determination of amphetamine-type substances and ketamine for forensic and toxicological analysis. *J. Chromatogr. B* 836(1–2): 111–115.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57(1): 289–300.

Bennet, R. D., and E. Heftmann. 1962. Thin-layer chromatography of sterols. *J. Chromatogr.* 9(9): 359–62.

Bijlsma, S., I. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen, and A. K. Smilde. 2005. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Anal.Chem.* 78(2): 567–574.

Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc.Natl. Acad. Sci. USA 97(1): 262–267.

Brugger, B., G. Erben, R. Sandhoff, F. T. Wieland, and W. D. Lehmann. 1997. Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem masspectrometry. *Proc. Natl. Acad.Sci. USA* 94(6): 2339–2344.

Caffrey, M., and J. Hogan. 1992. LIPIDAT: A database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis. *Chem. Phys. Lipids* 61(1): 1–109.

Cai, Y.-D., X.-J. Liu, X.-b. Xu, and G.-P. Zhou. 2001. Support Vector Machines for predicting protein structural class. *BMC Bioinformatics* 2(1): 3.

Churchwell, M. I., N. C. Twaddle, L. R. Meeker, and D. R. Doerge. 2005. Improving LC-MS sensitivity through increases in chromatographic performance: Comparisons of UPLC-ES/MS/MS to HPLC-ES/MS/MS. *J. Chromatogr. B* 825(2): 134–143.

Cotter, D., A. Maer, C. Guda, B. Saunders, and S. Subramaniam. 2006. LMPD: LIPID MAPS proteome database. *Nucl. Acids Res.* 34(suppl_1): D507–510.

Cutler, R. G., J. Kelly, K. Storie, W. A. Pedersen, A. Tammara, K. Hatanpää, J. C. Troncoso, and M. P. Mattson. 2004. Involvement of oxidative stress-induced abnormalities in ceramide and cholesterol metabolism in brain aging and Alzheimer's disease. *Proc.Natl. Acad. Sci. USA* 101(7): 2070–2075.

Druilhet, R. E., M. L. Overturf, and W. M. Kirkendall. 1975. Structure of neutral glycerides and phosphoglycerides of human kidney. *Int. J. Biochem.* 6: 893–901.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*, John Wiley & Sons, New York.

Ejsing, C. S., E. Duchoslav, J. Sampaio, K. Simons, R. Bonner, C. Thiele, K. Ekroos, and A. Shevchenko. 2006. Automated Identification and Quantification of Glycerophospholipid Molecular Species by Multiple Precursor Ion Scanning. *Anal. Chem.* 78(17): 6202–6214.

Ejsing C.S., J.L. Sampaio, V. Surendranath, E. Duchoslav, K. Ekroos, et al. 2009. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *PNAS* 106(7): 2136–2141.

Ekroos, K., I. V. Chernushevich, K. Simons, and A. Shevchenko. 2002. Quantitative profiling of phospholipids by multiple precursor ion scanning on a hybrid quadrupole time-of-flight mass spectrometer. *Anal. Chem.* 74(5): 941–949.

Fahy, E., S. Subramaniam, H. A. Brown, C. K. Glass, A. H. Merrill, Jr., R. C. Murphy, C. R. H. Raetz, D. W. Russell, Y. Seyama, W. Shaw, T. Shimizu, F. Spener, G. van Meer, M. S. VanNieuwenhze, S. H. White, J. L. Witztum, and E. A. Dennis. 2005. A comprehensive classification system for lipids. *J. Lipid Res.* 46(5): 839–862.

Fahy, E., S. Subramaniam, R. C. Murphy, M. Nishijima, C. R. H. Raetz, T. Shimizu, F. Spener, G. van Meer, M. J. O. Wakelam, and E. A. Dennis. 2009. Update of the LIPID MAPS comprehensive classification system for lipids. J. Lipid Res. 50(Supplement): S9–14.

Fahy, E., M. Sud, D. Cotter, and S. Subramaniam. 2007. LIPID MAPS online tools for lipid research. *Nucl. Acids Res.* 35(suppl_2): W606–612.

Farcomeni, A. 2008. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods Med. Res.* 17(4): 347–388.

Fernandis, A. Z., and M. R. Wenk. 2009. Lipid-based biomarkers for cancer. J. Chromatogr*. B* 877(26): 2830–2835.

Fisher Box, J. 1987. Guinness, Gosset, Fisher, and Small Samples. Stat. Sci. 2(1): 45–52.

