University of Alberta

Developing bioinformatics tools for metabolomics

by

Jianguo Xia

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Physiology, Cell and Developmental Biology

Department of Biological Sciences

Fall 2011

Edmonton, Alberta

Abstract

Metabolomics aims to study all small-molecule compounds (i.e. metabolites) in cells, tissues, or biofluids. These compounds provide a functional readout of the physiological, developmental, and pathological state of a biological system. The field of metabolomics has expanded rapidly over the last few years with increasing applications to disease diagnosis, drug toxicity screening, nutritional studies and many other life sciences. However, significant challenges remain in both collecting and understanding metabolomic data. The central objective of my thesis project is to develop novel bioinformatic tools to address some of the key computational challenges in metabolomic studies. In particular, my research is focused on three areas: (i) compound identification from complex biofluids, (ii) processing and statistical analysis of metabolomic data, and (iii) functional interpretation of metabolomic data.

In addressing these issues I have developed a number of efficient and user-friendly software tools, including MetaboMiner, MetaboAnalyst, MSEA and MetPA. Each of these software packages has required the development of novel algorithms, novel interfaces or the implementation of novel analytical concepts. MetaboMiner (http://wishart.biology.ualberta.ca/metabominer) is a standalone Java application for compound identification from 2D NMR spectra of complex biofluids. Based on a novel adaptive search algorithm and specially constructed spectral libraries, MetaboMiner is able to automatically identify ~80% of metabolites from good quality NMR spectra. MetaboAnalyst

(http://www.metaboanalyst.ca) is a web-based pipeline for metabolomic data processing, normalization, and statistical analysis. This application is based on a novel framework that combines the statistical and visualization power offered by R (http://www.r-project.org) with an enhanced graphical user interface enabled by Java Server Faces technology. It is currently the most comprehensive and popular data analysis web service in metabolomics. MSEA or metabolite set enrichment analysis (http://www.msea.ca) represents a novel application of the gene set enrichment analysis technique to metabolomics. In particular, MSEA is a web application for the identification of biologically meaningful patterns through enrichment analysis of quantitative metabolomic data. To create MSEA, I assembled a unique database of ~6300 groups of biologically related metabolites with association data on diseases, pathways, genetic traits, and cellular or organ localization. MetPA (http://metpa.metabolomics.ca) is a web-based tool for metabolic pathway analysis. It integrates functional enrichment analysis and pathway topology analysis through a novel Google-map style network visualization system. MetPA currently supports the analysis of ~1200 KEGG metabolic pathways for 15 model organisms.

These four software tools have become quite popular within the metabolomics community. Together, they offer a comprehensive bioinformatics toolkit that implements all the necessary steps to assist researchers in moving from raw data analysis to understanding relevant biology.

Acknowledgements

There are so many people whom I would like to thank that contributed to making my thesis research so enjoyable and meaningful.

Firstly, I would like to thank my supervisor, Dr. David S. Wishart. I feel immensely grateful for his openness, trust, and generosity in supporting my research during the past four years. He has provided me with many helpful suggestions and constant encouragement during the course of my research program. He has introduced me to the life of being a scientist and has motivated me to pursue intellectual and professional independence along this path. Without his guidance and support, I would not have been able to succeed in my graduate studies.

Dr. Wishart's lab has provided an excellent interdisciplinary environment for my bioinformatics research. I could easily talk to the people in the biofluids team on one side and to the IT people on the other side during my project development. Therefore, I would like to thank Peter Tang and Nelson Young, members of the IT support staff, who were always there sharing their ideas during my early days of software programming. I would also like to thank David Hau and Nick Psychogios, members of the biofluids group, who offered thoughtful opinions on how metabolomic data should be presented and accessed from a bench biologist's perspective. I would also like to thank Rupa Mandal, the lab's metabolomics manager, who has been serving as a diligent tester during my

software validation and deployment phase. I also appreciate the valuable advice from Roman Eisner and Dr. Russ Greiner regarding machine learning techniques.

I would like to thank my wife, Cindy, and Mia, my newborn daughter, for being part of my family and for being such a tremendous source of joy and motivation. They have given me real insight into what is most important in life. Much thanks to my in-laws, who have helped a great deal during the first several months after Mia was born.

Finally, I would like to thank my generous funding source - the Alberta Ingenuity Fund (AIF), now part of Alberta Innovates - Technology Futures, for financial support.

Table of Contents

List of Figures

List of Tables

List of Terms and Abbreviations

Terms Relevant to Metabolomics and Biology

APCI	atmospheric pressure chemical ionization
CID	collision-induced dissociation
CSF	cerebrospinal fluid
Da	Dalton, or atomic mass unit
DSS	2, 2-Dimethyl-2-silapentane-5-sulfonic acid (NMR reference standard)
ESI	electrospray ionization
EST	expressed sequence tags
EIC	extracted ion chromatograms
FID	free induction decay
GWAS	genome-wide association studies
HPLC	high performance liquid chromatography
HSQC	heteronuclear single quantum correlation spectroscopy
IEM	inborn errors of metabolism
MAS	magic angle spinning
MRS	magnetic resonance spectroscopy
MS	mass spectrometry
MRM	multiple reaction monitoring
NMR	nuclear magnetic resonance
ppm	parts per million
QTL	quantitative trait loci

SAGE	serial analysis of gene expression
TCA	tricarboxylic acid
TOCSY	total correlation spectroscopy
UPLC	ultra-high performance liquid chromatography
GC-MS	gas chromatography mass spectrometry
LC-MS	liquid chromatography mass spectrometry

Terms Relevant in Bioinformatics, Statistics, and Computing Science

Ajax	asynchronous JavaScript with XML
ANOVA	analysis of variance
BLAST	basic local alignment search tool
BLOSUM	block substitution matrix
ES	effect size
FDR	false discovery rate
FWER	family-wise rrror rate
GO	gene ontology
GSEA	gene set enrichment analysis
GUI	graphic user interface
JSF	java server faces
KEGG	Kyoto encyclopedia of genes and genomes
LOOCV	leave-one-out cross validation
MCMC	Markov chain Monte Carlo

NP	Nondeterministic polynomial time
ORA	over representation analysis
PAM	point accepted mutations
PCA	principal component analysis
PLS-DA	partial least squares – discriminant analysis
PNG	portable network graphics
RF	random forests
ROC	receiver operating characteristic
SAM	significance analysis of microarrays
SPIA	signaling pathway impact analysis
SVD	singular value decomposition
SVM	support vector machine
VIP	variable importance in projection

Chapter 1

General Introduction

Metabolomics is a relatively new member of the "omics" family. It is mainly concerned with comprehensive analysis of all small-molecular compounds (i.e. metabolites) found in a biological system such as cells, tissues or biofluids (1). The field of metabolomics has grown rapidly in recent years. This growth has been driven primarily by advances in analytical technologies such as high-resolution nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). It has also been facilitated by developments in computer-aided pattern recognition and bioinformatics (2-5). Metabolomics has become an increasingly important research tool in a wide range of life science disciplines including clinical chemistry, drug toxicity screening, nutritional research, environmental monitoring and many other fields (6-9). The rapid growth of metabolomics, coupled with its widespread applications in many life science disciplines is what led me to pursue my PhD research in this particular area. However to gain a real appreciation of both the strengths and weaknesses in metabolomics and to provide a clearer justification for my particular choice of research activities in metabolomics, it is important to have a more detailed understanding of this field.

This introductory chapter is intended to provide the reader with both the background and motivation for my research program. It is organized as follows: after a brief summary of the history of metabolomics and its current applications, I will provide a review of the three fundamental components of metabolomics – the metabolome, the analytical technologies, and bioinformatics, with particular focus given to the latter component as it is most relevant to my research. This

chapter will conclude with a description of my research objectives and an outline of this thesis.

1.1 A Brief History of Metabolomics

The practice of using bodily fluids for the study of disease can be traced back as far as 500 BC when ancient Greek and ancient Chinese physicians made medical diagnoses by assessing the color and smell of a patient's urine. The quantitative study of metabolites only became possible some 2400 years later with the development of chemical tests for key metabolites by early clinical chemists such as Archibald Garrod (1857–1936). In the 1960s and 1970s, researchers began to use chromatographic separation techniques and MS to investigate a relatively small number of metabolites in human blood and urine (10,11). In the late 1970s and 1980s, NMR-based approaches to monitoring and measuring isotopically labeled metabolites in biofluids, cells and tissues started to appear (12-14). The idea of measuring or comparing dozens of unlabeled metabolites simultaneously only started to emerge in the mid 1990s. This approach, also known as metabo***n***omics, was mainly developed by Dr. Jeremy Nicholson and his colleagues at the Imperial College, UK (15,16). In the late 1990s, papers describing NMR-based quantitative metabolomics started to appear (17-19), however these kinds of multi-metabolite studies were not formally labeled "metabolomics" studies. In fact, the term "metabolome" is a relatively recent invention being first suggested by Dr. Stephen Oliver (University of Manchester, UK) in 1998 (20). Soon afterwards, Dr. Oliver Fiehn (Max Plank Institute,

Germany) proposed the term "metabolomics" for the systematic study of metabolome changes in a biological system (1). With rapid advances in analytical NMR and MS technologies and the growing emphasis on high-throughput measurements of genes, proteins and metabolites, the field of metabolomics has since entered a period of active exploration, experimentation and expansion (8). In 2007, the Human Metabolome Project led by Dr. David S. Wishart (University of Alberta, Canada) published the first draft of the human metabolome (21).

There are two general approaches in current metabolomics studies – chemometric (or non-targeted) approaches and quantitative (or targeted) approaches (2). Chemometric approaches do not attempt to identify the compounds from the NMR or MS spectra of a complex mixture (such as a biofluid or tissue extract). Instead, they use the unannotated peaks or binned spectra combined with multivariate statistics to find important "features" or "peaks" that differentiate one sample cohort from another (i.e. disease from control). These important features may or may not be identified in the subsequent analysis steps. In contrast, quantitative metabolomics aims to formally identify and quantify all detectable metabolites from the NMR or MS spectra, *prior to subsequent data analysis*. Compound identification and quantification is usually achieved by comparing the spectra of the biological samples to a set of chemical standards or a spectral reference library.

Chemometric approaches were widely used during the early days of metabolomics, when compound identification was severely hampered by the lack of spectral databases and appropriate software support. However, without

identifying or quantifying which compounds have changed or by how much, it is difficult to identify useful biomarkers, to understand the pathways involved or to infer the mechanisms of action. With the availability of several comprehensive metabolomics databases and user-friendly bioinformatics tools (21-25), compound identification has become much easier and quantitative metabolomics is increasingly performed in today's metabolomics laboratories. In addition, researchers, particularly those involved in MS-base metabolomics, often combine both identified compounds and unknown peaks in their analyses, blurring the boundaries between these two approaches.

1.2 Applications of Metabolomics

Metabolomics is generally regarded as the end point of the "omics cascade". This is because metabolites are the final products of complex biological events governed by genetic, epigenetic, enzymatic and environmental changes that occur both within and outside the organism. Therefore metabolomics provides a wealth of information on the biochemical status of the cells, tissues or organisms, which is complementary to, yet distinct from, that generated by genomic and proteomic approaches. Metabolomics is particularly useful for understanding how a disturbance of a given biological system can be connected to mechanisms or processes on the cellular and molecular level.

Metabolomics has the potential to contribute significantly to both basic and clinical research. Recent advancements in the analytical techniques and the available bioinformatics resources have greatly expanded the number of

metabolomics applications and the type of applications for which metabolomics can be used. In particular, metabolomics now plays an increasingly important role in functional genomics, disease diagnosis, toxicogenomics, nutrition and nutrigenomics, systems biology, environmental monitoring and even cellular imaging (26-32).

1.2.1 Metabolomics in Functional Genomics

Functional genomics aims to elucidate the functions of previously unknown or uncharacterized genes or gene products in a high throughput manner (33). Traditional approaches have centered largely on expression studies of genes (transcriptomics) and proteins (proteomics). As perhaps the best indicator of an organism's phenotype, metabolomics has proven to be a powerful, complementary tool for elucidating the function of the unknown and novel genes.

In microbial studies, metabolomics has been used to reveal the phenotype of silent mutations through a technique known as the "Functional ANalysis by Co-responses in Yeast (FANCY) (26). FANCY compares the metabolic profiles of yeast strains deleted for known genes with metabolic profiles of strains deleted for unstudied genes that produce no overt phenotype. The underlying assumption is that the "unknown" genes that produce identical metabolic profiles (co-responses) will possess similar functions as the known genes (guilty-by-association). In plant functional genomics, metabolomics has been applied to quantitative trait loci (QTL) analysis for the elucidation of molecular networks (34) and for crop improvement (35). In this context, the metabolic profiles

(instead of phenotypes) are used as the quantitative traits of interest (m-traits), and these m-traits are used to test any linkages with genetic polymorphisms. In human studies, two recent large scale genome-wide associations studies (GWAS) have revealed distinctive "genetically determined metabotypes" (36,37). In both studies, the authors measured hundreds of endogenous serum metabolite concentrations and calculated their association with SNPs sequenced from the same cohort. They found that most of the significant genetic variants identified were located in or near genes that code for enzymes or solute carriers. Individuals with different genotypes in these genes were found to have significantly different metabolic capacities with respect to the synthesis of some of the associated metabolites.

1.2.2 Metabolomics for the Study of Diseases

Small molecules play key roles in a number of cellular processes such as intracellular signalling, energy transfer and cell-to-cell communication. As a result, changes in metabolite concentrations can serve as very good indicators of perturbations to these processes. Furthermore, the metabolome typically generates an amplified response to any changes in either the proteome or transcriptome. Indeed, small changes in enzyme concentrations can often have a manifold larger impact on metabolite concentrations (38,39). Consequently, metabolomics is generally thought to have a greater potential for the monitoring and early diagnosis of disease than transcriptomics or proteomics (40).

One of the first applications of metabolomics towards disease diagnosis was in the screening of inborn errors of metabolism (IEM) based on GC-MS (41).

GC-MS and MS-MS methods are now able to test for over 130 IEMs using blood or urine samples from newborn babies (27). Metabolomics has also proven to be a valuable tool in the study of type II diabetes (42,43), the diagnosis of diabetic nephropathy (44), as well as the monitoring of other diabetic complications (45). In cardiovascular disease research, metabolomics has been used to identify biomarkers of acute myocardial ischemia (46) and to predict cardiovascular events (47). More recent applications include the use of metabolomics to study obesity (48), aging (49) and cancer (50).

1.2.3 Metabolomics for Drug Toxicity Assessment and Environmental Monitoring

Early detection of drug toxicity has been one of the driving forces behind the early adoption of metabolomics technologies within the pharmaceutical industry. One of the most noteworthy efforts is the COMET (Consortium for Metabonomic Toxicology) project formed between five pharmaceutical companies and the Imperial College London (29). Using rat and mouse models, the project tested a total of 147 common toxins and collected around 35,000 NMR spectra of rat and mouse biofluids. The information has been used to develop an expert system that is able to predict the organ most likely to be affected based on the blood or urine spectra collected after the administration of a novel drug or toxin (51). Metabolomics has also proven to be valuable in environmental toxicity assessment. For instance, metabolic profiling of earthworms has identified potential biomarkers of soil contamination (52). A recent metabolomics study has

shown that the metabolic signatures of marine mussels can also be used to monitor the impact of chemical exposure in ocean water systems (53).

1.2.4 Metabolomics in Food Sciences and Nutritional Studies

Foods of plant origin contain a large number of phytochemicals. These phytochemicals are transformed into various secondary metabolites after digestion and absorption. Some of these secondary food metabolites can further modulate metabolism and thereby influence health. For instance, a number of studies have shown that antioxidants such as polyphenols (found in tea or wine) or carotenoids (found in tomato juice) provide many health benefits and prevent various diseases including cardiovascular diseases, cancers, neurodegenerative diseases, diabetes, osteoporosis, *etc.* (54,55). Metabolomics is now considered a key tool in understanding the biological mechanisms of phytochemicals and in the development of functional foods (30,56).

Over the past decade a number of metabolomic studies have been conducted to investigate the metabolic responses following dietary interventions in both animal models and human populations (57-61). Metabolomics is now recognized as an essential tool for monitoring dietary interventions and conducting personalized nutritional research (62). The recent introduction of the Nutritional Metabolomics Database (http://wiki.nugo.org/index.php/Nutritional_Metabolomics_Database) will greatly facilitate research in this area.

1.3 Defining the Metabolome

The metabolome consists a wide spectrum of low-molecular-weight compounds, including sugars, amino acids, lipids, nucleotides, vitamins and cofactors. These compounds have very diverse chemical and physical properties (i.e. molecular weight, polarity, solubility, or volatility) and occur at different concentrations that can vary over nine orders of magnitude from picomolar to millimolar levels. The metabolome was originally defined in the context of metabolic control analysis (MCA) as the set of all endogenous, low molecular weight compounds synthesized by an organism (20). It was later redefined to refer to all the small compounds that can be measured within a biological system (1). This broader definition implicitly includes both endogenous and exogenous compounds such as those derived from foods and xenobiotics (i.e. drugs, toxins, pollutants). In addition, as technology improves and detection limits decrease, it is likely that many more metabolites will be identified, making the potential size of any given organism's metabolome practically infinite.

Nevertheless, it is possible to give some estimates of the size of different metabolomes based on work from genome-scale reconstructions of different organism's metabolic networks as well as information from public databases and literature. For instance, the total number of metabolites in the *E. coli* metabolome was estimated to be ~450 (63); the yeast metabolome was predicted to contain ~600 metabolites (64); while the total number of metabolites within a mammalian cell was estimated to be ~650 (65). However, these network reconstructions appear to seriously underestimate the true size of these metabolomes. For

instance, literature derived data indicates that the *E. coli* metabolome consists of over 1,000 metabolites (66), a similar approach has shown that the yeast metabolome contains well over 2,000 metabolites (67), while the Human Metabolome Project has identified over 8,000 endogenous metabolites in humans (including ~4,000 lipid species) (25). Plant metabolomes tend to be somewhat larger as over 5,000 metabolites were reported from the rice metabolome (68); and the total number of metabolites in the plant kingdom has been estimated to be ~200,000 (31).

While the size of the metabolome varies considerably between species, it is also important to note that even within the same species, the distribution of metabolites is subject to considerable spatial and temporal variability. As a result, for multi-cellular organisms it is more common to report the metabolome for specific organs or specific compartments. For example, the human cerebrospinal (CSF) metabolome contains ~1005 metabolites, the human serum/plasma metabolome contains ~4,600 metabolites, and the size of human urine metabolome is ~800 compounds (25,69,70). Even within a given organ or a given physiological compartment, the metabolome is still highly dynamic and context-dependent, varying according to the physiological, developmental, or pathological state of the organism (71). In the case of human subjects, factors like age, gender, diet, diurnal variations, exercise, or disease conditions can all exert noticeable effects on metabolite concentrations and compositions (72). Therefore, the metabolome is often described as a "state function" of an individual at a particular time point (9). Despite these variations, recent studies have demonstrated that an

invariant distinctive “metabolic phenotype” can be ascribed to specific individuals (73) or geographically dispersed populations (74).

1.4 Metabolomics Platforms

Unlike genomics or proteomics, where the data collection procedures have mostly consolidated and standardized into a few platforms and protocols, the chemical complexity and heterogeneity of metabolome makes it extremely difficult to measure all compounds using a single analytical platform. Current metabolomics studies are largely based on the use of NMR or GC/LC-MS to detect, identify, and quantify small molecule compounds from biological samples (2). For historical reasons, NMR has been more commonly applied to mammalian samples, while MS-based methods have been used more often in plants and microbial studies. Due to their complementary nature, these analytical methods are increasingly used in combination to provide a more comprehensive coverage of the metabolome.

In the following sections, I will provide a basic overview covering the main characteristics and features of NMR and GC/LC-MS based technologies in the context of metabolomics studies. It is important to note that other analytical techniques are also being used in metabolomics, such as infrared spectroscopy, immunodetection, and capillary electrophoresis with fluorescent detection. However, time and space prevent me from discussing all of these technologies. For a more comprehensive introduction to metabolomics technologies, please refer to two recent and very excellent review articles (3,75).

1.4.1 NMR Spectroscopy

The basic principle behind NMR measurement is that when a biological sample is put in a constant magnetic field, those nuclei that possess non-zero magnetic spins (such as ^{1}H and ^{13}C) will be either align (low energy) or oppose (high energy) the external magnetic field. Each nucleus or class of nuclei spins with a characteristic frequency or *resonance frequency*, determined by its nuclear composition, its chemical environment and the strength of the applied magnetic field. When one applies a short high-power pulse of radio frequency (RF) energy that is close to the resonance frequency of the nuclei of interest, a small proportion of those nuclei will absorb the RF energy and will be excited into a high-energy state. As the system returns to equilibrium, the absorbed energy is released as a burst of radio waves with slightly different frequencies known as the free induction decay (FID). Each of the frequencies corresponds to the characteristic resonance frequency of the different nuclei in the sample. After Fourier transformation, the signal is converted to a conventional NMR spectrum characterized by multiple peaks of varying position (corresponding to chemical shifts) and varying height (corresponding to the relative abundance of each nuclear type). Using this information it is possible to determine the chemical structure of pure substances or the chemical composition of liquid mixtures.

The most abundant and sensitive NMR-active nucleus is the hydrogen nucleus. As hydrogen is found in nearly every organic molecule, proton NMR is one of the most commonly used approaches to characterize organic molecules. Given that most metabolites are organic molecules, it is not surprising to learn

that one-dimensional (1D) proton NMR spectroscopy has been used widely in metabolomic studies involving human biofluids. NMR can also be used to study solid tissue samples using a technique called high-resolution magic angle spinning (MAS) NMR spectroscopy of intact tissue (76). Current detection limits for proton NMR spectroscopy are on the order of 1-5 μM in biofluid samples, with a typical acquisition time of ~10 minutes (2). NMR can also be used in a non-invasive manner to detect or measure metabolites. This involves using *in vivo* magnetic resonance spectroscopy (MRS) of intact organisms (77).

While most NMR-based metabolomics practiced today is based on using 1D proton spectra (which are quick and easy to collect), it is also possible to analyze biological mixtures using slightly more advanced NMR spectroscopic techniques. In particular, two-dimensional (2D) NMR offers a robust approach to resolving excessively overlapped spectra commonly encountered in the 1D proton NMR of complex biofluid mixtures. 2D NMR can also be used to elucidate the structure of novel compounds that have been isolated or purified from biological mixtures (78). There are many different types of 2D NMR experiments. Among them, ^{1}H-^{1}H total correlation spectroscopy (TOCSY) and ^{1}H-^{13}C heteronuclear single quantum correlation spectroscopy (HSQC) are commonly used in NMR-based metabolomics. The TOCSY experiment can help link clusters of peaks that are thought to belong to the same compound, but it also reveals linkages to peak clusters that 1D NMR cannot resolve. This additional information can be very useful for compound assignments. The HSQC experiment provides increased resolution by utilizing the greater ^{13}C chemical shift dispersion on one axis of the

2D spectrum, which significantly reduces peak overlap and therefore increases metabolite specificity.

There are several desirable features associated with NMR-based metabolomics. The most attractive feature is that multiple small molecule metabolites can be measured simultaneously without prior separation, which greatly simplifies the sample preparation requirements. The other important feature is that NMR spectra are highly reproducible, and samples analyzed from one spectrometer will generate near-identical results to those measured on other types of spectrometers (79). These features have made NMR spectroscopy a platform of choice for many large-scale high-throughput collaborative metabolomics projects (51,80). A major drawback is that NMR is a relatively insensitive technique as only medium to high abundance metabolites can be detected with this approach. The limited coverage of the metabolome has made data interpretation very difficult.

1.4.2 MS-based Methods

In metabolomics studies, mass spectrometry (MS) is usually coupled with a chromatographic technique such as gas chromatography (GC) or liquid chromatography (LC) to form hyphenated GC-MS or LC-MS analytical platforms. With MS-based metabolomics compounds are first separated either in the gas (GC) or solution phase (LC), and subsequently ionized, detected and then sorted according to their mass-to-charge (m/z) ratio, which can be used to identify the metabolites.

GC-MS offers a very high degree of chromatographic resolution and reproducibility. GC-MS is most suitable for low-molecular weight (< 500 Da), volatile, and thermally-stable compounds such as sugars, fatty acids, and amino acids. For large and polar compounds, chemical derivatization is usually required to improve their volatility and thermal stability before analysis. The most commonly used ionization technique in GC-MS is electron impact (EI) ionization which is very robust and reproducible. The characteristic mass spectral fragmentation patterns produced for common metabolites can be used to build a spectral library that can be used to compare spectra from other samples and to accurately identify metabolites from mixtures.

Compared to GC-MS, LC-MS methods typically have somewhat lower chromatographic resolution and reproducibility. However, LC-MS techniques can typically access a much broader mass range (100-2000 Da) because volatilization or derivatization is not necessary. LC-MS is also a better choice for separating and identifying polar and non-volatile compounds. Electrospray ionization (ESI) and atmospheric pressure chemical ionization (APCI) are the two most common ionization methods in LC-MS (81). A major concern with both ESI and APCI is ion suppression in which a compound suppresses (or enhances) the ionization of a co-eluting compound. Therefore, for complex samples, high quality separations are necessary in order obtain reliable LC-MS results. This is usually achieved by using multidimensional HPLC, capillary HPLC, or ultra-performance liquid chromatography (UPLC). Both ESI and APCI techniques will generate a molecular ion whose mass can be searched against a database of known

metabolites for possible identification. LC-MS systems can also be extended to perform tandem MS (MS/MS) or MS^n. Tandem mass spectrometry uses collision-induced dissociation (CID) of precursor or parent ions to produce reproducible fragment patterns for the identification or structural elucidation of the metabolite of interests. Recent advances in LC-MRM (multiple reaction monitoring)-MS can schedule up to 1,000 MRMs per run, making it a sensitive and robust platform for high-throughput quantitative or semi-quantitative metabolomics studies (82,83).

MS-based methods are widely used in metabolomics research. They are in general more sensitive than NMR-based approaches, and can usually detect metabolites at a concentration at least two orders of magnitude below that of NMR (9). Current detection limits for MS-based approaches are of the order of 100 nM, allowing the detection of ~1,000 metabolites, with typical acquisition times of ~ 30 minutes (81). However, compared to NMR-based approaches, MS-based approaches can sometimes be more time-consuming, especially if metabolite identification and quantification need to be performed. MS techniques also suffer from problems of reproducibility and require compound-specific calibration curves to perform any kind of absolute quantification. The introduction of selective isotope labeling for MS-based metabolomics should help resolve some of these limitations (84).

Whether metabolomic data is collected via LC-MS, GC-MS or NMR, it eventually has to be collated and analyzed. This aspect of data analysis represents one of the greatest challenges in metabolomics as it often requires the intelligent use of a full suite of bioinformatic tools. In other words, bioinformatics plays a

part in metabolomics that is almost equal to the analytical platforms used to collect the data.

1.5 Bioinformatics

1.5.1 Overview

The evolution of molecular biology from a bench-intensive, low-throughput science into a high-throughput, data-driven science has shifted the focus from data gathering towards data analysis. As a result, bioinformatics has become an integral part of almost every molecular biology experiment done today. Bioinformatics can be loosely defined as the application of statistics and computer science to the field of molecular biology. The field of bioinformatics has co-evolved closely with every advance in molecular biology.

The earliest applications of bioinformatics were focused on protein 3D structure determination from X-ray crystallographic data (85) and computing evolutionary trees from protein sequence data (86). These “niche” disciplines have since evolved into the field of structural bioinformatics and computational evolutionary biology. However, it wasn’t until the advent of the Human Genome Project (1989-2003) that the discipline of bioinformatics truly came into the spotlight. Thanks to the tools, databases and techniques developed by bioinformaticians, such as BLAST (87), GenBank (88), Phred-Phrap-Consed (89,90), it became possible to compare and annotate dozens of genomes from model organisms. Bioinformatics has continued to evolve as the focus has shifted

from analyzing sequence data towards analyzing microarray data, proteomic data and other kinds of omics data.

Omics data can be conceptually grouped into three major types - sequence data, expression data, or a combination of both. Sequence data includes gene or genome sequences, mRNA sequences and protein sequences. The bioinformatics tasks associated with sequence analysis usually involve similarity searches, sequence alignment, sequence annotation and structure prediction. These procedures typically involve text manipulation and string matching and are often performed using scripting languages such as Perl (http://www.perl.org) or modules from the BioPerl library (91). Expression data includes gene-expression data measured via microarray techniques, protein expression data measured from two-dimensional gel electrophoresis or isotopic labeling methods such as iTRAC or iCAT (92), as well as metabolomics data measured by a variety of different techniques. The bioinformatics tasks associated with expression data analyses typically include differential expression analysis, pattern discovery, classification and pathway analysis. These procedures often involve a significant amount of statistical analysis and machine learning. The programming language of choice is often R (http://www.r-project.org) and packages from the Bioconductor project (93). Expressed sequence tag (ESTs) data, serial analysis of gene expression (SAGE) data, or RNA-seq data generated by next-generation sequencing technology (94) can be considered as both sequence and expression data. The bioinformatics tasks for analyzing this type of data involve mapping and aligning

the sequence reads to the underlying genes/exons and then using the resulting sequence count data to perform expression analysis.

A detailed discussion of sequence analysis is outside the scope of this thesis. For a more comprehensive discussion of EST data analysis and annotation, please refer to my Master's thesis (95). The procedures for analyzing RNA-seq data are still rapidly evolving (96) and will not be discussed here. In keeping with the central subject of this thesis, I will review some of the important bioinformatic procedures associated with metabolomic expression data analysis.

1.5.2 Raw Data Processing in Metabolomics

The purpose of raw data processing in metabolomics is to convert raw spectral data generated by off-the-shelf NMR or MS instruments into a data format suitable for downstream statistical analysis or machine learning. The detailed procedures are highly dependent on the instruments used. In many cases, the low-level raw data are stored in proprietary formats and can only be processed using the software supplied by instrument vendors. Therefore, one of the first steps for metabolomic data processing is the conversion of raw (proprietary) data files from different machines or different vendors into a common and open-access format. Software supplied by instrument vendors usually contains scripts that can be used for this task. The development of data processing tools for metabolomics has been an active area of bioinformatics research in recent years. Many commercial and open source software tools are now available for raw data conversion and processing. A comprehensive list of tools can be found at the MS-Utils website

(http://www.ms-utils.org/wiki/pmwiki.php/Main/SoftwareList) and at the Fiehn laboratory website (http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment/). Some of these tools and procedures are discussed briefly below.

Processing NMR Data

Raw NMR data is usually saved as an FID file. The basic processing steps include fast Fourier transform, phase correction, baseline correction, and chemical shift referencing. Users can then either perform spectral binning or compound profiling (the latter requiring a comprehensive reference spectral library). The final output is typically a data table of spectral bins (corresponding to small chemical shift regions with integrated areas) or compound concentration values.

A variety of software tools are available for processing and interpreting NMR spectra for metabolomics. Certainly all three major NMR instrument manufacturers (Varian, Bruker and JEOL) produce instrument-specific software for processing and visualizing 1D proton spectra. However, these spectral processing tools are not ideal for identifying or profiling the compounds that might be found in complex mixtures – as is the case with NMR-based metabolomics. As a result several specialized software tools have been developed independently and introduced to the market. The Chenomx NMR Suite (Chenomx, Edmonton, Canada) is probably the most widely used metabolomics software tool for processing and profiling 1D proton NMR spectra. Another widely used commercial tool is AMIX (Bruker Biospin, Germany) which offers similar features. More recently, several open source software tools have also been

developed for processing NMR spectra for metabolomics studies, such as HiRes (97) and Automics (98). Compared to the commercial ones, these tools lack comprehensive spectra libraries to support compound identification and quantification. Instead, they tend to focus on spectral alignment, binning, and batch processing.

Software tools that support 2D NMR spectra processing and peak picking include NMRPipe (99), Sparky (100), AMIX, and rNMR (101). Peak alignment tools for 2D-NMR spectra have also been described (102). As yet there have been relatively few tools described that permit compound identification or quantification via 2D NMR spectroscopy, although this is changing (103-105).

Processing MS Data

Compared to NMR-based metabolomics, the field of MS-based metabololomics is somewhat more vibrant and diverse given the much larger variety of MS instrument vendors and MS instrument types. Indeed, there are almost as many MS data formats as instrument vendors. A necessary preliminary step with MS-based metabolomics is to convert the raw MS data from its instrument-specific or proprietary format into an open format such as netCDF or mzXML. Once the data has been converted to a readable format, users have a large number of both commercial and free software tools at their disposal including: AMDIS (106), AnalyzerPro (SpectralWorks, UK), Binbase (107), ChromaTOF (LECO, USA), MassFrontier (HighChem, Slovakia), MetAlign (108), Met-IDEA (109), MSFACTS (110), Tagfinder (111), XCMS (112), and MZmine (113).

The basic steps for MS data processing include noise filtering, feature detection, and spectral/peak alignment. The first two steps are usually integrated with each other. Different tools usually employ different algorithms to perform these steps. The popular open source XCMS package was developed primarily for processing LC-MS spectra for global, untargeted metabolite profiling (112). XCMS accepts raw LC-MS spectra in either mzXML or netCDF format. It first creates extracted ion base-peak chromatograms (EIBPC) by cutting each spectrum into slices in the m/z dimension. Each EIBPC is then processed by a matched filter using a second-derivative Gaussian as the model peak shape. After filtration, peaks are selected using a signal-to-noise cutoff. Peak intensities are calculated by integrating the unfiltered chromatogram between the peak boundaries (defined by zero-crossing points of the filtered chromatogram) without background subtraction. These peaks are subsequently binned by mass and grouped according to their retention time distribution estimated by a Gaussian kernel density. Some "well-behaved" peak groups are then selected to build a deviation profile to correct the retention time drifts of the original peak lists using a local regression fitting method, *loess*. The process of peak matching and retention time alignment can be performed iteratively by successively detecting more and more well-behaved peak groups to improve the overall alignment. The final output from XCMS is a peak intensity table with peaks identified by their retention time and m/z values for each spectrum.

Similar to NMR-based metabolomics tools, commercial MS data processing tools are usually shipped with comprehensive reference spectral

libraries to help metabolite identification. On the other hand, open-source tools for MS-based metabolomics are mainly designed for spectra processing and peak picking. Public MS spectra libraries have started to appear such as METLIN (23) and HMDB (25).

1.5.3 Metabolomic Data Normalization

Before performing any kind of statistical analysis with NMR or MS-derived metabolomic data, it is often necessary to perform some kind of data normalization. The purpose of data normalization is to reduce any systematic bias within the data and to improve overall data consistency so that meaningful biological comparisons can be made.

Systematic bias in experimental data is often a measurement error or bias that is unrelated to the biological changes of interest. Bias can be introduced during the many steps involved in the experimental setup such as patient selection, sample collection and preparation, spectral acquisition, and so on. Common sources of systematic bias include differences in sample quantity, dilution effects, technical variations due to imperfect instrument calibration or changes in measurement conditions, *etc.* A variety of methods have been developed to address the systematic biases encountered in metabolomics studies. Some commonly used methods include normalization by the sum (i.e. dividing by the sum of all concentrations or integrated spectral area), normalization by using internal controls, normalization by a physiological constants (i.e. dividing by *creatinine* concentration) or probabilistic quotient normalization (114).

Metabolites are present over a wide concentration range, which can differ by several orders of magnitude depending on the conditions to which the organism is exposed. However, it is important to remember that the magnitude of these concentrations or concentration changes is not necessarily proportional to their biological relevance. In many cases, the within-group metabolite concentration variances are higher in those groups where the mean concentration is also higher. Proper data transformation can usually improve the within-group data consistency and thereby increase the probability of detecting meaningful differences between groups. Metabolite concentrations are usually not normally (i.e. Gaussian) distributed, yet many statistical tests assume that data values are normally distributed. It is therefore important to perform certain data transformations to make metabolite data normally distributed. Commonly used methods include centering, autoscaling, pareto scaling, or log transformations (115). These procedures aim to reduce the impact of very large values and to make all metabolite concentrations (absolute or relative) more comparable. No consensus has been reached on the best data normalization procedures that will perform well on all types of metabolomic data sets. Different methods tend to emphasize different data features and each method has its own merits and drawbacks. The impact of these data transformation procedures have been discussed in detail by van den Berg *et al.* (115).

1.5.4 Metabolomic Data Analysis

Metabolomic data sets are usually high-dimensional, meaning that the number of variables (peaks, spectral bins or metabolites) is often very large, ranging from a

few dozen to hundreds or even thousands. Consequently it is difficult to manually examine each and every data point. This situation is very similar to what researchers have experienced with microarray data analysis. To deal with this kind of high dimensional data, a variety of statistical and machine learning approaches have been developed and tested. Below I will review some of the most commonly used approaches organized under three general categories: biomarker identification, pattern discovery, and class prediction.

Biomarker Identification

In large scale expression studies, it is often assumed that most of the observed metabolite or gene expression changes are a result of normal physiological variations (background noise) and that only a small proportion of them are actually associated with the experimental condition of interest. Identification of those "key" features is typically the first step toward finding useful biomarkers or understanding the biological processes involved in the condition under investigation. In microarray gene expression studies, this procedure for identifying these key genes is known as *differential analysis*. A variety of approaches for differential analysis have been developed for this task, with the majority of them being based on classical univariate approaches and their variants. In the following section, I will explain and review a number of univariate methods that can be used to identify important features from high dimensional data.

Univariate methods, by definition, consider each feature separately and treat that feature as an independent variable. Features are ranked according to

some measure related to their association to the conditions of interests. After this ranking step, the first few variables in the list are selected for further analysis. The majority of univariate methods used today are based on t-tests and ANOVA (analysis of variance) methods, or their non-parametric counterparts such as the Wilcoxon's rank sum tests and Kruskal-Wallis tests. Among these methods, the t-statistic used in the t-test is the simplest and most commonly used. The general form of the t-statistic is:

$$t = \bar{X} / S \qquad \textbf{(1)}$$

where $\bar{X}$ is the group difference and S is the (pooled) sample standard deviation (SD). For two groups of size n_1 and n_2 with variance of $\mathbf{S_1}$ and $\mathbf{S_2}$, the pooled variance for the classical t-test is defined as:

$$S_t = \sqrt{S_1^2 / n_1 + S_2^2 / n_2} \qquad \textbf{(2)}$$

The t-statistic can be considered as a measure of the signal-to-noise ratio in which the difference between the two sample means is the signal, and the noise is measured by the SD. The SD indicates the scatter or the dispersion of the sample values.

One major concern for high-dimensional data analysis is that when the sample size is small, parameters estimation tends to be unstable or prone to large errors. As a result, standard t-test and ANOVA methods often perform poorly with too many false positives. This issue has led to the development of several highly successful and widely used alternative tools such as Cyber-T (116), SAM

(117), and Limma (118). These approaches are generally called *regularized or moderated t-statistic* methods. The basic idea is to "borrow information" from other variables (i.e. genes) to improve the estimation of variance S. For example, SAM uses an *ad hoc* permutation approach to estimate a small "fudge factor" (c_0) to add to the standard SD for each gene expression value as shown here:

$$S_{SAM} = c_0 + S_t \qquad \textbf{(3)}$$

where S_t is defined by equation (2). In Limma, an empirical Bayes approach is used to borrow information from all the genes in the experiment, with

$$S_{\text{limma}} = \sqrt{(d_0 S_0^2 + d S_t^2)/(d_0 + d)} \qquad \textbf{(4)}$$

where d is the degree of freedom, S_t is defined by formula (2), d_0 and S_0 are estimated from the data using an empirical Bayes approach. These regularized t-statistic methods are generally considered superior to the standard t-test in microarray data analysis in the sense that they usually produce better results with less false positives.

Researchers in metabolomics face similar data dimensionality and data analysis issues as those in transcriptomics, but to a somewhat smaller extent. As the cost of a metabolomics experiment is approximately an order of magnitude lower than that of a microarray experiment, more samples are usually analyzed (>10 per group) in metabolomics studies than in microarray studies (3~10 per group). With more replicates, the variance estimate becomes fairly stable and accurate using standard t-test or ANOVA methods, which usually give a similar

performance compared to other more complicated methods for improved variance estimation (119).

Although approaches based on the *moderated t-statistic* contain some multivariate components in the sense that they use (or borrow) information from other variables, they do not take variable correlations or interactions into consideration. These methods are generally considered univariate approaches. In general, univariate methods are simple to use and results are easy to understand. They are widely used for exploratory data analysis. However, univariate approaches are considered suboptimal for high-dimensional data analysis in biology as they tend to ignore the correlations that are known to be present among genes or metabolites. In addition, the high-dimensional nature of the data inevitably leads to multiple tests which would inflate the false positives. As a result, multivariate methods which simultaneously take all variables into consideration are generally considered superior to simple univariate approaches.

However, the classical multivariate statistical methods - Hotelling's T^2 test and multivariate analysis of variance (MANOVA), which represent direct extensions of the t-test and ANOVA to their multivariate counterparts, have not gained much use in the expression analysis community. This is primarily because it becomes much more difficult to ascertain the nature of the underlying multivariate distributions with a small number of samples (*the curse of dimensionality*). In addition, these two multivariate tests simply tell the user whether a difference exists or not and require complicated follow-up analyses to know which features are significant. Nevertheless, some modified variants of

MANOVA have been proposed (120). Two other widely used multivariate approaches for analyzing high-dimensional data are based on dimensional reduction - principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). These two methods will be discussed in more detail, particularly as they relate to class discovery and class prediction, respectively.