Gopalacharyulu, P. V., E. Lindfors, C. Bounsaythip, T. Kivioja, L. Yetukuri, J. Hollmén, and M. Orešič. 2005. Data integration and visualization system for enabling conceptual biology. *Bioinformatics* 21(Supplement 1): i177–185.

Gray, S., and A. Vidal-Puig. 2007. Adipose Tissue Expandability in the Maintenance of Metabolic Homeostasis. *Nutr. Rev.* 65(s1): S7–S12.

Griffiths, W. J. 2003. Tandem mass spectrometry in the study of fatty acids, bile acids, and steroids. *Mass Spectrom. Rev.* 22(2): 81–152.

Gupta, S., M. R. Maurya, D. L. Stephens, E. A. Dennis, and S. Subramaniam. 2009. An Integrated Model of Eicosanoid Metabolism and Signaling Based on Lipidomics Flux Analysis. *Biophys. J.* 96(11): 4542–4551.

Hackl, H., M. Maurer, B. Mlecnik, J. Hartler, G. Stocker, D. Miranda-Saavedra, and Z. Trajanoski. 2004. GOLD.db: genomics of lipid-associated disorders database. *BMC Genomics* 5(1): 93–98.

Han, X., and R. W. Gross. 2001. Quantitative Analysis and Molecular Species Fingerprinting of Triacylglyceride Molecular Species Directly from Lipid Extracts of Biological Samples by Electrospray Ionization Tandem Mass Spectrometry. *Anal. Biochem.* 295(1): 88–100.

Han, X., and R. W. Gross. 2005a. Shotgun lipidomics: Electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples. *Mass Spec. Rev.* 24(3): 367–412.

Han X, Gross RW. 2005b. Shotgun lipidomics: multidimensional MS analysis of cellular lipidomes. *Expert Review of Proteomics* 2: 253–64

Hand, D. J., H. Mannila, and P. Smyth. 2001. *Principles of Data Mining*. MIT Press.

Haynes, C. A., J. C. Allegood, H. Park, and M. C. Sullards. 2009. Sphingolipidomics: Methods for the comprehensive analysis of sphingolipids. *J.Chromatogr. B* 877(26): 2696–2708.

Hermansson, M., A. Uphoff, R. Kakela, and P. Somerharju. 2005. Automated Quantitative Analysis of Complex Lipidomes by Liquid Chromatography/Mass Spectrometry. *Anal. Chem.* 77(7): 2166–2175.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* 24: 417–441.

Houjou, T., K. Yamatani, M. Imagawa, T. Shimizu, and R. Taguchi. 2005. A shotgun tandem mass spectrometric analysis of phospholipids with normal-phase and/or reverse-phase liquid chromatography/electrospray ionization mass spectrometry. *Rapid Comm. Mass Spectrom.* 19(5): 654–666.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proc. ECML-98, 10th European Conference on Machine Learning*: 137–142.

Jolliffe, I. T. 1986. *Principal Component Analysis*. Springer-Verlag, New York.

Joyce, A. R., and B. O. Palsson. 2006. The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell. Biol.* 7(3): 198–210.

Junker, B., C. Klukas, and F. Schreiber. 2006. VANTED: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* 7(1): 109.

Kaluzny, M., L. Duncan, M. Merritt, and D. Epps. 1985. Rapid separation of lipid classes in high yield and purity using bonded phase columns. *J. Lipid Res.* 26(1): 135–140.

Kanehisa, M., and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res.* 28(1): 27–30.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucl. Acids Res.* 32(Database issue): D277–280.

Katajamaa, M., J. Miettinen, and M. Orešič. 2006. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22(5): 634–636.

Katajamaa, M., and M. Orešič. 2007. Data processing for mass spectrometry-based metabolomics. *J. Chromatogr. A* 1158(1–2): 318–328.

Katajamaa, M., and M. Orešič. 2005. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics* 6(1): 179–190.

Keseler, I. M., J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp. 2005. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucl. Acids Res.* 33(suppl_1): D334–337.

Khaselev, N., and R. C. Murphy. 2000. Structural characterization of oxidized phospholipid products derived from arachidonate-containing plasmenyl glycerophosphocholine. *J. Lipid Res.* 41(3): 564–572.

Koutnikova, H., T.-A. Cock, M. Watanabe, S. M. Houten, M.-F. Champy, A. Dierich, and J. Auwerx. 2003. Compensation by the muscle limits the metabolic consequences of lipodystrophy in PPAR gamma hypomorphic mice. *Proc. Natl. Acad. Sci. USA* 100(24): 14457–62.