Pattern Discovery

Identifying biologically interesting patterns has been one of the central activities in exploratory analysis of omics data. As omics technologies are generally considered hypothesis generating tools, researchers typically want to find patterns in the data that are not predicted by their current knowledge or pre-conceptions. In gene expression data analysis, some typical goals include the identification of groups of genes whose expression patterns are tightly related across samples (i.e. co-expression analysis); or to find unknown subgroups among samples (i.e. disease subtypes). Here I will focus on two approaches that are commonly used in metabolomic data analysis - dimensional reduction and clustering. As the human eye is very proficient at discerning patterns, both approaches are designed to provide excellent visualization support to facilitate the process.

Dimensional Reduction

The basic idea in dimensional reduction is to "summarize" a large number of variables into a small number of new components with minimal loss of information. One of the most widely used dimensional reduction methods is

principal component analysis (PCA), which aims to project or transform the data into a new coordinate system such that most data variance lies in the first few components. These components (also known as principal components or PCs) are constructed using linear combinations of the original variables. The results of a PCA are usually discussed in terms of *scores* and *loadings*. Scores are the values of the original data projected to the new coordinate system while loadings are the weights applied to original data during the projection process. Once the PCA transformation is complete, researchers then visually scan the PCA score plots to look for inherent groupings or patterns in the data. The most significant features contributing to the observed clusters can then be identified from the corresponding PCA loading plot.

PCA provides a convenient summary of data with respect to data variance. However, there is no guarantee that the direction of maximum variance will align with the direction of the biological changes of interest. For instance, the first principal component may simply reflect systematic error in the data, especially when data processing and normalization are not performed properly. In addition, PCA's reliance on users to identify patterns through visual inspection is subjective (and sometimes dangerous) as the human eye can often spot or "create" patterns even they are not there. Finally, as PCA is based on eigenvalue decomposition of a data covariance matrix or singular value decomposition (SVD) of a data matrix, it is quite sensitive to the presence of outliers and artifacts arising from data transformation procedures.

Clustering

Clustering analysis aims to reveal interesting patterns by directly generating or seeking natural clusters within the data. Clusters are subgroups in a data set that are more similar to each other than any other subgroup in the data set. One necessary component in clustering analysis is how to measure the distance between any two objects. Many distance measures have been used for clustering microarray data. Some common ones include Euclidean distance, Manhattan distance, Pearson's correlation, Spearman's rank correlation, and cosine-angle (121). The other critical component is the clustering algorithm, namely, how the clusters are constructed given a distance matrix. In bioinformatics, three types of cluster analysis are commonly used - hierarchical clustering, partitional clustering, and biclustering.

Hierarchical clustering is probably the most widely used clustering method in biology. Indeed, hierarchical clustering with heat maps has become almost routine for microarray gene expression data analysis, beginning with the first description of heat maps by Eisen et al. (122). Hierarchical clusters can be constructed through either an agglomerative or a divisive algorithm. For instance, the agglomerative approach begins with each sample being considered as a separate cluster and then proceeds to combine them until all samples belong to one cluster. The order by which each cluster is combined is determined by the distance among the clusters. The intercluster distance can be calculated with different measures such as single-linkage, complete linkage, average linkage, or Ward’s method (121). The net result of a hierarchical clustering process is a tree

of nested clusters usually accompanied with heat maps as illustrated in **Figure 1.1A**. As with most image- or graph-based methods, users need to decide at which level the clusters are most biologically meaningful. Another issue with hierarchical clustering is that once an assignment has been made, it cannot be changed in later stages if it is found to be non-optimal.

In contrast to hierarchical clustering, partitional clustering approaches attempt to directly decompose the data set into a user-specified number of disjoint clusters. K-means and self-organizing map (SOM) are two widely used partitional clustering methods. For instance, k-means clustering aims to partition the data into a set of *k* user-specified clusters such that the sum of squares from points to the assigned cluster center is minimized. This can be achieved by considering every possible partition of p data points into k groups and then selecting the one that yields the lowest within-group sum of squares. However, for a high number of instances, it is impossible to enumerate all possible partitions. As a result, the method is usually implemented in iterations by rearranging existing partitions until no improvement is seen. One issue associated with partitional clustering is how to determine the optimal number of clusters when there is no prior knowledge about the number of clusters that should be seen in the data. In practice, many researchers choose the initial cluster number based on results from PCA or hierarchical clustering.

Both hierarchical and partitional approaches fail to accommodate several important biological characteristics inherent in high dimensional “omics” expression data. For example, some genes or metabolites can be involved in more

than one active biological process, which means that clustering algorithms should ideally allow partial overlap among different clusters. More importantly, many genes or metabolites may not be involved in any of the active processes and mainly serve as "background noise" – especially if we assume that most observed expression changes are a result of normal physiological variations. Including this "background noise" in the clustering process will dilute the signal and obscure any kind of useful functional interpretation. It is also quite possible in biological systems that some active processes may only be "turned-on" under some conditions. Therefore the ideal clustering algorithm should focus only on these local "patches" of data that exhibit interesting patterns while leave the remaining data unclustered as illustrated in **Figure 1.1B**. Different clustering algorithms have been proposed over the last few years to perform this kind of biologically intelligent clustering (123). These approaches are generally referred to as biclustering or subspace clustering methods. The local "patches" are called biclusters as they are two-dimensional clusters - i.e. gene clusters that are only defined over an associated sample cluster. The algorithms for biclustering are very complex and will not be described here. In general, the problem of estimating a set of biclusters is considered to be NP-hard and a globally optimal solution is unlikely to be obtained for high-dimensional data. Despite these difficulties, a number of tools have been implemented for biclustering analysis of gene expression (124-126).

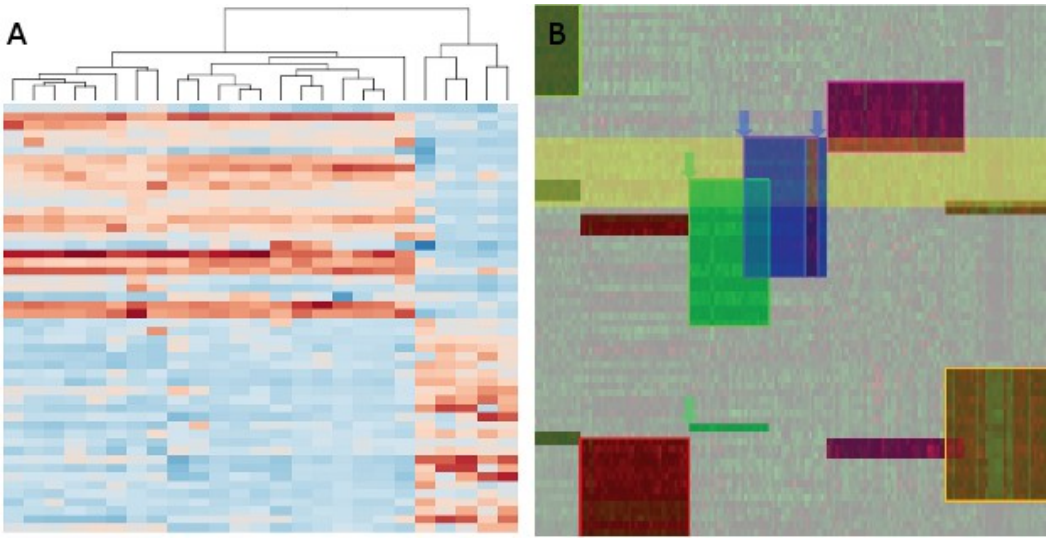

Figure 1.1 Screenshots illustration of hierachical clustering with heat maps and biclustering. The hierachical clustering in (A) was performed on columns, and the colored patches in (B) represent the biclusters identified, adapted from Luscher, *et al.* (125).

Class Prediction

Classification or class prediction using data from gene expression or metabolite concentration data has important applications in disease diagnosis, prognosis and therapy. As a result, data classification and prediction has been the subject of intensive studies among statisticians and machine learning researchers. A great variety of classification methods have been invented or adapted for analyzing or learning from high dimensional data. In this section, I will focus on three widely used methods that directly handle a large number of variables. They include partial least squares discriminant analysis (PLS-DA) which is based on dimension reduction, random forests classification, which uses an ensemble method for classification, and soft-margin support vector machines (SVM), which are based on using a soft penalty or shrinkage function to facilitate classification.

PLS-DA has been widely used by the chemometrics and metabolomics communities for many years and has served as the primary classification workhorse for several large-scale human metabolomics studies (127). Dimensional reduction in PLS-DA is achieved, like PCA, by projecting data to a small number of latent variables (LVs) via linear combinations of the original predictor variables that explain most of the *covariance* with the response (Y). These LVs are ranked by how well they explain the Y-variance. PLS-DA also produces variable importance measures such as variable importance in projection (VIP) which is a weighted sum of squares of the PLS loadings that takes into account the amount of explained Y-variance for each LV. An important issue with using PLS-DA is deciding on the number of LVs to be used to build the model. Commonly used criteria include the sum of squares captured by the model or the prediction accuracies with different numbers of LVs. Another challenge with PLD-DA is that it tends to easily overfit data (128). In practice, rigorous validation procedures including both cross validation and permutation testing are usually performed to evaluate PLS-DA results.

Random forest classification (RF) is a powerful non-parametric method and can be used for both classification and feature selection (129,130). RF uses an ensemble of classification trees, each of which is grown by random feature selection from a bootstrap sample at each branch. Class prediction is based on the majority vote of the ensemble. During tree construction, about one-third of the instances are left out of the bootstrap sample. These “left-out” data are then used as test sample to obtain an unbiased estimate of the classification error, known as

the ‘out-of-bag’ (OOB) error. This procedure makes RF classification very robust and, as a consequence, RF does not suffer from the overfitting problems seen in PLS-DA. Variable importance is evaluated by measuring the increase of the OOB error when the values of the variable are permuted. Because RF classification takes into account the context of other variables when scoring the relevance of an individual variable, RF is often applied to genome-wide association studies (GWAS) to identify genetic interactions (131). One drawback associated with the RF ensemble approach is that the classification rules are obscured as predictions are based on many different classification trees. Another known issue with RF is that when variables have very different measurement scales (i.e. one in 10s and other in 1000s, for instance) or different numbers of categories (i.e. one contains 3 groups and other 15 groups, for example), the computation of variable importance can be biased (132).

SVM is another classification method frequently used in high-dimensional data analysis (133,134). The SVM uses a set of mathematical functions, known as kernels, to map the data points into a higher dimensional feature space and then separate them by means of a maximum-margin hyperplane (135). In theory, by projecting the data into a sufficiently high dimensional space, any data set can be separated by a hyperplane. However, this strict hyperplane or hard margin approach tends to overfit data. As a result, a certain number of misclassified training examples can be accepted (soft margin approach). An important issue associated with the application of SVM methods is how to choose the best kernel function, the optimal kernel parameter(s) and the best soft margin parameter. It is

often necessary to successively increase kernel complexity (using a grid search) until an appropriate classification is achieved. Another major limitation is that SVM cannot perform automatic feature selection, although some methods have been proposed based on their weights in a linear SVM classifier (136-138).

1.5.5 Omics Data Interpretation

The output from the data analysis step in metabolomics is usually a long list of features (or metabolites) that have changed significantly under the different conditions (differential expression) or show interesting patterns of coordinated changes (co-expression). Obtaining such data is usually not the end point of the analysis; rather it is the starting point for data interpretation.

Data interpretation in metabolomics is traditionally a manual process. It normally involves manually browsing through related databases, reading published literature, and finally integrating the information into a justifiable biological "story" based on the researcher's background knowledge. This manual approach has become somewhat impractical now that so much data is being generated in this era of high-throughput omics science. As a result there is a greater reliance on computers to help with this task. Consequently, computer-assisted data interpretation is now among the hottest areas in bioinformatics research. Over the past decade, many approaches to computer-assisted data interpretation have been explored and tested. Among them, *group-based significance tests* and *pathway analysis* methods have gained wide acceptance among researchers involved in transcriptomics data analysis. These two

approaches allow the incorporation of pre-existing biological knowledge into the data analysis process and have greatly facilitated the data interpretation process.

Testing a Group of Related Variables

The basic idea in group-based significance tests is to shift the unit of analysis from a single variable (gene, protein, or metabolite) to sets of biologically related variables. This kind of biologically-reasoned approach will bring at least three immediate advantages: a) dimension reduction - reducing the number of tests by allowing one to test multiple variables simultaneously; b) taking the variable correlations into consideration by assessing their behavior simultaneously; and c) ease of interpretation as features are grouped under some biological themes thereby linking statistical significance with biological interpretation. In theory, biologically-driven and group-based significance tests will have increased statistical power to detect *subtle but consistent changes* among a group of related variables, which may fail with conventional approaches.

Many different algorithms have been developed for testing or analyzing groups of related genes based on different assumptions and statistical methods (139-143). Goeman and Buhlmann (144) suggested that these group-based significance test methods could be generally classified into two types: 1) competitive or 2) self-contained. The competitive methods test whether the genes in the gene set are *more strongly associated* with the phenotype than a random set of the same number of genes. It assumes genes are independent and randomly sampled from a *complete gene universe*. It compares the genes within the gene set with the remaining genes (background) in the gene universe to determine if the

given gene set is more closely associated with the phenotype than the background. On the other hand, the self-contained methods test whether there are genes in gene set that correlate with the phenotype. This approach directly assesses the association of the gene sets with the phenotype. Consequently, the result does not depend on the measurement of genes outside the gene set under consideration. Below I will briefly review several commonly used approaches for group-based significance tests - over-representation analysis or ORA (139), GlobalTest (143), and Gene Set Enrichment Analysis (GSEA) (142).

Over-representation analysis (ORA) is a competitive approach. It tests whether the observed proportion of genes identified as being differentially expressed in a gene set is significantly different from the corresponding proportion in the complementary set. ORA starts with a list of differentially expressed genes and tests whether a gene set is over-represented in this list more than expected by random chance. This type of analysis can be performed using Fisher's exact test, a Chi-square test, a hypergeometric test, or its binomial approximation (139,145). One major concern with ORA is its requirement for a strict cut-off in selecting differential expression of individual genes. Because this cutoff can be chosen arbitrarily, it can lead to different results with different users.

The GlobalTest (143) is a self-contained method. It tests whether a group of genes is significantly associated with a specific phenotype or clinical outcome. The null hypothesis is that none of the genes in the gene set are correlated with a clinical outcome. GlobalTest uses a logistic regression model to test this null hypothesis. Unlike ORA, it directly uses the expression data matrix without the

requirement for pre-selecting differentially expressed genes. For a significant result to come from GlobalTest, it is not necessary that the genes in the gene set have similar expression patterns; it only requires that many of them are correlated with the phenotype.

The widely used GSEA (142) is considered a hybrid approach between competitive and self-contained methods. It tests whether the dataset contains any gene set that is associated with the phenotype. GSEA first uses a univariate method (i.e. t-tests) to rank all the genes, and then tests whether the ranks in the gene set differ from a uniform distribution, using a weighted Kolmogorov-Smirnov test. The p-value for each gene set is calculated via sample permutation.

Pathway Analysis

Although the simple concept of gene sets can conveniently cover a large body of knowledge in various forms, one limitation associated with this knowledge representation is that all the members within each gene set are treated equally with no further information about their interactions or inter-relationships. For certain subsets of gene sets, substantially more detailed knowledge is available regarding their relationships in the form of transcriptional regulatory pathways, protein-protein interaction networks, and metabolic pathways. These “knowledge-rich” genes, proteins or metabolites are better tackled using pathway analysis.

Pathway analysis is a formal computer-aided analytical procedure that aims to reveal important clusters of genes/proteins/metabolites or functional modules at a higher level to facilitate biological understanding. Pathway analysis

has proven to be an invaluable tool in describing cellular responses in the context of available knowledge frameworks. Most current pathway databases focus mainly on visually displaying and highlighting matched genes, proteins or metabolites in the context of pathways, such as KEGG (146), SMPDB (147), BioCyc (148), Reactome (149), BioCarta (www.biocarta.com). Tools that support quantitative pathway analysis for gene expression data have only started appearing very recently (150-152). For instance, Tarca *et al.* described a novel Signaling Pathway Impact Analysis (SPIA) approach which combines the evidence obtained from classical enrichment analysis with a novel type of evidence that utilizes the pathway topology to measure the actual perturbation on a given pathway under a given condition (150). This hybrid approach was shown to provide increased sensitivity and specificity when compared to other methods based only on enrichment analysis.

In this section, I have reviewed the development of bioinformatics within the context of metabolomics and other related omics fields, with particular focus on various approaches for processing, normalization, statistical analysis and functional interpretation of high-dimensional expression data. These techniques and theories have provided fertile ground for my PhD research projects which I will introduce in the next section.

1.6 Research Objectives

A typical metabolomics study involves several steps: 1) researchers first design their metabolomic studies and collect the necessary samples (i.e. urine, plasma,

plant tissues, etc); 2) these samples are then measured by NMR, GC-MS or LC-MS either locally or by some metabolomics core facility; 3) the raw spectral data are then processed to generate peak lists or compound concentration tables, which are 4) subsequently subject to various statistical analyses to identify significant features or patterns; 5) finally, biological interpretations are given to the biological questions posed in the study design. During the process, researchers often need to consult various metabolite databases to help them with both metabolite identification and biological interpretation. These basic steps are summarized in **Figure 1.2**.

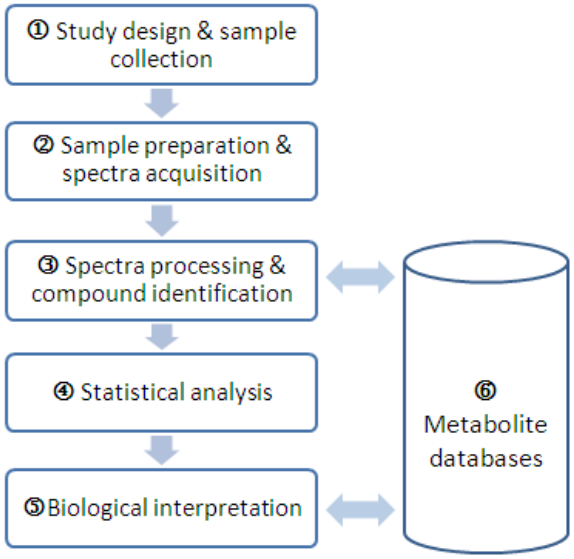

Figure 1.2 A typical workflow of a metabolomics study

Early efforts in the field of metabolomics were focused on technology development and technological refinement to establish high-throughput platforms and protocols for rapid and cost-effective data generation. Later more emphasis was placed on developing some of the bioinformatics infrastructure to help

process this flood of data. Bioinformatics efforts in metabolomics have focused on two major areas. The first one involved the construction of centralized metabolite databases to fill the equivalent role of GenBank (153) for sequence data. The past few years have seen a growing number of comprehensive and high-quality metabolomic databases emerge such as HMDB (25), BMRB (154), PubChem (155), ChEBI (156), KEGG (157), BiGG (158), METLIN (23), *etc.* The other area where bioinformatics has played a key role has been in the creation of software tools to facilitate raw data processing and compound identification. These tools have been reviewed in the section **1.4.2**. In brief, theses efforts focused primarily on step②, ③ and ⑥ as described in **Figure 1.2**, while step ④ and ⑤ remain largely underdeveloped. In addition, compound identification from raw spectra is still mainly a manual process, one of the major bottlenecks in metabolomics. Given these limitations and bottlenecks, the central objective of my thesis is to develop bioinformatics tools to facilitate high-throughput metabolomics studies, with particular focus on (1) compound identification in complex mixtures; (2) efficient metabolomic data processing and statistical analysis; and (3) high-level functional interpretation of metabolomic data.

One of the advantages of being last in the "omics" race is the benefit of hindsight. Many of the approaches developed from other omics field are not domain-specific and can be potentially adapted for metabolomics applications. The underlying rationale for the work described in this thesis is that metabolomic data analysis can be greatly accelerated by following the successes while avoiding potential pitfalls experienced in other omics fields.

1.7 Thesis Outline

This document represents a compilation of the work I have done related to the development of novel bioinformatics tools for metabolomic data analysis. The thesis is organized as follows: Chapter 1 serves as a general introduction on metabolomics technologies and provides a literature review that summarizes the current progress on omics data analysis methods and techniques. Chapters 2, 3, 4, 5 contain the detailed descriptions of the four software tools I developed and implemented: MetaboMiner, MetaboAnalyst, MSEA, and MetPA. Chapter 6 provides details on the validation, examples of real-world applications and user statistics concerning these tools. Chapter 7 is the general conclusion and future work. Appendix I contains a protocol with step-by-step instructions on how to use these web-based tools.

Chapter 2

Compound Identification from Spectra of Complex Biofluid Mixtures[1]

[1] A version of this chapter has been published previously:

Xia, J., Bjorndahl, T.C., Tang, P. and Wishart, D.S. (2008) MetaboMiner: semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, 9, 507.