Krieger, C. J., P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. 2004. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucl. Acids Res.* 32(suppl_1): D438–442.

Leandro, C. C., P. Hancock, R. J. Fussell, and B. J. Keely. 2006. Comparison of ultra-performance liquid chromatography and high-performance liquid chromatography for the determination of priority pesticides in baby foods by tandem quadrupole mass spectrometry. *J. Chromatogr. A* 1103(1): 94–101.

Leavell, M. D., and J. A. Leary. 2006. Fatty Acid Analysis Tool (FAAT): An FT-ICR MS Lipid Analysis Algorithm. *Anal. Chem.* 78(15): 5497–5503.

Lesnefsky, E. J., M. S. K. Stoll, P. E. Minkler, and C. L. Hoppel. 2000. Separation and Quantitation of Phospholipids and Lysophospholipids by High-Performance Liquid Chromatography. *Anal. Biochem.* 285(2): 246–254.

Lu, Y., S. Hong, and C. Serhan. 2006. Lipid Mediator Informatics-Lipidomics: Novel Pathways in Mapping Resolution. *AAPS Journal* 8(2): E284–E297.

Lusis, A. J. 2000. Atherosclerosis. *Nature* 407(6801): 233–241.

Mann, H. B., and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18: 50–60.

Matthew Barker, W. R. 2003. Partial least squares for discrimination. *J. Chemometr.* 17(3): 166–173.

McHowat, J., J. H. Jones, and M. H. Creer. 1997. Gradient elution reversed-phase chromatographic isolation of individual glycerophospholipid molecular species. *J.Chromatogr. B* 702(1–2): 21–32.

Menendez, J. A., and R. Lupu. 2007. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat. Rev. Cancer* 7(10): 763–777.

Merrill, J., Alfred H., M. C. Sullards, J. C. Allegood, S. Kelly, and E. Wang. 2005. Sphingolipidomics: High-throughput, structure-specific, and quantitative analysis of sphingolipids by liquid chromatography tandem mass spectrometry. *Methods* 36(2): 207–224.

Michalec, C., M. Sulc, and J. Mestan. 1962. Analysis of Cholesteryl Esters and Triglycerides by Thin-Layer Chromatography. *Nature* 193: 63–64.

Miller, R. G. 1981. *Simultaneous statistical inference*. Springer Verlag, New York.

Mitchell, T. 1997. *Machine Learning*. McGraw Hill, New York.

Myoung Hee Lee, G. H. L., Jong Shin Yoo,. 2003. Analysis of ceramides in cosmetics by reversed-phase liquid chromatography/electrospray ionization mass spectrometry with collision-induced dissociation. *Rapid Comm. Mass Spectrom.* 17(1): 64–75.

Navas-Iglesias, N., A. Carrasco-Pancorbo, and L. Cuadros-Rodríguez. 2009. From lipids analysis towards lipidomics, a new challenge for the analytical chemistry of the 21st century. Part II: Analytical lipidomics. Trac-trend Anal. Chem. 28(14): 393–403.

Ogiso, H., T. Suzuki, and R. Taguchi. 2008. Development of a reverse-phase liquid chromatography electrospray ionization mass spectrometry method for lipidomics, improving detection of phosphatidic acid and phosphatidylserine. *Anal. Biochem.* 375(1): 124–131.

Orešič, M., V. A. Hänninen, and A. Vidal-Puig. 2008. Lipidomics: a new window to biomedical frontiers. *Trends Biotechnol.* 26(12): 647–652.

Panganamala, R. V., L. A. Horrocks, J. C. Geer, and D. G. Cornwell. 1971. Positions of double bonds in the monounsaturated alk-1-enyl groups from the plasmalogens of human heart and brain. *Chem. Phys. Lipids* 6(2): 97–102.

Pearson, K. 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. Royal Soc.* 187: 253–318.

Pietiläinen, K. H., M. Sysi-Aho, A. Rissanen, T. Seppänen-Laakso, H. Yki-Järvinen, J. Kaprio, and M. Orešič. 2007. Acquired Obesity Is Associated with Changes in the Serum Lipidomic Profile Independent of Genetic Effects- A Monozygotic Twin Study. *PLoS ONE* 2(2): e218.

Pulfer, M., and R. C. Murphy. 2003. Electrospray mass spectrometry of phospholipids. *Mass Spectrom. Rev.* 22(5): 332–364.

Rosen, E. D., P. Sarraf, A. E. Troy, G. Bradwin, K. Moore, D. S. Milstone, B. M. Spiegelman, and R. M. Mortensen. 1999. PPAR[gamma] Is Required for the Differentiation of Adipose Tissue In Vivo and In Vitro. *Mol. Cell* 4(4): 611–617.

Schwudke, D., J. Oegema, L. Burton, E. Entchev, J. T. Hannich, C. S. Ejsing, T. Kurzchalia, and A. Shevchenko. 2005. Lipid Profiling by Multiple Precursor and Neutral Loss Scanning Driven by the Data-Dependent Acquisition. Anal.Chem. 78(2): 585–595.

Shi, Y., and P. Burn. 2004. Lipid metabolic enzymes: emerging drug targets for the treatment of obesity. *Nat. Rev. Drug Discov.* 3(8): 695–710.

Snedecor, G. W., and W. G. Cochran. 1989. Statistical Methods, Iowa State University Press.

Snyder, F. 1999. The ether lipid trail: a historical perspective. *Biochimica et Biophysica Acta (BBA) – Molecular and Cell Biology of Lipids* 1436(3): 265–278.

Song, H., F.-F. Hsu, J. Ladenson, and J. Turk. 2007. Algorithm for Processing Raw Mass Spectrometric Data to Identify and Quantitate Complex Lipid Molecular Species in Mixtures by Data-Dependent Scanning and Fragment Ion Database Searching. *J. Am. Soc. Mass Spectrom.* 18(10): 1848–1858.

Spearman, C. 1904. The proof and measurement of association between two things. Amer. J*. Psychol.* 15: 72–101.

Spiegelman, B. M. 1998. PPAR-gamma: adipogenic regulator and thiazolidinedione receptor. *Diabetes* 47(4): 507–514.

Ståhlman, M., C. S. Ejsing, K. Tarasov, J. Perman, J. Borén, and K. Ekroos. 2009. High-throughput shotgun lipidomics by quadrupole time-of-flight mass spectrometry. *J. Chromatogr. B* 877(26): 2664–2672.

Sud, M., E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, Jr, R. C. Murphy, C. R. H. Raetz, D. W. Russell, and S. Subramaniam. 2007. LMSD: LIPID MAPS structure database. *Nucl. Acids Res.* 35(suppl_1): D527–532.

Taskinen, M. R., T. Kuusi, E. Helve, E. A. Nikkila, and H. Yki-Järvinen. 1988. Insulin therapy induces antiatherogenic changes of serum lipoproteins in noninsulin-dependent diabetes. *Arteriosclerosis* 8(2): 168–177.

Thomas M.C., T.W. Mitchell, D.G. Harman, J.M. Deeley, J.R. Nealon, and S.J. Blanksby 2007. Ozone-Induced Dissociation: Elucidation of Double Bond Position within Mass-Selected Lipid Ions. *Analytical Chemistry* 80: 303–11

van Meer, G. 2005. Cellular lipidomics. *EMBO J.* 24(18): 3159–3165.

Vance, D. E., and J. E. Vance. 2008. *Biochemistry of lipids, lipoproteins and membranes.* Elsevier, Hungary.

Vapnik, V. 1995. *The nature of statistical learning theory*. Springer-Verlag, New York.

Wang, C., H. Kong, Y. Guan, J. Yang, J. Gu, S. Yang, and G. Xu. 2005. Plasma Phospholipid Metabolic Profiling and Biomarkers of Type 2 Diabetes Mellitus Based on High-Performance Liq uid Chromatography/Electrospray Mass Spectrometry and Multivariate Statistical Analysis. *Anal. Chem.* 77(13): 4108–4116.

Watson, A. D. 2006. Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders. Lipidomics: a global approach to lipid analysis in biological systems. *J. Lipid Res.* 47(10): 2101–2111.

Weininger, D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inform. Comput. Sci. 28(1): 31–36.

Wenk, M. R. 2005. The emerging field of lipidomics. Nat. Rev. Drug Discov. 4(7): 594–610.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bull.* 1: 80–83.

Wilson, I. D., J. K. Nicholson, J. Castro-Perez, J. H. Granger, K. A. Johnson, B. W. Smith, and R. S. Plumb. 2005. High Resolution Ultra Performance Liquid Chromatography Coupled to oa-TOF Mass Spectrometry as a Tool for Differential Metabolic Pathway Profiling in Functional Genomic Studies. *J.Proteome Res.* 4(2): 591–598.

Wold, S., K. Esbensen, and P. Geladi. 1987. Principal Component analysis. *Chemometr. Intell. Lab. Syst.* 2: 37–52.