Introduction

Over the past 15 years NMR has emerged as an ideal platform for studying metabolites in biofluids. It is a rapid, highly precise, non-destructive, and quantitative technique that allows one to compare, identify and quantify a wide range of compounds without the need for prior compound separation or derivatization (159-162). NMR is particularly amenable to compounds that are less tractable to GC-MS or LC-MS analysis, such as sugars, amines, volatile ketones and relatively non-reactive compounds. A key disadvantage of NMR is that it is a relatively insensitive technique, with a lower limit of detection of 1~5 μM and a requirement of relatively large sample sizes (~500 μL).

Currently, most NMR-based metabolomic studies involve the analysis of 1D ^{1}H NMR spectra, although 1D ^{13}C and ^{31}P NMR spectra may also be analyzed (163-166). There are generally two routes to analyzing NMR spectra for metabolomic studies. In one method (called the chemometric approach), the compounds are not initially identified – only their spectral patterns and intensities are recorded and statistically compared in order to identify the relevant spectral features that distinguish sample classes. Once these features have been located, a variety of approaches may then be used to identify the corresponding metabolites (167). In the other approach (often called quantitative metabolomics or targeted profiling), compounds are first identified and quantified by comparing the NMR spectrum of the biofluid of interest to a spectral reference library obtained from pure compounds (168). Once these compounds are identified and quantified, the

data can be analyzed in many different ways to identify the most relevant biomarkers or informative pathways.

A variety of protocols and software tools have been recently developed for conducting quantitative metabolomics via 1D ^{1}H NMR (161,162,168,169). In most cases, a manual peak-fitting process is required in order to perform compound identification and quantification. However, this manual fitting process can become particularly difficult and prone to frequent errors, especially for very complex biofluid mixtures (such as urine or tissue extracts) due to severe spectral overlap. In contrast to 1D NMR, 2D NMR offers a robust approach to resolving excessively overlapped spectra. Indeed 2D (and 3D) NMR has long been used to resolve and identify individual resonances from large macromolecules such as DNA, RNA and proteins. 2D NMR experiments such as TOCSY, HSQC and J-resolved spectroscopy are also increasingly being used in metabolomic studies in order to resolve spectral ambiguities to aid in the identification of specific compounds in complex biofluid mixtures (103-105,170-175).

A number of small-molecule NMR databases have been developed in recent years to support metabolomics research, including the Human Metabolome Database (HMDB) (176), the BioMagResBank Database (BMRB) (154), the Madison Metabolomics Consortium Database (MMCD) (177), the Magnetic Resonance Metabolomics Database (MRMD) (22), and the Platform for RIKEN Metabolomics (PRIMe). These resources, which contain significant numbers of reference NMR spectra of metabolites, also support metabolite identification through web-based submission of 1D and 2D NMR peak lists. However, these

on-line tools do not generally provide graphical support for peak filtering, processing, comparative display or annotation. Furthermore, they don't exploit additional constraints such as knowledge about biofluid composition or metabolite concentration ranges to make compound identification even more robust, more accurate, and more efficient.

My hypothesis is that, by creating a comprehensive 2D spectral reference library along with some biological constraints (i.e. a more detailed knowledge of metabolite compositions of different biofluids), it is possible to develop an algorithm that largely automates compound identification from 2D NMR spectra. To validate and test this hypothesis I have developed a stand-alone program called MetaboMiner, to perform semi-automated metabolite identification from 2D TOCSY and HSQC-spectra of complex biofluid mixtures. The remaining sections of this chapter provide detailed descriptions of the program implementation, followed by a performance evaluation and comparison with other available software tools. The chapter ends with a discussion about the limitations and challenges in the 2D NMR based metabolomics.

Implementation

Data Collection and Curation

Key to the development of this software package was the creation of an extensive 2D spectral library containing TOCSY and HSQC spectra of pure metabolites. We used several publicly available sources in constructing this library. The

majority of the raw ^{1}H-^{1}H TOCSY spectra were collected from the standard compound spectral library available at the BMRB database (154). A few additional compound spectra were obtained from the MRMD (22). The ^{1}H-^{13}C HSQC spectral library was downloaded from the HMDB (176). These raw spectra contained a number of spectral artifacts (noise, water bands, asymmetries, peaks from TSP or DSS, contaminants, etc.). Consequently it was necessary to convert these raw spectra into "synthetic" or "simplified" spectra corresponding to the peaks specific to the pure compounds of interest. This conversion was done manually, with each of these simplified, noise-free spectra being examined for inconsistencies by comparing them to the original raw spectra and the compound's known resonance assignments. In total, the MetaboMiner TOCSY reference library includes spectra from 223 common metabolites and the MetaboMiner HSQC library contains spectra from 502 metabolites. The compounds in both libraries were further catalogued into three sub-libraries corresponding to the three common human biofluids – cerebrospinal fluid (CSF), plasma and urine. The classification was based on their respective metabolic compositions listed in the HMDB. Since the presence of these biofluid-specific metabolites was determined by a variety of technologies not limited to NMR, we further investigated the appearance of these metabolites in a large number of 1D ^{1}H spectra collected in-house from human CSF, plasma and urine samples under various conditions. The combined collection of compounds (and spectra) was used to create corresponding "common biofluid" 2D NMR spectral libraries that effectively represent a generic biofluid or cell extract. The "CSF", "plasma",

"urine", "biofluid" and "total" spectral libraries are stored as XML files and are editable via MetaboMiner's graphical user interface (GUI).

After the spectral libraries were constructed, each peak for each compound in each library was assigned a series of uniqueness values that are specific for that reference library. A unique peak in MetaboMiner is defined as a relatively isolated peak around which no peak from any other compound is observed based on the spectral library of the given biofluid. For any given peak, its uniqueness value is calculated as the total number of surrounding peaks from other compounds within a given chemical shift "distance". Five distance levels were used to measure peak uniqueness. For ^{1}H chemical shifts, the distance thresholds are 0.01, 0.02, 0.03, 0.04, and 0.05 ppm. For ^{13}C chemical shifts, the distance thresholds are set at 0.05, 0.10, 0.15, 0.20, and 0.25 ppm. For instance, an HSQC peak with a series of assigned uniqueness values of 0-0-0-1-2 indicates that no peak from any other compound in the reference library is observed within 0.03 ppm (^{1}H dimension) and 0.15 ppm (^{13}C dimension) of that peak. It also indicates that one peak from another compound in the spectra library was observed within 0.03 ~ 0.04 ppm (^{1}H dimension) and 0.15 ~ 0.20 ppm (^{13}C dimension) and another peak from another compound was observed within 0.04 ~ 0.05 ppm (^{1}H dimension) and 0.20 ~ 0.25 ppm (^{13}C dimension). See **Figure 2.1** for a more complete description of the uniqueness value concept. These uniqueness values are automatically updated after any spectral library change using MetaboMiner's GUI.

Peak Processing, Peak Matching and Compound Identification

As part of its input, MetaboMiner requires peak lists corresponding to the peaks that were identified in either the TOCSY or HSQC spectra collected from the biofluid(s) of interest. While it is possible for users to provide manually picked peak lists, MetaboMiner also supports processing of multidimensional NMR peak lists obtained from automatic peak peaking programs. Multidimensional NMR spectra typically contain substantial numbers of spectral artifacts such as baseline distortions, intense solvent lines, ridges, sinc wiggles (truncation artifacts obtained by Fourier transforming truncated time-domain signals, usually caused by too short acquisition time), *etc*. Automatic peak-picking programs tend to mistake these noise signals for real resonances. Therefore, any raw 2D spectra collected from biofluids must be processed appropriately before attempting to match them to MetaboMiner's reference spectral library. Two automated procedures were found to be very effective in cleaning up raw 2D spectra: 1) streak removal and 2) symmetrical editing. Note that the latter processing technique is only applicable for TOCSY spectra. Spectral streaks are usually caused by residual solvent signals (i.e. water) or the presence of other compounds at extremely high concentrations. Streaks can be recognized by their specific locations and prominent shapes in the NMR spectra. Streak removal was implemented by searching for groups of peaks at these common locations and eliminating them from the peak list. Symmetrical editing exploits the fact that real TOCSY peak signals form a symmetrical square pattern along the diagonal line. Off-diagonal peaks without any corresponding symmetrical peaks can be

considered to be artifacts. Both peak positions and intensities (if provided by the user) of the corresponding peaks are examined for symmetry. Since TOCSY cross peaks are frequently not of equal intensity we require that the intensity ratio between the upper and lower-diagonal peaks should be within an empirical range of 0.8~2.5 of each other to be considered symmetrical.

In order to accommodate small chemical shift differences between the observed NMR spectra and the reference NMR spectra, an adaptive threshold method was implemented based on the uniqueness values (described above) of each reference peak. During the peak searching/matching process, the search threshold varies automatically based on the maximum uniqueness value of the current peak. For instance, when searching for potential matches for a TOCSY peak with uniqueness values of 0-0-0-0-1, MetaboMiner will automatically set its threshold to 0.04 ppm. The peak matching and adaptive thresholding employ two processes: a reverse search strategy and a forward search strategy. In the reverse search strategy, the library peaks are searched and matched against the query peaks. Typically most query peaks find their potential matches during this reverse search step. However there are usually some peaks left without any matches. In order to assign these unmatched peaks a forward search is performed in which the unmatched query peaks are searched against the reference library with expanded but fixed thresholds – 0.08 ppm for TOCSY and 0.12 ppm (^{1}H) and 0.4 ppm (^{13}C) for HSQC spectra. A match is identified if only a single reference peak is identified within this range.

In MetaboMiner a compound is considered to be present only if its matched pattern satisfies the requirements of what we call “minimal signatures”. A minimal signature is defined as the minimum peak set that can uniquely identify a compound from all others in a given spectral library. Based on the complete peak set of the reference spectral library, many minimal signatures can be derived through different combinations of unique peaks. A single peak match may be considered a minimal signature if it is completely unique. More peaks are required to define a minimal signature for less unique ones. For instance, in our current implementation, the presence of a single peak with uniqueness values 0-0-0-0-x (x>=0) will determine the presence of the corresponding compound (subject to authenticity checks as discussed later); while at least two peaks with uniqueness values 0-0-0-x-x are required to reach the decision.

Since query spectra (i.e. real spectra from biofluids) usually contain substantial levels of spectral noise, even after pre-processing, we found that we could reduce MetaboMiner’s false positive rate even further by implementing several authenticity checks. These include: 1) having a minimum number of matched peaks (3 for TOCSY spectra and 1 for HSQC spectra), 2) having a minimum matched fraction of peaks (1/2 for TOCSY spectra and 1/6 for HSQC spectra), 3) ensuring the presence of certain peaks for certain compounds (determined by manual testing and validation for each compound), and 4) ensuring that the identified compounds were known to be in a given biofluid. It is important to note that these procedures aim to reduce false positives when the algorithm is used for automatic compound identification. Compounds that fail to

meet these criteria but otherwise have matched peak(s) will be presented for user inspection via the GUI as described below.

User Interface Description

MetaboMiner's graphical user interface was implemented using Java Swing technology. The spectral visualization and manipulation tools were built using the JGraph library (Java open source graph visualization library, http://www.jgraph.com). **Figure 2.2** illustrates a flowchart describing the MetaboMiner GUI. There are four main functional views, 1) a Processing View, 2) a Search View, 3) an Annotation View, and 4) a Library View. All these views share the same component arrangement, with panels on the right side being used for visualizing and manipulating peaks, and the panels on the left being used for displaying parameters, compound lists, structure images, *etc.* Navigation to each view is readily accessible by clicking an appropriate menu item.

When the program launches, the default view is the "Processing View" where users can copy and paste the automatically picked peak list. The input format must be either a two or three-column list, with numbers separated by a space or a semicolon. The first two columns must be the x and y chemical shift coordinates of each peak in the 2D spectrum and the optional third column must be the peak height or peak intensity. After processing the raw peaks, both the original and the processed spectra will be displayed on MetaboMiner's spectral viewing panel (located on the right). With this viewing panel, users can directly edit peaks on the spectrum if necessary. For manually picked peaks, this step can

be skipped by turning the processing options off. By clicking the “Search” button, MetaboMiner’s “Search View” will be displayed with its initial, automated compound identification results. Users can adjust the search threshold or switch the reference library to further refine the result. A compound is marked as identified if the matched pattern passes the authenticity checks and satisfies the minimum signature requirement. The raw matched scores are also displayed. MetaboMiner’s interface allows users to visually inspect the matched peaks of any metabolite against the corresponding reference spectrum. By right clicking any peak displayed on the spectrum, users can search the library for this particular peak. The identified compound list can be saved in three different formats by clicking the “Export” button. A screenshot of MetaboMiner’s “Search View” is shown in **Figure 2.3.**

Users can further refine the automated search results by manually annotating the raw 2D spectrum. By clicking the “Refine” button in the “Search View”, the “Annotation View” will be launched with the identified compounds being transferred as the starting point. Users can also directly enter the “Annotation View” mode by clicking the “Annotate” button from the “Console” menu. In order to perform manual annotation, users first need to load a high resolution spectral image in PNG format and set up the spectral axes properly. Peak searching is performed by right clicking the peak position on the spectrum to search the reference library as shown in **Figure 2.4**. All compounds that generate peaks within the search threshold will be checked. The compound with the closest peak match will be highlighted with its database reference spectrum displayed on

the uploaded “raw” spectrum. Users can perform peak annotation for any currently displayed compound. Double clicking any database peak will open a small text editor where users can enter the peak assignment or a comment**.** The peak pattern of the identified compounds can also be edited to match the experimental spectrum. For example, users can insert, delete, or drag a database peak to match the observed peak in the raw spectrum. These changes will be valid only for the current session. To make permanent changes, MetaboMiner’s “Library View” must be used.

The “Library View” is intended for browsing and managing MetaboMiner’s spectral libraries. To view all the available reference spectra in MetaboMiner’s libraries, users must click the “Browse” button in the “Library” menu. Double clicking any compound in the compound list will open a popup window for peak editing. Any changes will be reflected on the spectrum at real time. New compounds can be introduced by clicking the “New” button at the bottom of the compound list. A new compound can be either exported from another library or be created from scratch through the wizard dialog. Both peak editing or adding new compounds will trigger updating of the uniqueness values of the affected peaks. For researchers who study other types of biological samples (e.g. plant or microbial extracts), they may either use MetaboMiner’s generic spectral reference library or create a new library customized for that particular type of biofluid. Library creation or deletion can be easily accomplished by clicking the appropriate menu items in the “Library” menu. The compounds in the default reference library are linked to PubChem, HMDB, and the BMRB via the

hyperlink under their structure icon. The "Graphics" menu enables users to change the size, shape, or color of the synthetic peaks to suit their preferences.

It is important to note that MetaboMiner does not support spectral processing such as phasing, baseline correction or chemical shift referencing. There are many other high-quality NMR-processing software available for this task, including NMRPipe (99), Felix (Molecular Simulations, Inc.,San Diego, CA), VNMR (Varian, Inc., Palo Alto, CA), and XWinNMR (Bruker Analytik GmbH, Karlsruhe, Germany), to name a few. These tools should be used prior to loading spectral images into MetaboMiner. In other words, MetaboMiner is not a spectral processing tool, but a NMR-based metabolomics tool that facilitates automatic peak processing, rapid compound identification, and facile spectrum annotation capabilities through an intuitive graphical interface. MetaboMiner is available at: http://wishart.biology.ualberta.ca/metabominer.

Evaluation

MetaboMiner was assessed in a variety of ways using both synthetic and experimental NMR spectra. The synthetic spectra were generated from the 162 compounds that have both TOCSY and HSQC spectra in the reference library. The experimental spectra were collected from three defined compound mixtures (totalling 72 compounds) and a biofluid sample of known composition (plasma). These evaluations allowed a complete and comprehensive assessment of MetaboMiner's performance as well as its potential strengths and limitations.

The Effects of Different Spectral Noises on Compound Identification

The performance of the minimal signature method and the adaptive threshold method were evaluated under two common types of spectral noise – missing peaks and "drifting" peaks (i.e. peaks that have drifted from their canonical positions due to temperature, pH or solvent effects). The missing peaks were simulated by deleting peaks of each compound at random with 0%, 10%, 20%, 30%, 40%, 50% probabilities. The chemical shift drift effects were simulated by adding random values of ±0.01, ±0.02, ±0.03, ±0.04, ±0.05 ppm for each ^{1}H chemical shift, and ±0.05, ±0.10, ±0.15, ±0.20, ±0.25 ppm for each ^{13}C chemical shift. The spectra of each synthetic query mixture were generated by first pooling the peaks from 50 compounds that were randomly selected from the MetaboMiner reference spectral library (162 compounds). After introducing this artificial spectral noise, the query mixtures were searched against the reference spectral library with and without using the adaptive threshold method. Two compound identification strategies were compared - the minimal signature method (MS) and the percentage match method (PM) with 75% as the cut-off value. The F-measure was used for performance evaluation, where $F = 2 \times (precision \times recall) / (precision + recall)$ where recall is the proportion of true positives in the returned result ($recall = TP/(TP+FN)$) and precision is a measure of the percentage of positive or correct results ($precision = TP/(TP+FP)$). The values were obtained as the averages of TOCSY and HSQC search results over 50 iterations. **Figure 2.5A** summarizes MetaboMiner's performance using data with different fractions of

missing peaks. **Figure 2.5B** shows the results using data with increasing chemical shift drift effects.

The Effects of Different Data Types on Compound Identification

We further investigated the usefulness of different NMR data types for compound identification based on our concept of a minimal spectral signature. Four NMR data types were compared - 1D ^{1}H, 1D ^{13}C, ^{1}H TOCSY, and ^{1}H-^{13}C HSQC spectra. For this particular evaluation, reference 1D ^{1}H and 1D ^{13}C spectra were obtained from the corresponding ^{1}H and ^{13}C chemical shifts of MetaboMiner's HSQC spectral library. For a small number of compounds, these artificial 1D spectra lacked some of the expected ^{1}H or ^{13}C signals that might be seen in a real 1D NMR spectrum, but their absence also helped to simulate the fact that some peaks in 1D NMR spectra are broadened or washed out due to signal overlap or solvent suppression.

Synthetic 2D NMR spectra (query spectra) representing different biofluids of increasing molecular complexity were generated by pooling peaks of 20, 30, 40, 50, 60, 70, and 80 compounds randomly selected from MetaboMiner's reference spectral library. To further simulate noise or pH/salt effects, 10% of the peaks from the query spectra were deleted at random, followed by the introduction of random chemical shift changes (±0.01 ppm for ^{1}H and ±0.05 ppm for ^{13}C) to the remaining peaks. The resulting peaks were subsequently searched against MetaboMiner's reference spectral library using the adaptive threshold

method. The F measures were averaged over 50 iterations. The result is summarized in **Figure 2.6**.

Compound Identification Using Experimental Spectra

Twelve 2D NMR experiments (six TOCSY and six HSQC) were collected under different pH conditions using three synthetic mixtures and a plasma sample. The three synthetic mixtures were composed of 27, 21, and 24 common metabolites, respectively, with concentrations ranging from 40 to 60 mM. The plasma sample contained 35 identifiable metabolites (ranging in concentration from 0.1 to 10 mM) as determined by independent profiling of its 1D 1H NMR spectra by several experienced individuals using Chenomx's NMR Suite software. These results were further confirmed by spiking/doping authentic standards into the plasma sample and by GC-MS analysis. The plasma sample was prepared by first lyophilizing and then dissolving the remaining solids in distilled water to its 1/5 original volume. Deuterium oxide (D_2O) was added to make a final concentration of 90% H_2O and 10% D_2O. All spectra were acquired at 25 °C. Six spectra were collected on a Varian INOVA 800 MHz spectrometer equipped with a 5 mm triple axis gradient cryoprobe. The other six spectra were collected on a Varian INOVA 500 MHz spectrometer with a 5 mm triple-resonance z-gradient probe. The TOCSY experiments were performed using the wgtocsy pulse sequence, and the HSQC experiments were performed using the gChsqc pulse sequence, both provided by Varian's BioPack. For the TOCSY experiments, the spectral width was set to 11990 Hz and a mixing time of 0.05 seconds. Sixteen transients were collected for each t1 interval using an acquisition time of 0.085 seconds with a

relaxation delay of 2.0 seconds. One increment was collected for every 46.825 Hz in t1 dimension. For the 13C-HSQC experiments, the spectral widths of the proton and carbon dimensions were 11990 Hz and 28160 Hz respectively. Sixty four transients were acquired for each t1 interval using an acquisition time of 0.085 seconds and a relaxation delay of 1.0 seconds. The spectra were collected with 2048*256 complex points for the 1H and 13C dimensions respectively. The total spectral acquisition time was ~5 hours. Sample TOCSY and HSQC spectra are shown in **Figure 2.7** and **2.8**.