Yetukuri, L., K. Ekroos, A. Vidal-Puig, and M. Orešič. 2008. Informatics and computational strategies for the study of lipids. *Mol. BioSyst.* 4(2): 121–127.

Zemski Berry, K. A., and R. C. Murphy. 2004. Electrospray ionization tandem mass spectrometry of glycerophosphoethanolamine plasmalogen phospholipids. *J. Am. Soc. Mass Spectrom.* 15(10): 1499–1508.

Publications II and VI are not included in the PDF version. Please order the printed version to get the complete publication (http://www.vtt.fi/publications/index.jsp).

Series title, number and
report code of publication

VTT Publications 741
VTT-PUBS-741

Author(s)
Laxmana Rao Yetukuri

Title

Bioinformatics approaches for the analysis of lipidomics data

Abstract

The potential impact of lipid research has been increasingly realised both in disease treatment and prevention. Recent advances in soft ionization mass spectrometry (MS) such as electrospray ionization (ESI) have permitted parallel monitoring of several hundreds of lipids in a single experiment and thus facilitated lipidomics level studies. These advances, however, pose a greater challenge for bioinformaticians to handle massive amounts of information-rich MS data from modern analytical instruments in order to understand complex functions of lipids**.** The main aims of this thesis were to 1) develop bioinformatics approaches for lipid identification based on ultra performance liquid chromatography coupled to mass spectrometry (UPLC/MS) data, 2) predict the functional annotations for unidentified lipids, 3) understand the *omics* data in the context of pathways and 4) apply existing chemometric methods for exploratory data analysis as well as biomarker discovery.

A bioinformatics strategy for the construction of lipid database for major classes of lipids is presented using simplified molecular input line entry system (SMILES) approach. The database was annotated with relevant information such as lipid names including short names, SMILES information, scores, molecular weight, monoisotopic mass, and isotope distribution. The database was tailored for UPLC/MS experiments by incorporating the information such as retention time range, adduct information and main fragments to screen for the potential lipids. This database information facilitated building experimental tandem mass spectrometry libraries for different biological tissues.

Non-targeted metabolomics screening is often get plagued by the presence of unknown peaks and thus present an additional challenge for data interpretation. Multiple supervised classification methods were employed and compared for the functional prediction of class labels for unidentified lipids to facilitate exploratory analysis further as well as ease the identification process. As lipidomics goes beyond complete characterization of lipids, new strategies were developed to understand lipids in the context of pathways and thereby providing insights for the phenotype characterization. Chemometric methods such as principal component analysis (PCA) and partial least squares and discriminant analysis (PLS/DA) were utilised for exploratory analysis as well as biomarker discovery in the context of different disease phenotypes.

ISBN
978-951-38-7402-5 (soft back ed.)
978-951-38-7403-2 (URL: http://www.vtt.fi/publications/index.jsp)

Series title and ISSN	Project number
VTT Publications 1235-0621 (soft back ed.) 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)	

Date	Language	Pages
May 2010	English	75 p. + app. 106 p.

Name of project	Commissioned by

Keywords
Lipids, Lipidomics, Bioinformatics, Lipid pathways, High density lipoproteins, k-nearest neighbours, Liquid chromatography/mass spectrometry, Principal component analysis, Partial least squares and discriminant analysis, Obesity, Support vector machines, LipidDB

Publisher
VTT Technical Research Centre of Finland
P.O. Box 1000, FI-02044 VTT, Finland
Phone internat. +358 20 722 4520
Fax +358 20 722 4374

Lipid research has recently gained increased attention due to their involvement in numerous diseases including diabetes, obesity, atherosclerosis and Alzheimer's disease. Modern technological advancements in mass spectrometry allowed us to study several hundreds of lipids at a time. This level of investigation not only provides insights into the specific roles of lipid molecular species in the health and disease, but also facilitates in identifying potential biomarkers for prevention and treatment of human health. These developments, however, comes with a set of informatics challenges in terms of handling the data. This thesis mainly deals with some of the challenges associated with this kind of lipid research in the context of liquid chromatography/mass spectrometry methods. The presented informatics methods herein assist in identification of molecular species, their functional class prediction, and data interpretation in biological pathway context and data analysis

ISBN 978-951-38-7402-5 (soft back ed.)
ISSN 1235-0621 (soft back ed.)

ISBN 978-951-38-7403-2 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN 1455-0849 (URL: http://www.vtt.fi/publications/index.jsp)