The raw NMR spectra were first processed using NMRPipe (99) and the peaks were subsequently picked using Sparky's (100) automatic peak picking program. The resulting "raw" peak lists were copied and pasted to the processing view of MetaboMiner. Both peak processing and compound identification were performed using MetaboMiner's default parameter sets. The reference library used for the synthetic mixtures was the biofluid (common) library. For plasma data, the plasma (common) library was used. To assess the degradation in performance assuming no prior knowledge of the sample source (urine, plasma, cell extract or generic biofluid) the complete spectral reference library (223 compounds for TOCSY, 502 compounds for HSQC) was also used to identify compounds. To assess the performance of the web-servers that support 2D NMR mixture analysis -- the HMDB (176), the MMCD (177), the BMRB (154), and the SpinAssign (178) of PRIMe (http://prime.psc.riken.jp) – the same set of peak lists were submitted. For PRIMe, the default search parameters were used. For other

web services, the search threshold for 1H was set to 0.03 ppm and 0.10 ppm for 13C. The results are summarized in **Table 2.1**.

Results

Using synthetic query spectra constructed as described previously, the performance characteristics of MetaboMiner were first assessed under different levels of spectral noise. Secondly, the utility of different NMR data types were also investigated for our approach of compound identification. Finally, we evaluated MetaboMiner's performance using a total of 12 real NMR spectra collected from defined compound mixtures and a plasma sample of known composition.

After creating the spectral reference libraries and calculating the uniqueness values for each peak, we first investigated the performance of the minimal signature (MS) method versus the adaptive threshold method under different types of spectral noise. As an additional comparison, the percentage match (PM) method was also included. As shown in **Figure 2.5**, the MS method performed consistently better than the PM method when missing peak and chemical shift variations are present. The adaptive threshold method appears to be most effective for data exhibiting large chemical shift variations. When chemical shift variation is negligible, as in the test with missing peak data, this method performs exactly the same as the fixed threshold method.

The utility of different NMR data types was also investigated using synthetic spectra of increasing complexity. As illustrated in **Figure 2.6**, HSQC, TOCSY and ^{13}C -based methods worked very well over the full range of compound mixtures. The number of compounds in the query mixture has very little effect on their overall performance. The slight increases in F scores with increasing spectral complexity are not statistically significant. In general, MetaboMiner's performance for compound identification was best using the HSQC dataset. It is also apparent that the minimal signature method favours compound identification with HSQC spectra as these spectra tend to have more unique peaks than other types of NMR spectra. Also evident from **Figure 2.6** is the fact that MetaboMiner's performance using 1D ^{1}H data, alone, is the poorest. In contrast to the ^{1}H spectra (both 1D and 2D), it is quite clear that ^{13}C chemical shifts (even in 1D spectra) can provide sufficient information for robust compound identification. This is mainly due to the much wider chemical shift dispersion (0~200 ppm) seen in ^{13}C spectra compared to ^{1}H spectra. Most ^{13}C chemical shifts remain unique even in mixtures of 162 compounds.

While ^{13}C spectra (1D and 2D) provide excellent data sets for compound identification we found that by focusing on off-diagonal peaks originating from the coupling between pairs of protons, the utility of TOCSY spectra could be greatly improved. As indicated in **Figure 2.6**, MetaboMiner's metabolite identification performance based on TOCSY spectra was better than that based on 1D ^{13}C spectra and was only slightly outperformed by ^{13}C HSQC data. These

results underscore the utility of using 2D spectra in NMR-based compound identification of complex (>20 compounds) mixtures.

We also assessed MetaboMiner's performance on its own and against several other web services using eight experimental NMR spectra collected from four different mixtures of known composition. As indicated in **Tables 2.1A** and **2.1B**, the best performance for MetaboMiner was obtained when a biofluid-specific reference library was used to analyze TOCSY or HSQC data. On average, MetaboMiner was able to correctly identify (recall, precision and F-measure) an average of 81% of the compounds from both TOCSY and HSQC data. When the entire spectral library (223 TOCSY, 502 HSQC) was used, the performance (F-measure) decreased by an average of 15%. Among the four web services evaluated using the same data (**Table 2.1A**), the SpinAssign program performs the best (F-measure = 49%) but this is still about 30% worse than MetaboMiner when it uses a biofluid-specific library and 15% worse than MetaboMiner when it uses its entire spectral library. Overall, the other web servers did not perform particularly well with average F-measures of 15-25% for HSQC data and 6-12% for TOCSY data. Both the HMDB and MMCD servers performed better when analyzing HSQC data than TOCSY data. The performance for all web servers was essentially the same regardless of whether "clean" peak lists (no noise peaks) or "raw" peak lists were used as input. Note that TOCSY mixture analysis is not currently supported by either PRIMe or BMRB.

In an effort to understand the influence of pH on the efficacy of compound identification by 2D NMR, we also collected TOCSY and HSQC data at pH 4.2

and pH 8.8. This is approximately 3 pH units below (and 1.5 pH units above) the pH at which the spectral library standards were collected. As seen in **Table 2.2**, a significant pH change in the sample (relative to the pH of the spectral libraries) can negatively impact the performance of compound identification. For instance, at pH 4.2, the F-measure drops by more than 20% for the HSQC spectra.

Discussion

In this chapter, we described the development and assessment of a software tool (MetaboMiner) to facilitate compound identification from 2D NMR spectra of biofluids or small molecule mixtures. We first created a series of "clean" spectral reference libraries based on publicly available spectral databases. Secondly we developed and tested several algorithms to facilitate robust and automatic peak processing, peak matching and compound identification. Finally, we integrated these resources into an easy-to-use application and evaluated its performance using a variety of synthetic and real spectra.

MetaboMiner's spectral reference library covers most NMR-detectable metabolites present in human biofluids (176). To improve the reliability of the compound identification, the larger spectral library was further partitioned into three smaller libraries based on the composition of different biofluids (CSF, plasma and urine). In addition, we also created "common" libraries for each type of biofluid that contain the most common or abundant metabolites found in these biofluids (as ascertained from previous experience and from data contained in the HMDB). We found that the creation of biofluid-specific libraries significantly

improved compound identification by reducing the spectral search space. As indicated in **Tables 2.1A** and **2.1B**, a 15% reduction in MetaboMiner's performance occurred if the entire compound library was used instead of the biofluid-specific "common" library. For researchers who wish to study other types of biofluids, MetaboMiner provides intuitive interfaces that allow users to easily expand and customize their spectral reference libraries.

Performance Assessment

The performance of MetaboMiner was evaluated using a variety of synthetic and experimental datasets. In all cases, our strategies for peak matching and compound identification showed robust performance under various noise (real and synthetic) conditions. Using synthetic data, the best compound identification performance was ~90% (F-measure) under moderate noise levels as indicated by **Figure 2.6**. Further inspection of the compound identification lists showed that the most common problem was the identification of false positives. In particular, for certain compounds it is inherently difficult to uniquely identify them by NMR based on their matched peaks. For example, the TOCSY peaks of citrate and serine cluster very tightly around the diagonal. They also overlap with peaks of other more abundant compound species. As a result, they are often misidentified. Another source of false positives comes from the existence of structurally similar compounds such as *asparagine/aspartate, inosine/adenosine*, or *creatine/creatinine*, which have nearly identical NMR spectral features.

When we assessed the performance of MetaboMiner using experimentally collected data, the performance was reduced by ~10% for both TOCSY and HSQC data. Close examination of the lists of identified compounds as well as the spectra used in the evaluation indicated two sources of problems. The first relates to the fact that several compounds known to be in the mixtures failed to be identified (false negatives). Manual inspection of the actual spectra (TOCSY or HSQC) indicated that in every case, no peaks or very weak peaks were visible for these compounds. These compounds were obviously below the detection threshold of the instrument. The second problem was the existence of several false positives. Again, manual inspection of the spectra showed that the false positives were mainly caused by spectral artefacts. As illustrated in **Figure 2.7**, real spectra can contain a significant amount of spectral noise. In the case of TOCSY spectra, most of the automatically picked peaks (>60%) are from these artefacts. Obviously if cleaner spectra could be collected or if more manual intervention was used to eliminate some of the spectral artefacts prior to submitting the data to MetaboMiner, a better performance could be achieved. In addition, we also observed that some compounds are exquisitely sensitive to small pH variations such as *lactate*, *uracil* and *histidine.*

The Challenges in Automated Compound Identification

There are three major challenges facing automatic compound identification for NMR-based metabolomics. The most common and perhaps the most vexing is the so-called spectral overlap problem. The spectral complexity inherent in many biofluids can lead to a large number of peaks confined to a relatively narrow

chemical shift range (~10 ppm for ^{1}H spectra). With the increased availability of high-field NMR spectrometers, the problem of spectral overlap is diminished somewhat for simple biofluids such as CSF. However, for complex mixtures like plasma, urine, or tissue extracts, the problem is still quite severe. The second challenge in automated compound identification is the handling of chemical shift changes induced by the variation of pH, temperature, ionic strength, *etc.* This effect combined with the first issue makes it difficult to perform automated compound identification based solely on chemical shifts. The third challenge in automated compound identification is the low signal-to-noise (S/N) ratio for the NMR resonances of low abundance compound species. Consequently, the wide range of metabolite concentrations found in many biofluids poses a serious problem for most automatic peak-picking programs. In particular, many low intensity peaks are likely to be missed by most peak picking programs.

To address these challenges we introduced the notion of "uniqueness" in the reference library in order to deal with the problem of missing peaks and chemical shift variations. The uniqueness values were calculated for every peak based on their relative distances to each other in a given reference library. Although a finer scale may perform better for more complex mixtures, we found that the use of five levels of uniqueness works well for most situations. In MetaboMiner, both peak matching and compound identification rely heavily on these uniqueness values. During peak matching, an adaptive threshold method was used to adjust the current search threshold to its maximum uniqueness scope. This approach significantly improves MetaboMiner's performance when chemical

shift variations are nontrivial. By using an adaptive threshold method we found the effect of chemical shift changes or chemical shift drift could be tolerated to a greater extent. These uniqueness values also allow us to derive minimal signatures based on a relatively small portion of unique peaks. This approach turned out to be both robust and flexible since it did not require “stable” spectral patterns. In contrast, the common percentage match (PM) method weights each peak equally and uses a fixed threshold for compound identification. This approach suffers greatly if a nontrivial proportion of peaks are not matched. Note that the minimal signature method in MetaboMiner is complemented by authenticity checks with due consideration of the total matched patterns.

It should be noted that the performance of MetaboMiner strongly depends on the quality of the upstream spectral collection and processing work. As a general rule, for optimal performance, the NMR experiments on biofluid mixtures should be carried out under the same (or at least similar) conditions as the conditions used to collect the spectra for the reference library (i.e. neutral pH). The automatic peak-picking process should be closely monitored and an iterative approach is recommended in order to pick up most signals while avoiding obvious spectral noise. In our testing process, a typical TOCSY spectrum usually generated 2,000~3,000 peaks with a high proportion of noise peaks. We found that these noisy signals were handled quite efficiently by MetaboMiner. As a general rule, we would suggest that users employ a low threshold during the peak picking stage for TOCSY spectra. For HSQC spectra, because of the difficulty associated

with detecting and removing noisy signals, we would suggest more manual intervention during the peak picking process.

Comparison to other spectral analysis software tools

There are several commercial software tools available for analysing complex metabolite mixtures using NMR. Chenomx Inc. (an Edmonton-based metabolomics company) has developed a commercial bioprofiling software package called the Chenomx NMR Suite that allows semi-automated identification of compounds from 1D ^{1}H NMR spectra. The Chenomx software package provides an excellent interface for compound identification and quantification via a manual peak-fitting process using a spectra reference library containing 260 compounds. However, the requirement for manual fitting and analysis leaves the process open to inconsistent interpretation or inconsistent assignment by different individuals. Furthermore, the analysis can take upwards of one hour per sample and the software is relatively expensive. Bruker's AMIX (Bruker BioSpin) software is another powerful tool that offers support for compound identification and quantification for both 1D and 2D NMR. It used a method called AutoDROP to facilitate compound identification and structure verification (179). The key idea is the systematic decomposition of reference spectra into spectral patterns of molecular fragments. Compound identification is based on recognition of such patterns in the target spectra. However, similar limitations pertaining to cost, the reliance on manual analysis, processing time and inconsistent interpretation appear to apply to AMIX as well.

In addition to these commercial packages there is at least one other non-commercial system that has been described. Xi *et. al.* (105) developed a statistical and chemical model for automatically identifying compounds in mixtures using 2D COSY spectra. Like MetaboMiner, the Chenomx NMR Suite and AMIX, Xi *et al*'s method uses a library of pre-collected NMR spectra to assist with compound identification. These authors reported experimental results using spectra collected from abalone muscle and digestive gland extracts. Their method was able to identify 12 out of 15 amino acids and 6 out of 9 amino acids as determined with Chenomx's NMR Suite. However, it appears that this system has a spectral library of only 19 COSY spectra, so it is very limited in terms of practical applications.

In addition to these stand-alone, graphically based software packages, several metabolomics database websites, including the HMDB (176), MMCD (177), BMRB (154), and PRIMe now allow direct querying of their databases using peak lists obtained from compound mixtures. However, as web servers they are somewhat limited in their graphic capabilities and user-interface interactions. In particular, most of these sites typically return long lists of potentially matched compounds without a graphic display of the matched spectra to help users make their decisions. Further, most of the servers appear to be designed to handle single compound or simple mixture queries and are not optimized for compound identification from complex biofluid mixtures. This is particularly evident from the results shown in **Tables 2.1A** and **2.1B**. There are some possible explanations regarding the different performance of the web services under comparison. In

particular, the PRIMe database only contains ~80 common compounds for the NMR search, which significantly reduces number of false positives. The HSQC reference spectra used by MetaboMiner as well as the test samples were collected under pH 7.0~7.2 with 10% D_2O. This is quite different from the conditions used by MMCD and BMRB (pH 7.4 in solvent D_2O, which corresponds to a pH 7.8 in H_2O given the deuterium isotope effect). The pH mismatch may partially explain the poor performance using our test samples. Please note that in most cases, using the whole library instead of biofluid-specific library will increase of both type I and type II errors. This is because when more compounds are included in the library, NMR spectral peaks become less unique due to significant chemical shift overlap. As a result, the minimal signature approach becomes more error-prone when judging the presence or absence of a particular compound.

Overall, MetaboMiner combines many of the useful interactive graphic features and high levels of performance of the stand-alone commercial packages such as Bruker's AMIX and Chenomx's NMR Suite with the relatively simple automation or semi-automation seen with NMR-based metabolomics web servers. Key to MetaboMiner's success are its large and carefully constructed spectral libraries, its robust spectral filtering and peak matching routines, and its use of biofluid-specific spectral libraries to rationally limit the spectral search space. Given the extensive testing and the availability of many built-in tools for spectral manipulation, viewing and annotation we believe MetaboMiner is well designed and ready for practical metabolomics applications.

Limitations

MetaboMiner is not without its limitations. In particular, MetaboMiner does not support compound quantification. Efforts are underway to add this functionality to the program but it appears that quantification via 2D NMR spectra is intrinsically more difficult and less reliable than with 1D NMR spectra. Lewis *et al.* (180) recently described a quantification method from HSQC spectra based on standard curves calibrated for each selected unique peak of a given compound. Quantification by TOCSY still remains difficult because the specific transfer functions for complex spin systems are common for many metabolites. When these techniques are established, it is our next step to provide more functions to better support the research community. Another limitation (which is also shared by other NMR analysis programs including Chenomx's NMR Suite and Bruker's AMIX) is the fact that MetaboMiner's performance is highly dependent on the spectral quality and spectral pre-processing of the query (i.e. experimental) spectra. High signal-to-noise, good phasing, minimal baseline distortion and the elimination of spectral artefacts will always improve the performance. However, some biofluid samples may be refractory to good spectral processing or some users may lack sufficient experience/skill to properly process their spectra. Under these circumstances, MetaboMiner's results may prove to be unreliable or non-reproducible. A third limitation to MetaboMiner is its limited sensitivity. In particular, MetaboMiner's exclusive reliance on 2D NMR spectra generally reduces its sensitivity limit by a factor of ~10 over what might be detected via 1D spectrum. Obviously the use of more concentrated samples, longer collection

times or isotopic labelled samples can overcome these problems, but the trade-off between time, cost and convenience may not always be in MetaboMiner's favour. It is worth noting that Methods such as the acceleration by sharing adjacent polarization (ASAP) HMQC (181) now allow very rapid acquisition of 2D NMR data. This may make the issues of sensitivity and time much less important, particularly for 2D heteronuclear experiments. A fourth limitation is that fact that MetaboMiner's spectral libraries (esp. the TOCSY library) are still missing a number of important compounds. Efforts are underway to expand the TOCSY library over the coming months and the MetaboMiner website will provide periodic reference spectral updates. Alternately, users are invited (and encouraged) to add their own spectra to MetaboMiner's reference libraries. Finally, unlike Chenomx's NMR Suite and Bruker's AMIX, MetaboMiner's spectral libraries do not cover a broad range of pH values. As a result, MetaboMiner is largely restricted to analyzing spectra from biofluids or metabolite mixtures that are titrated to pH 7.0 +/- 0.5.

Conclusions

In this chapter we have demonstrated that by utilizing the extra information found in 2D NMR spectra as well as prior knowledge about the composition of the biofluid itself, it is possible to semi-automatically identify a significant number of compounds in complex aqueous mixtures (both defined mixtures and biofluids) with excellent (>80%) accuracy. In particular, the quality and degree of metabolite identification achieved by MetaboMiner certainly matches that of what

a skilled NMR spectroscopist could do - but in significantly less time. Overall, we have shown that by using a comprehensive reference library coupled with robust algorithms for peak processing, peak matching and compound identification, the process of metabolite identification from 2D NMR spectra can be greatly simplified.

Figures

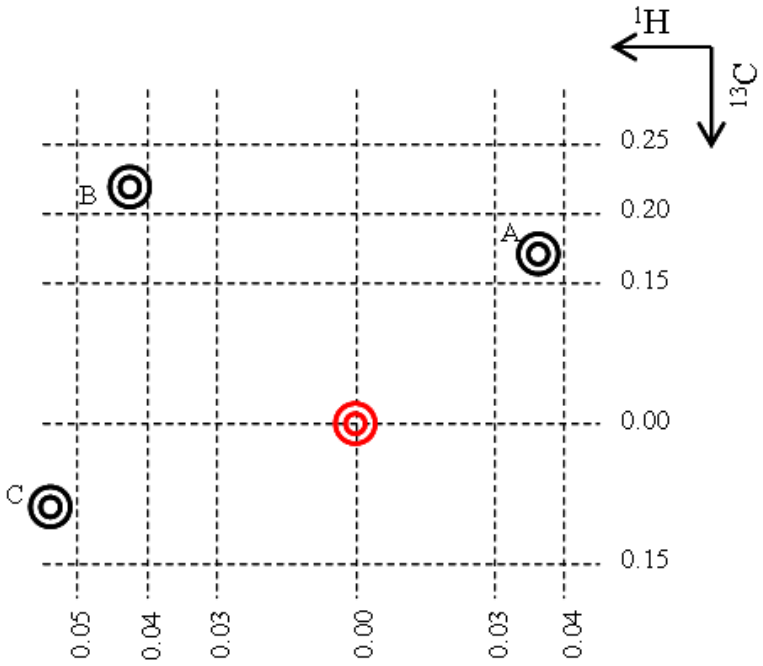

Figure 2.2 An illustration of the calculation of uniqueness values

The red peak represents the peak of interest and three peaks are in its immediate vicinity. The calculations are only performed at five chemical shifts distance levels - 0.01, 0.02, 0.03, 0.04, 0.05 ppm along the ^{1}H dimension, and 0.05, 0.10, 0.15, 0.20, 0.25 ppm along the ^{13}C dimension. No peak is observed in the first three distance levels. So the maximum unique scope for this peak is (0.03, 0.15) ppm. Peak A is found within 0.03~0.04 ppm (^{1}H dimension) and 0.15~0.20 ppm (^{13}C dimension) of the red peak; Peak B is found within 0.04~0.05 ppm (^{1}H dimension) and 0.20~0.25 ppm (^{13}C dimension) of the red peak; Peak C is not considered since the chemical shift distance is more than 0.05 ppm along the ^{1}H dimension. Therefore, the assigned uniqueness values are 0-0-0-1-2. Note that the distance is not drawn to scale for illustration purposes.

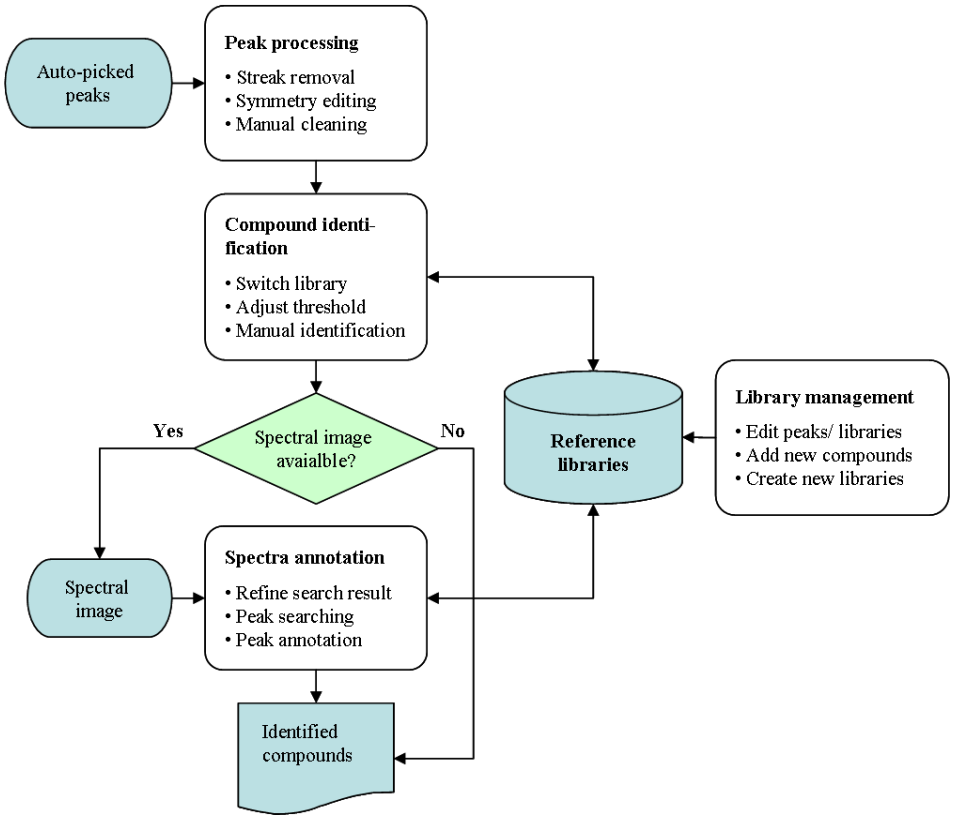

Figure 2.3 MetaboMiner flowchart

The query peaks obtained from an automatic peak-picking program are first processed to remove streaks and other artefacts. The cleaned peak list is then scanned for the presence of peak patterns of compounds in a spectral reference library corresponding to the biofluid that has been identified by the user. Spectral images can be used to further refine the search result.

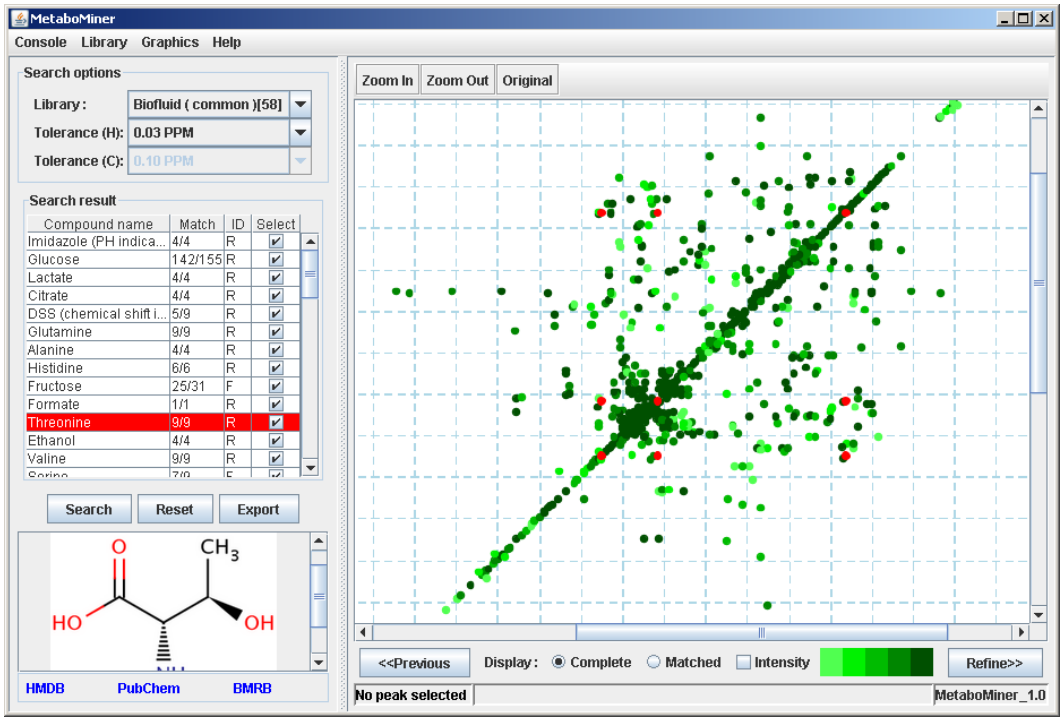

Figure 2.3 Screenshot of MetaboMiner's "Search View".

The left panel shows the library compounds that have matches in the query peaks. The selected checkbox indicates the corresponding compound is considered to be present by MetaboMiner. On the right panel, the reference peaks (in red) of the current selected compound is displayed with query peaks as background. The color variations represent the peak intensities with the dark green corresponding to the strongest peak intensities. When the mouse is placed over any synthetic peak, all its information (name, position, uniqueness values, *etc.*) will be displayed on the view panel. Right clicking on any peak will allow users to search the spectral library for this particular peak.

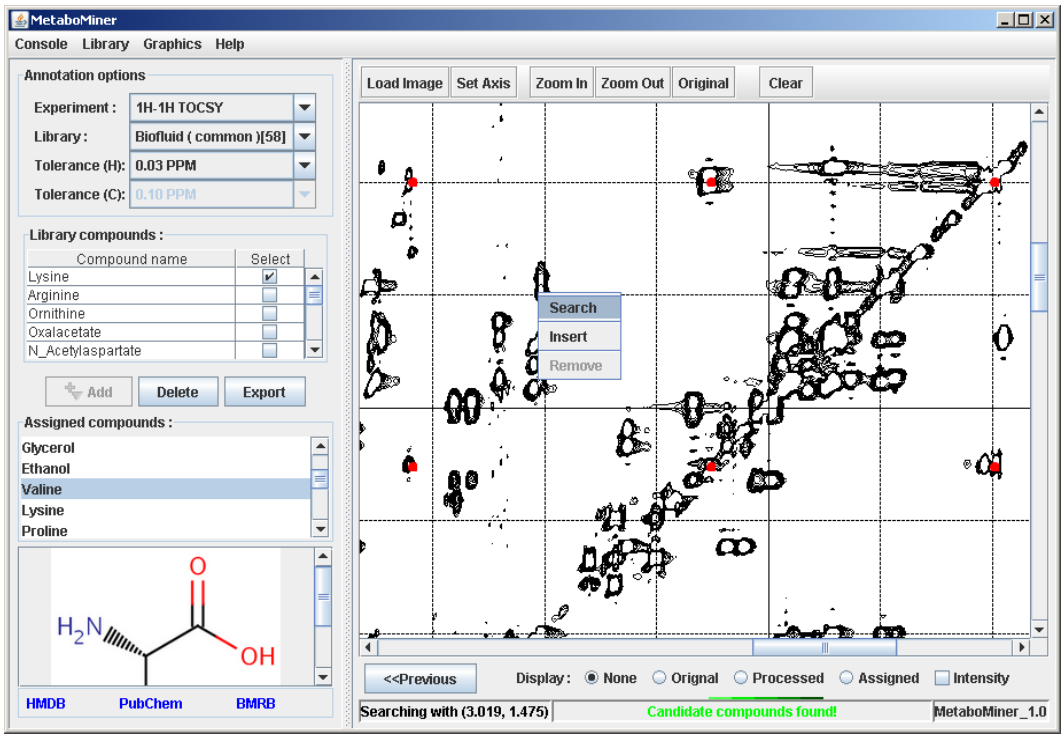

Figure 2.4 Screenshot of MetaboMiner's "Annotation View".

The contents of the reference spectral library and the identified compound list are shown on the left panel. The spectral image is displayed on the right panel. The red peaks correspond to the current compound being annotated (Lysine). Peak searching is carried out by right clicking on a corresponding Lysine peak. The user can also directly edit the current compound by inserting, removing, or dragging its peaks to match the exact pattern of the reference spectrum.

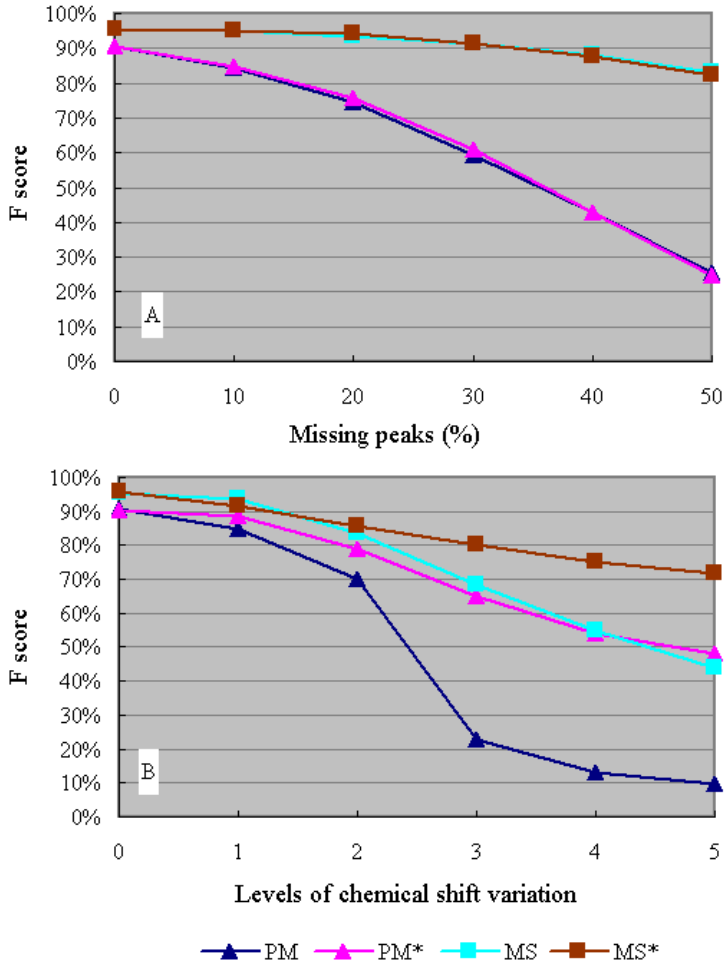

Figure 2.5 Comparative performances of different search strategies.

Synthetic mixture query spectra were generated by pooling the peaks of 50 randomly selected compounds from the reference spectral library. Different levels of spectral noise were added to these peaks and then compounds were identified with (*) and without using the adaptive threshold method. The Figure 5A, the query peaks were deleted at random with 0%, 10%, 20%, 30%, 40% and 50% probabilities; Figure 5B, the query peaks were subject to five levels of random chemical shift variations (±0.01, ±0.02, ±0.03, ±0.04, ±0.05 ppm for each ^{1}H chemical shift, and ±0.05, ±0.10, ±0.15, ±0.20, ±0.25 ppm for each ^{13}C chemical shift). The F scores were averaged over 50 iterations. (Abbreviations: PM, percentage match method; MS, minimal signature method).

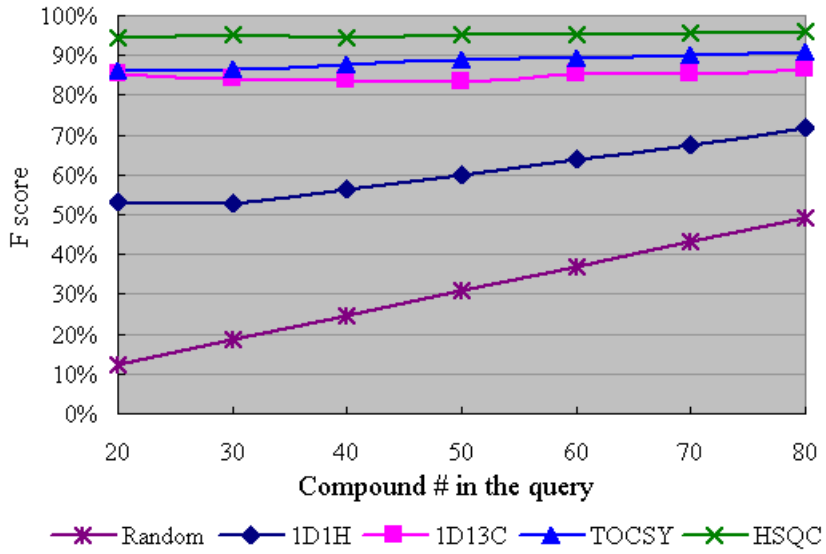

Figure 2.6 Evaluation of MetaboMiner using simulated datasets

Synthetic mixture query spectra were generated by pooling peaks from 20, 30, 40, 50, 60, 70, and 80 compounds randomly selected from MetaboMiner's spectral library. Spectral noise was introduced via random (10%) peak deletion and random chemical shift changes within ±0.01 ppm for each ^{1}H chemical shift, and within ±0.05 ppm for each ^{13}C chemical shift. Compound identification was based on minimal signatures using the adaptive threshold method. The F-measures were averaged over 50 iterations. The random results were calculated as the numbers of true positive hits selected by random chance using the same thresholds.

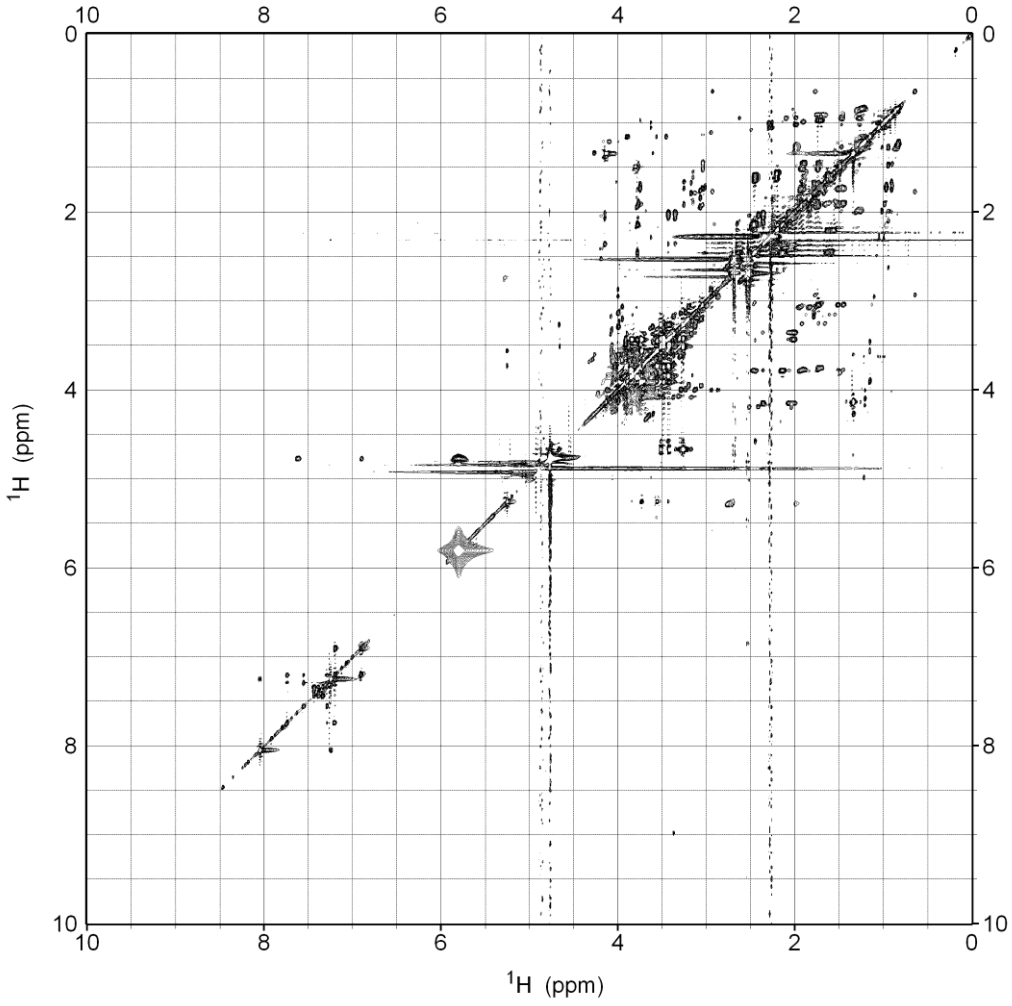

Figure 2.7: An example of a TOCSY spectrum for a biofluid mixtures.

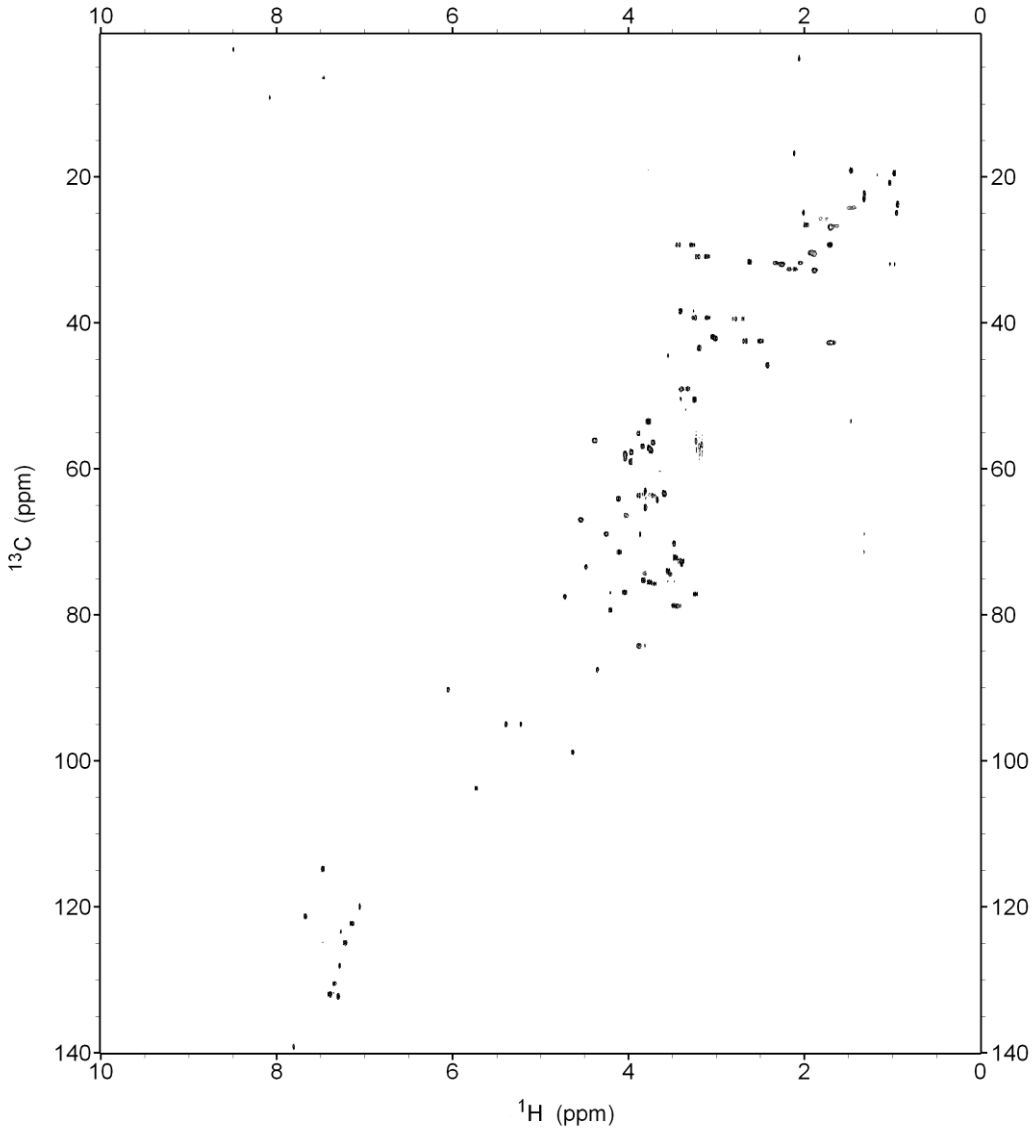

Figure 2.8: An example of a ^{1}H-^{13}C HSQC spectrum for a biofluid mixture.

Tables

Table 2.1A: Performance evaluation using HSQC data collected at pH ~7.2.

Samples A, B, and C are synthetic cocktail mixtures and sample D is a plasma sample. (Annotation: MetaboMiner-sp = searched using the biofluid-specific library; MetaboMiner-all = searched using the entire spectral library; TP = true positives; FN = false negatives; FP = false positives)

Method	Sample	# Cmpds	TP	FN	FP	Precision (%)	Recall (%)	F score
MetaboMiner-sp	A	27	21	6	6	77.8	77.8	77.8
MetaboMiner-all	A	27	15	12	8	65.2	55.6	60.0
HMDB	A	27	8	19	19	29.7	29.7	29.7
MMCD	A	27	8	19	6	57.1	29.7	39.1
BMRB	A	27	6	21	21	22.2	22.2	22.2
PRIMe	A	27	14	13	6	70.0	51.9	59.6
MetaboMiner-sp	B	21	16	5	5	76.2	76.2	76.2
MetaboMiner-all	B	21	9	12	0	100	42.9	60.0
HMDB	B	21	1	20	20	4.8	4.8	4.8
MMCD	B	21	0	21	0	0	0	0
BMRB	B	21	1	20	20	4.8	4.8	4.8
PRIMe	B	21	6	15	4	60.0	28.6	38.7
MetaboMiner-sp	C	24	22	2	2	91.7	91.7	91.7
MetaboMiner-all	C	24	16	8	3	84.2	66.7	74.4
HMDB	C	24	9	15	15	37.5	37.5	37.5
MMCD	C	24	2	22	0	100	7.4	13.8
BMRB	C	24	3	21	21	12.5	12.5	12.5
PRIMe	C	24	8	16	5	61.5	33.3	43.2
MetaboMiner-sp	D	35	29	6	6	82.9	82.9	82.9
MetaboMiner-all	D	35	16	19	7	69.6	45.7	55.2
HMDB	D	35	9	26	26	25.7	25.7	25.7
MMCD	D	35	4	31	3	57.1	11.4	19.0
BMRB	D	35	7	28	28	20.0	20.0	20.0
PRIMe	D	35	14	21	5	73.7	40.0	51.8

Table 2.1B Performance evaluation using TOCSY data collected at pH ~7.2.

Samples A, B, and C are synthetic cocktail mixtures and sample D is a plasma sample. (Annotation: MetaboMiner-sp = searched using the biofluid-specific library; MetaboMiner-all = searched using the entire spectral library; TP = true positives; FN = false negatives; FP = false positives).

Method	Sample	# Cmpds	TP	FN	FP	Precision (%)	Recall (%)	F score
MetaboMiner-sp	A	27	23	4	4	85.2	85.2	85.2
MetaboMiner-all	A	27	21	6	6	77.8	77.8	77.8
HMDB	A	27	2	25	25	7.4	7.4	7.4
MMCD	A	27	1	26	26	3.7	3.7	3.7
MetaboMiner-sp	B	21	16	5	5	76.2	76.2	76.2
MetaboMiner-all	B	21	12	9	2	85.7	57.1	68.5
HMDB	B	21	2	19	19	9.5	9.5	9.5
MMCD	B	21	2	19	19	9.5	9.5	9.5
MetaboMiner-sp	C	24	17	7	7	70.8	70.8	70.8
MetaboMiner-all	C	24	15	9	8	65.2	62.5	63.8
HMDB	C	24	4	20	20	16.7	16.7	16.7
MMCD	C	24	2	22	22	8.3	8.3	8.3
MetaboMiner-sp	D	35	30	5	5	85.7	85.7	85.7
MetaboMiner-all	D	35	23	12	12	65.7	65.7	65.7
HMDB	D	35	3	32	32	8.6	8.6	8.6
MMCD	D	35	2	33	33	5.7	5.7	5.7

Table 2.2 Performance evaluation of MetaboMiner under different pH conditions.

Sample	# Cmpds	NMR Exp.	pH	TP	FN	FP	Precision (%)	Recall (%)	F score
A	27	HSQC	4.2	15	12	12	55.6	55.6	55.6
			7.2	21	6	6	77.8	77.8	77.8
		TOCSY	4.2	21	6	6	77.8	77.8	77.8
			7.2	23	4	4	85.2	85.2	85.2
D	35	HSQC	7.3	29	6	6	82.9	82.9	82.9
			8.8	24	11	4	85.7	68.6	76.2
		TOCSY	7.3	30	5	5	85.7	85.7	85.7
			8.8	25	10	6	80.6	71.4	75.7

Note: Tests were performed using biofluid-specific library

.

Chapter 3

A Metabolomics Data Analysis Pipeline[2]

[2] A version of this chapter has been publisher previously

Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res,* 37, W652-660.

Introduction

Over the past decade, many robust algorithms and programming tools have been developed for analyzing high-dimensional data produced from transcriptomic or microarray studies. In particular, R (http://www.r-project.org) and the Bioconductor project (93) probably represent the most complete collection of up-to-date statistical and machine learning algorithms for microarray data analysis. However, programming in R or using command-line R programs is particularly challenging for bench biologists – especially for those without computer or programming experience. To address this issue, many bioinformatics tools have been developed to provide user-friendly interfaces based on R. Among these tools, the most visible is probably the GenePattern server (182) from the Broad Institute (http://www.broad.mit.edu/genepattern). GenePattern is an open source and web-based tool for genomics and microarray data analysis. The three key concepts behind GenePattern's programming design are: 1) *Comprehensive Analysis and Visualization*; 2) *Pipeline* and 3) *Servers*.

Data from metabolomics and microarray gene expression experiments share a great deal in common. For example, both kinds of studies aim to identify important features associated with certain conditions (biomarker discovery) or to predict these conditions based on the measured data (classification). Furthermore the data matrices generated by both metabolomics and transcriptomics experiments are characterized by a high-dimensional feature space and a relatively small sample size. The differences are mainly in the procedures used for

data processing and data annotation. Therefore, by utilizing the available resources originally developed for microarray analysis, coupled with several new algorithms specific for metabolomics, it is possible to develop an efficient pipeline optimal for metabolomics data processing and analysis. Following this approach, I have created MetaboAnalyst – a web-based server for processing, analyzing, visualizing and annotating high throughput metabolomic data. I have designed MetaboAnalyst so that it is able to process a wide variety of metabolomic data types including compound concentration tables (for quantitative metabolomics) as well as spectrally binned data, NMR/MS peak lists and GC/LC-MS spectra (NetCDF, mzXML, mzDATA – for chemometric metabolomics). It also provides a comprehensive list of analysis options for normalization, feature identification, dimensional reduction clustering and classification. Furthermore, MetaboAnalyst produces colorful graphical output and it supports a number of compound identification and pathway mapping tools for data annotation. The remaining sections of this chapter provide information about the implementation of MetaboAnalyst, followed by detailed descriptions of its analysis features. The chapter ends with a discussion and comparison with related software tools.

Implementation

MetaboAnalyst's web interface was developed using Java Server Faces (JSF) technology (http://java.sun.com/javaee/javaserverfaces). The backend statistical computing and visualization operations were carried out using functions from the R and Bioconductor packages. The integration between Java and R was

established through the Rserve package (http://www.rforge.net/Rserve). Spectral matching and pathway identification software was developed in Java using the spectral libraries and pathway libraries developed for the Human Metabolome Project (21) and MetaboMiner (183).

JSF is a very powerful technology for developing Java-based web applications. It is designed to simplify the development of user interfaces for Java Enterprise Edition (Java EE) applications by automatic handling of low level HTTP requests and user input processing. JSF uses a component-based model for web development. Using the visual JSF web application tool offered by the NetBeans (http://www.netbeans.org) integrated development environment (IDE), components can be literally "painted" on a virtual JSF page by dragging-and-dropping them from a palette of JSF component library. Event handlers can then be defined for each component the same way as for developing standalone Java graphic user interface (GUI) application. Finally, navigation rules are specified for each page from a central XML configuration file (*faces-config.xml*). User actions on a web interface will trigger an event whose return value determines which page is to be displayed subsequently based on the navigation rules specified for that page. This approach facilitates modular and flexible design, making web application development much simpler and faster.

MetaboAnalyst consists of several functional modules that will be discussed, in detail, later. These functions are carried out by several R scripts and Bioconductor function calls. Detailed information about the individual packages used and the R scripts can be downloaded from the MetaboAnalyst home page.

When MetaboAnalyst is run, the executed R commands are recorded to a temporary text file. During the summary report generation, this R command history is examined and the last call for each analysis performed is re-evaluated using the R *Sweave* function that executes the R commands and writes text descriptions along with tabular and graphical results into a LaTeX file. Finally, the file is converted into a PDF report describing the analysis, which is available to the user for download.

MetaboAnalyst is currently hosted on GlassFish (https://glassfish.dev.java.net) installed on a Linux operating system (Fedora Core 10). The server is equipped with two Intel Pentium 4 processors (2.8 GHz each) and 4 GB of physical memory. The web application is platform independent and has been tested successfully under both Linux and Windows operating systems. R (version 2.8.0) is currently installed on the same machine with latest Bioconductor release 2.3 and Rserve 0.5-2.

A diagram illustrating MetaboAnalyst's workflow is shown in **Figure 3.1**. MetaboAnalyst is not a "single-click" analysis tool, but rather it is an on-line analysis pipeline similar in concept to several existing on-line microarray analysis tools such as GEPAS (184) and CARMAweb (185). It is primarily designed to allow users to conduct two-group discriminant analysis (i.e. control vs. non-control - the most common type of metabolomic analysis) for classification and "significant feature" identification. MetaboAnalyst also supports both paired and unpaired data analyses. A typical MetaboAnalyst run consists of six steps – 1) data upload, 2) processing, 3) normalization, 4) statistical analysis, 5) annotation,

and 6) summary report download. Users are guided through these steps by MetaboAnalyst's intuitive interface and the navigation bar on the left panel of each page. Completed steps are indicated by a change in color. Certain downstream analyses may not be allowed depending on the context or type of analyses previously performed. Detailed descriptions, help files, and helpful hints are either shown on the corresponding web pages or are provided as mouse-over pop-up balloons. This support is further enhanced by the availability of several step-by-step tutorials, sample data sets (NMR, GC/LC-MS, binned data, etc.), sample summary files and frequently asked questions (FAQs) available on MetaboAnalyst's web site.

Step 1: Data Upload

Users can begin a MetaboAnalyst analysis by pressing the "Click Here to Start" link on the MetaboAnalyst's home page. This takes users to the data upload page. Because there is no widely-accepted standard format for reporting metabolomics experiments MetaboAnalyst has been designed to accept diverse data types including compound concentration tables (from quantitative metabolomic studies), binned spectral data, NMR or MS peak lists, as well as raw GC-MS and raw LC-MS spectra. For compound concentration or binned spectral data, MetaboAnalyst requires that they be uploaded as a CSV (comma separated values) table with class labels (control and abnormal, say) immediately following the sample names. For peak list data, MetaboAnalyst requires that they be uploaded as two zipped folders containing peak list files from the two respective groups. Each file should be a two or three-column CSV list indicating peak

positions (chemical shift for NMR peaks, mass and/or retention time for MS peaks) and intensities, respectively. Examples of these formats and more detailed explanations of the formatting requirements are provided on the MetaboAnalyst home page. Vendor-specific, proprietary GC-MS or LC-MS spectra should be first converted to open exchange file formats (NetCDF, mzXML, mzDATA) and uploaded as two zipped folders corresponding to the two groups being analyzed. Detailed instructions on how to specify paired information (for paired data analysis) as well as examples for each data type are available through MetaboAnalyst's "Data Formats" link on the home page.

Step 2: Data Processing and Data Integrity Checking

Depending on the type of uploaded data, different processing strategies can be employed to convert the raw numbers into a data matrix suitable for downstream analysis. For compound concentration lists, the data can be used immediately after MetaboAnalyst's data integrity check. For binned spectral data, a linear filter is first applied in order to remove baseline noise. This is done because most data processing algorithms do not work properly with many near-zero values. For NMR and/or MS peak lists, MetaboAnalyst first groups the peaks across all samples based on their positions. For GC-MS and LC-MS spectra or total ion chromatograms, the program performs peak detection, peak grouping, and retention time correction sequentially using the popular XCMS package (112). Users can adjust the default parameters for each processing step.

Often there are large numbers of missing values in a typical quantitative metabolomics dataset (10 - 40% in our experience). Most of these missing values are due to various compounds in certain samples being below the instrument detection limits. In untargeted approaches, however, other factors may come into play when missing values are introduced during spectral processing and feature detection. To allow selected analyses to proceed (i.e. without divide-by-zero problems), these missing values are replaced by the half of the minimum value found in the dataset by default. We also implemented a variety of methods which enable users to manually or automatically perform missing value exclusion, missing value replacement, as well as missing value imputation (186,187). In addition, as part of the data integrity check, MetaboAnalyst also checks class labels and pair specification (if applicable) to make sure all the required information is present and consistent before proceeding to the next step.

Step 3: Data Normalization

At this stage, the uploaded data is compiled into a table in which each sample is formally represented by a row and each feature identifies a column. With the data structured in this format, two types of data normalization protocols - row-wise normalization and column-wise normalization -- may be used. These are often applied sequentially to reduce systematic variance and to improve the performance for downstream statistical analysis. Row-wise normalization aims to normalize each sample (row) so that it is comparable to the other. Four commonly used metabolomic normalization methods have been implemented in MetaboAnalyst, including normalization to a constant sum, normalization to a

reference sample (probabilistic quotient normalization) (114), normalization to a reference feature (creatinine or an internal standard) and sample-specific normalization (dry weight or tissue volume). In contrast to row-wise normalization, column-wise normalization aims to make each feature (column) more comparable in magnitude to the other. Four widely-used methods are offered in MetaboAnalyst - log transformation, auto-scaling, Pareto scaling, and range scaling. Given the vast dynamic range of many features (compound concentration or ion abundance) in metabolomics data, normalization is highly recommended. The effects and utility of these different normalization strategies have been discussed in detail elsewhere (115) and are described further in MetaboAnalyst's online tutorials.

Step 4: Data Analysis

MetaboAnalyst's data analysis module is a collection of well-established statistical and machine learning algorithms that have been shown to be particularly robust for high-dimensional data analysis. These algorithms are organized into five analysis "paths" for users to explore.

a) Univariate Analysis Path. Because of their simplicity and interpretability, univariate analyses are often first used to obtain an overview or rough ranking of potentially important features before applying more sophisticated analyses. Univariate analysis examines each variable separately and does not consider the effect of multiple comparisons. MetaboAnalyst's univariate analysis path supports three commonly used methods - fold-change analysis, t-tests, and volcano plots.

In a t-test, one attempts to determine whether the means of two groups are distinct. Once a t-value (refer to formula (1) in section 1.5.4) is determined, a p-value can be calculated that can be used to determine whether this distinction is statistically significant. Both paired (same individuals measured before and after an intervention) and unpaired (individuals randomly assigned to two groups) analyses are supported. Volcano plots are used to compare the size of the fold change to the statistical significance level. The horizontal axis plots the fold change between the two groups (on a log scale), while the vertical axis represents the p-value for a t-test of differences between samples (on a negative log scale).

b) Chemometric Analysis Path. This analysis path offers the two most commonly used chemometric methods – principal component analysis (PCA) and partial-least squares discriminant analysis (PLS-DA). PCA is an unsupervised method aiming to find the directions of maximum variance in a data set (X) without referring to the class labels (Y). PLS-DA is a supervised method that uses multiple linear regression technique to find the direction of maximum covariance between a data set (X) and the class membership (Y). For both methods, the original variables are summarized in many fewer variables using their weighted averages. These new variables are called scores. The weighting profiles are called loadings. MetaboAnalyst provides various views commonly used for PCA and PLS-DA analysis. Users can specify each axis to view the patterns between different components. Both two-dimensional (2D) and three-dimensional (3D) views are implemented. A 3D PLS-DA score plot is shown in **Figure 3.2A**.

As a supervised method, PLS-DA can perform both classification and feature selection. The algorithm uses cross-validation to select an optimal number of components for classification. Two feature importance measures are commonly used in PLS-DA. Variable importance in projection or VIP score is a weighted sum of squares of the PLS loadings. The weights are based on the amount of explained Y-variance in each dimension. The other importance measure is based on the weighted sum of PLS-regression coefficients. The weights are a function of the reduction of the sums of squares across the number of PLS components. Both importance measures are implemented in PLS-DA analysis for selecting important features. MetaboAnalyst's implementation of PLS-DA also supports several options for cross-validation including leave-one-out (LOOCV) and 10-fold cross validation. We also implemented PLS-DA permutation tests to help user determine the importance of class separation (188).

c) Feature Selection Path. This analysis path provides two well-established methods widely used for identification of differentially expressed genes in microarray experiments - Significance Analysis of Microarrays (and Metabolites) (SAM) (117) and Empirical Bayesian Analysis of Microarrays (and Metabolites) (EBAM) (189). However, these methods are very general for identification of significant features in high-dimensional data and are not restricted to the analysis of microarray data. SAM is designed to address false discovery rate problems (FDR) when running multiple tests on high-dimensional data. It first assigns a significance score to each variable based on its change relative to the standard deviation of repeated measurements. Then it chooses variables with scores greater

than an adjustable threshold and compares their relative difference to the distribution estimated by random permutations of the class labels. For each threshold, a certain proportion of the variables in the permutation set appear to be significant by chance. The number is used to calculate the FDR. In this way, SAM is able to perform permutation testing, something that is not done in MetaboAnalyst's t-tests. The EBAM algorithm is essentially a variation of the SAM method. The only difference is that EBAM uses a modified t-statistic in calculating the score. Typical SAM and EBAM plots are provided to assist users in choosing the best parameters and viewing the results. Tables containing numeric details are also available through hyperlinks in addition to these graphical presentations. A SAM plot is shown in **Figure 3.2B**.

d) Cluster Analysis Path. MetaboAnalyst's cluster analysis allows a closer interrogation of samples with similar abundance profiles. This path includes two major approaches of clustering analysis - hierarchical clustering and partitional clustering. Hierarchical (agglomerative) clustering begins with each sample considered as separate cluster and then proceeds to combine them until all samples belong to one cluster. A variety of dissimilarity measures (Euclidean distance, Pearson's correlation, and Spearman's rank correlation) and clustering methods (average linkage, complete linkage, single linkage, and Ward's linkage) have been implemented in MetaboAnalyst. The result of hierarchical clustering is usually presented as a dendrogram or heat map, both of which are available in MetaboAnalyst. A heat map view is presented in **Figure 3.2C** using one of our test data sets. Partitional clustering attempts to directly decompose the data set

into a user-specified number of disjoint clusters. Two widely used methods, k-means clustering and self-organizing maps (SOM) have been implemented in MetaboAnalyst. K-means clustering aims to create k clusters such that the sum of squares from points to the assigned cluster centers' is minimized. SOM is an unsupervised neural network based around the concept of a grid of interconnected nodes, each of which contains a model. The model clusters begin as random values, but during the iterative training process, they are updated to represent different subsets of the training set. Users indicate the number of clusters by specifying the expected dimension of the grid. The clusters from both k-means and SOM are presented as aggregated expression profiles in which samples in each cluster are plotted as line graphs on top of each other using their feature values.

e) Supervised Classification Path. Class prediction using metabolomics data is increasingly important in studies aiming for early diagnosis, prognosis or treatment outcomes. MetaboAnalyst offers three powerful supervised classification methods - PLS-DA, random forest (130), and support vector machine (SVM). These methods have proved to be robust for high-dimensional data and are widely used for other 'omics' data analysis. In addition, they can also help prioritize features that contribute significantly to the performance.

PLS-DA based feature selection and classification was discussed in the chemometrics path. Random forest uses an ensemble of classification trees, each of which is grown by random feature selection from a bootstrap sample at each branch. Class prediction is based on the majority vote of the ensemble. During

tree construction, about one-third of the instances are left out of the bootstrap sample. This data is then used as test sample to obtain an unbiased estimate of the classification (OOB) error. Variable importance is evaluated by measuring the increase of the OOB error when it is permuted. **Figure 3.2D** shows the important features ranked by random forest. The SVM classification algorithm aims to find a decision function in the input space by mapping the data into a higher dimensional feature space and separating it by means of a maximum margin hyperplane (190). MetaboAnalyst's SVM analysis is done through recursive feature selection and sample classification using a linear kernel (136). Features are selected based on their relative contribution in the classification using cross validation error rates. The least important features are eliminated in the subsequent steps. This process creates a series of SVM models. The features used by the best model are considered to be important and are ranked by their frequencies of being selected in the model.

Step 5: Data Annotation

A key step in placing statistically significant findings from chemometric analyses (as opposed to quantitative metabolomic analyses) into a biological context is to identify significantly altered compounds represented by certain spectral bins or certain clusters of spectral peaks. Once a user has identified lists of MS or NMR peaks that exhibit statistically significant changes, he may use one of several spectral comparison routines and spectral libraries to attempt to identify the compound(s) based on either lists of MS peaks (from MS or MS/MS data), GC-MS peaks (from EI mass values and retention indices) or NMR peaks (from ^{1}H,

^{13}C or heteronuclear NMR spectra). These compound identification routines and spectral reference libraries were originally developed for the HMDB and for MetaboMiner (183). While not as comprehensive as some commercial libraries or commercial software, these freely available tools have been shown to be quite useful in identifying many common compounds. Once compound information becomes available (via quantitative routes or via MetaboAnalyst's metabolite ID software), more insight can be obtained by which metabolic pathways are involved. Pathway mapping has been implemented in MetaboAnalyst using more than 70 pathway diagrams and metabolite libraries derived from the HMDB. Users simply type the names (or synonyms) of the metabolites identified and MetaboAnalyst provides the list of pathways in which these metabolites are found, along with hyperlinks to their pathway images. All results are linked to the HMDB where users can obtain more detailed information for each metabolite or pathway.

Step 6: Summary Report Download

When users finish their analyses and click the download link, a comprehensive report will be generated containing a detailed description of each step performed embedded with graphical and tabular outputs. In addition, the processed numeric data, high-resolution images (PNG format), R scripts, as well as the R command history are also available for downloading. Users familiar with R can easily reproduce the results on their local machine after installation of R and the required packages. Users have the option of providing an email address (to which the summary report is sent) or simply downloading the compressed file that contains

all the data (graphs, tables, etc.) produced during the analysis. A sample summary report is available for download from MetaboAnalyst's homepage. MetaboAnalyst creates a temporary folder for each job it received. The folder will remain on the server for three days (72 hours) before being automatically deleted.

Tutorials and Sample Data Sets

The inherent complexity of many data processing techniques combined with lack of familiarity that many users may have with some of the analytical approaches used by MetaboAnalyst led us to develop a number of tutorials and sample data sets. This was also done so that new users could become more familiar with MetaboAnalyst's expected inputs and outputs. Under the "Try our test data" in the data upload window, users will find eight different data sets labeled as 1) Concentrations (a metabolite concentration table); 2) NMR spectral bins; 3) NMR peak lists; 4) Concentrations (paired, time series); 5) MS peak intensities; 6) MS peak lists; 7) LC-MS spectra in NetCDF format; and 8) GC-MS spectra in NetCDF format. Users may process these data by clicking on the radio button beside a given data set and pressing the Submit button. Alternately, these example data sets can be downloaded and subsequently "uploaded" using the "Upload your data" section. Once a test data set is submitted (or uploaded) the user may navigate through MetaboAnalyst in any way they choose.

MetaboAnalyst also has four step-by-step tutorials describing several analysis paths using a number of different data sets. These tutorials are available by clicking the "Tutorials" link from homepage. Tutorial #1 uses the Metabolite

concentration list (data set #1). Tutorial #2 uses the Binned NMR spectra (data set #2), Tutorial #3 uses the paired concentration data (data set #4) and Tutorial #4 uses the LC-MS spectra in NetCDF format (data set #7). MetaboAnalyst also has ~20 FAQs to complement the information found in the tutorials. These tutorials and FAQs will be updated frequently based on user feedback.

Comparison to Other Software and Limitations

Many metabolomic analyses are currently done using local installations of commercial statistical software packages such as MatLab, MS-Excel, SigmPlot and SIMCA-P. SIMCA-P (Umetrics), in particular, is very widely used by the metabolomics community. While quite expensive, SIMCA-P offers excellent graphic capabilities and comprehensive analysis options for three multivariate methods (PCA, PLS/OPLS, and SIMCA). MetaboAnalyst supports two of these multivariate methods (PCA and PLS) but it also offers many other methods (i.e. volcano plots, SAM, k-means, SOM, random forest, SVM) not found in SIMCA-P. While MetaboAnalyst does not have the graphical flexibility of SIMCA-P, it is designed to be more accessible (via the web), freely available, and easier to use. In addition, MetaboAnalyst provides its own metabolite and pathway identification tools – something that is not found in any dedicated statistical software package. However, MetaboAnalyst's dependence on the HMDB infrastructure means that its coverage of plant and microbial metabolism is somewhat incomplete.

To the best of our knowledge, the only other web application that offers a

similar service to MetaboAnalyst is MeltDB (191). MeltDB is centered on MS-based metabolomics data storage, administration, analysis, and annotation. Unfortunately, this server is incompatible with a number of common browsers (Firefox, Netscape) and requires a user login and password to obtain access. According to the paper, MeltDB appears to offer some of the features found in MetaboAnalyst such as t-tests, volcano plots, principal component analysis, and heat maps. However, these analyses are restricted to GC/LC-MS data only. MetaboAnalyst provides support for many more diverse data types, more advanced data analysis methods, more comprehensive data annotation tools as well as automated report generation utilities.

The current implementation of MetaboAnalyst primarily supports 1) biomarker discovery and 2) two-group discrimination. We believe these kinds of analyses are most relevant to the widest range of metabolomics studies. Multiclass problems can always be converted into a series of two-class problems through pair-wise decomposition. Temporal studies (more than two time points) can be treated as a special case of multi-class problem and decomposed into a series of paired two-group analyses (see Tutorial #3 for an example of a time series analysis). We hope to add more functions to support simultaneous analysis of multiple time-points in the near future.

MetaboAnalyst makes extensive use of high-level data inputs (i.e. concentrations, peak lists) that requires users to perform some manual processing steps prior to uploading the data. The support for raw or partially processed GC-MS and LC-MS spectra is currently achieved through the XCMS package (112).

However, software to handle unprocessed NMR spectra is not so readily available. Steps such as phasing, baseline correction, referencing, peak detection and deconvolution must be manually checked by an experienced analyst to ensure the integrity of the results. As a result, MetaboAnalyst does not accept raw NMR spectra. Likewise, MetaboAnalyst is not (yet) capable of handling or interpreting capillary electrophoretic (CE) data, FTIR data, coulometric electrode array (CEA) data or raw chromatographic (HPLC or UPLC) data. Certainly if the user community grows significantly in these areas, efforts will be made to accommodate these analytical platforms. Indeed, MetaboAnalyst's modular and flexible framework should facilitate future development efforts to keep up with this fast-changing field.

Conclusions

MetaboAnalyst is a comprehensive, web-based tool designed to facilitate high-throughput metabolomics studies. It accepts a variety of input data (NMR peak lists, binned NMR or MS spectra, MS peak lists, compound/concentration data) in a wide variety of formats. It also offers a number of options for metabolomic data processing, data normalization, multivariate statistical analysis, graphing, metabolite identification and pathway mapping. Through its intuitive interface and high quality graphics, users are presented with data overviews from different perspectives (i.e. PCA plots, heat maps), lists of candidate biomarkers identified by simple univariate analysis (i.e. volcano plots), as well as estimated classification performances by several powerful algorithms (i.e. random forest,

SVM). Further biological insight can be gained by tapping into the HMDB using MetaboAnalyst's annotation tools. MetaboAnalyst's structured navigation, extensive documentation, as well as its comprehensive analysis reports should allow new users to analyze their data without significant training or without significant likelihood of statistical misadventure.

Figures

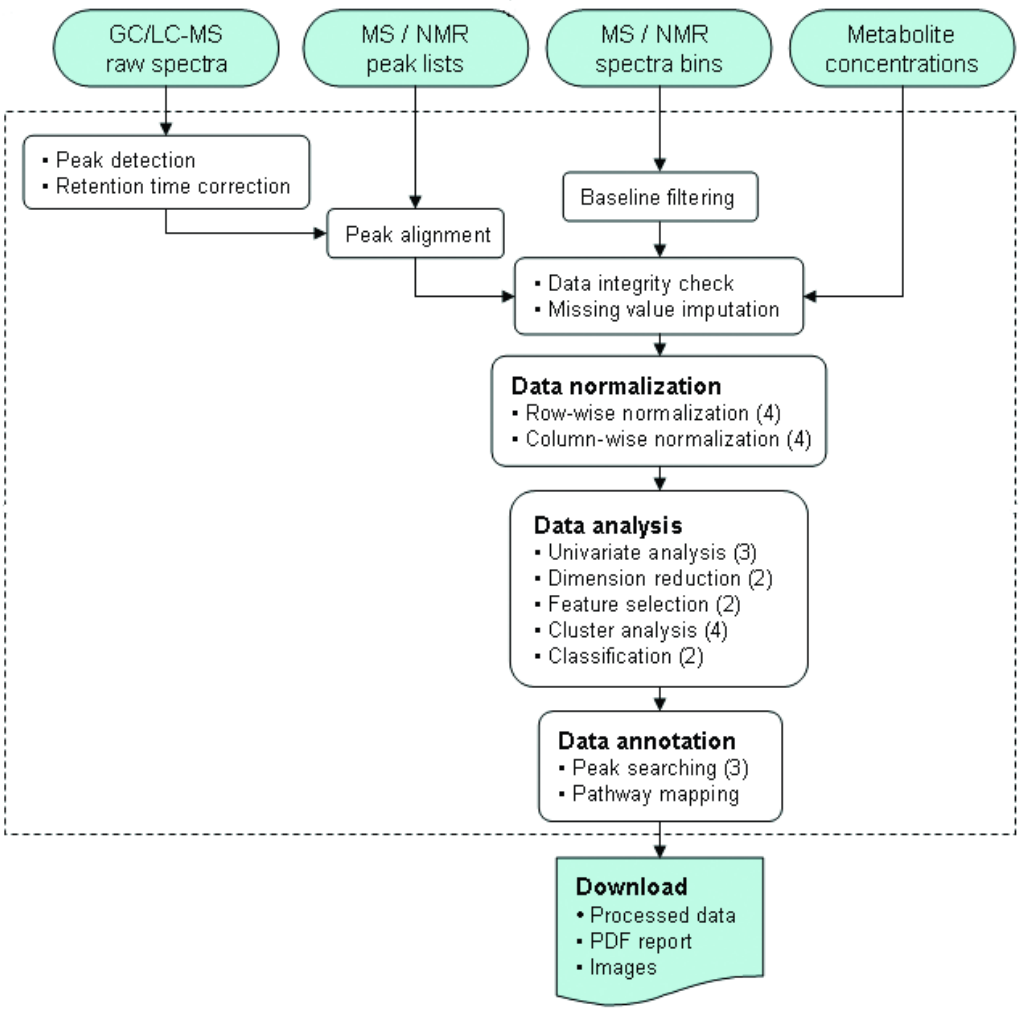

Figure 3.1. MetaboAnalyst's workflow and data processing options.

Different data inputs are first transformed into compatible data matrices using several different processing methods. A variety of algorithms are implemented for data normalization, analysis, and annotation. The number of available options is shown inside the round brackets for each category. At the end of any given analysis, a comprehensive PDF report, the processed data, and high-resolution images are available for download.

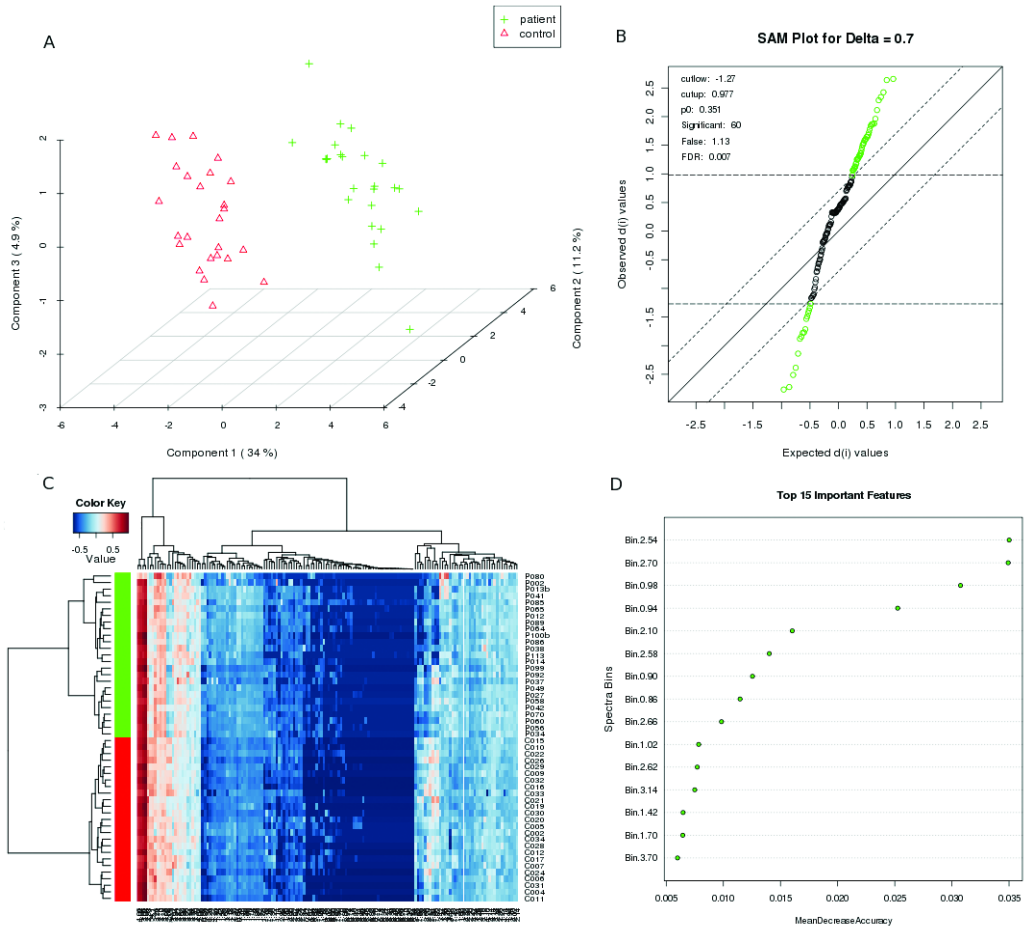

Figure 3.2 Examples of some graphical outputs from MetaboAnalyst.

Figure A shows the PLS-DA class separation based on the top three components. Figure B shows the significant features identified by SAM analysis. Figure C shows the heat map generated from hierarchical clustering. Figure D shows the features ranked by random forest. The binned NMR spectral data (test data #2) was used to generate these graphs.

Chapter 4

Metabolite Set Enrichment Analysis[3]

[3] A version of this chapter has been published previously:

Xia, J. and Wishart, D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*, 38, W71-77.

Introduction

As reviewed in Section 1.5.5, gene set enrichment analysis (GSEA) (142) has turned out to be very successful in helping interpret microarray gene expression data. The key idea behind GSEA is to directly investigate the enrichment of pre-defined groups of functionally related genes (or gene sets) instead of individual genes. This group-based approach does not require pre-selection of genes with an arbitrary threshold. Instead, functionally related genes are evaluated together as gene sets, allowing additional biological information to be incorporated into the analysis process. The GSEA approach has proven to be remarkably successful in deriving new information from genome-wide expression studies, having been cited over 1500 times since its initial description (141,142).

The success of GSEA has inspired many extensions, improvements and variations (143,192-197). These methods can be classified into competitive or self-contained methods (see Section 1.5.5 for more details). The choice of which method to use in other "omics" disciplines is highly dependent on the type of data being generated. In particular, the competitive model (which assumes random sampling from a complete "omic" universe) is not suitable for today's metabolomics technologies. Although metabolomics is defined as the non-biased identification and quantification of all metabolites in a biological system, there are currently no analytical techniques or combinations of techniques that have achieved complete, unbiased coverage of the microbial, plant or mammalian metabolomes. Indeed, most metabolite measurements are biased either towards

more abundant compound species (NMR-based approaches) or compounds with better ionization efficiencies (MS-based approaches). Because the random sampling assumption does not hold for competitive models, self-contained approaches are actually more suitable for metabolomics applications.

In addition to the choice of an appropriate statistical procedure, the other essential component to using enrichment analysis is a knowledge base with pre-defined sets of related features. This knowledge base is essential to properly carry out group-based significance tests. The most widely used databases or knowledge bases of functionally related genes are based on KEGG pathways (146) and Gene Ontologies (198). The Molecular Signature Database (MSigDB) (142) is another example of a knowledge base that has been created primarily to support gene set enrichment analysis for human gene expression data. In addition to these resources, MSigDB is another database that includes gene sets co-expressed under the same experimental conditions. A complete list of compatible or useful databases for GSEA analyses is available at http://www.broadinstitute.org/gsea/msigdb/collections.jsp.

To our knowledge, no tools similar to GSEA have been developed to support this group-based approach for metabolomic data analysis. This is likely because both enrichment analysis and quantitative metabolomics are relatively new techniques. However, it is also likely due to the fact that in order to use this approach, one needs an extensive and biologically meaningful metabolite set library. Such a library is very laborious and time-consuming to create. Given the increasing availability of many public metabolite databases and a large body of

literature on metabolic pathways and metabolic diseases, it should be possible and relatively straightforward to create a comprehensive knowledge base in the form of metabolite sets for functional enrichment analysis. Therefore, by collecting and compiling knowledge about metabolites into computable forms (metabolite sets), and utilizing suitable algorithms originally developed for enrichment analysis, it is possible to develop a useful tool for metabolite set enrichment analysis. Following this approach, I have implemented a web-based application, named MSEA (Metabolite Set Enrichment Analysis), to support group-based enrichment analysis for human and/or mammalian metabolomic studies. The main features of MSEA include: 1) a collection of five metabolite set libraries containing ~6,300 biologically meaningful groups of metabolites; 2) three enrichment analysis methods – over-representation analysis (ORA), single sample profiling (SPP), and quantitative enrichment analysis (QEA), to support common data forms generated in metabolomic studies; 4) support for enrichment analysis with discrete and continuous phenotypes; 5) support for enrichment analysis using customized (non-mammalian) metabolite sets; 6) support for conversions between metabolite common names, synonyms, and identifiers of nine major metabolomic databases; and 7) comprehensive analysis report generation. Through MSEA and its accompanying databases, it is possible to take a list of altered metabolites from a biofluid or tissue sample and use it to suggest a biological pathway or disease condition that can be further investigated. The MSEA server and all of its accompanying databases are freely available at http://www.msea.ca. The remaining sections of this chapter provide information about the implementation

of MSEA, followed by a detailed description of its data analysis features. The chapter ends with discussion about its current limitations and future developments.

Implementation

Creation of Metabolite Set Libraries

A group of metabolites are considered to constitute a meaningful metabolite set if they are known to be: a) involved in the same biological processes (i.e., metabolic pathways, signaling pathways); b) associated with genetic traits (i.e. SNPs); c) changed significantly under the same pathological conditions (i.e., various metabolic diseases); and d) present in the same locations such as organs, tissues or cellular organelles. These data were collected through manual curation from books and journals as well as through text mining of public databases. The resulting metabolite sets were manually validated/edited and then further organized into three categories: pathway-associated, disease-associated, and location based. MSEA's pathway-associated metabolite library contains 84 entries based on the 84 human metabolic pathways found in the Small Molecular Pathway Database (SMPDB) (199). MSEA's SNP-associated metabolite sets were derived from the two recent genome-wide association studies between genetic variations and metabolite profiles in human (36,37). MSEA's disease-associated metabolite sets were mainly collected from the literature. Metabolites associated with different diseases were manually identified, merged and subsequently refined by reading the original publications listed in the Human Metabolome Database

(HMDB) (21), the Metabolic Information Centre (MIC), and SMPDB. Using these resources, a total of 851 physiologically informative metabolite sets were created. These disease-associated metabolite sets were further divided into three sub-categories based on the biofluids in which they were measured: 398 metabolite sets in blood, 335 in urine, and 118 in cerebral-spinal fluid (CSF). MSEA's location-based library contains 57 metabolite sets based on the "Cellular Location" and "Tissue Location" listed in the HMDB. A summary of these metabolite set libraries is shown in **Table 4.1**.

Creation of a Metabolite Dictionary and Concentration Database

In order for the MSEA server to accept a range of metabolite names, synonyms or identifiers as input, it was also necessary to develop a local metabolite dictionary that could be used to perform facile name conversion or "normalization". Information contained in the HMDB was used to extract common names, synonyms, as well as identifiers (ID) used in nine major metabolomic databases (HMDB, PubChem (155), ChEBI (156), KEGG (157), BiGG (158), METLIN (23), BioCyc (200), Reactome (149), and Wikipedia). Examples of MSEA's supported IDs are listed in **Table 4.2**. In order for MSEA to perform single sample profiling (SSP) analysis it was also critical to obtain reference concentrations for as many metabolites as possible. These concentration data were collected primarily from the HMDB with additional values being added through manual curation. MSEA's reference concentrations are organized based on the biofluids in which they were measured. Concentrations are presented in the form of *mean (minimum – maximum)*. For concentrations reported as mean and

standard deviation (SD), their 95% confidence intervals (mean ± 2 SD) were used to define the concentration ranges. One compound may have multiple concentration values as reported from different studies.

Implementation of Enrichment Analysis Programs

Over the past 5 years, many different algorithms have been developed for group-based enrichment analysis, including GSEA (142), GSEA-P (201), PAGE (202), globaltest (143), SAFE (193), SAM-GS (194) and GSA (195). Based on a thorough review of the literature, we decided to adapt the *globaltest* algorithm as the backend for MSEA. The globaltest is originally designed for testing association between gene sets and a clinical outcome. It uses a generalized linear model to compute a "Q-stat" for each gene set. For a group of *m* genes, the Q-stat is calculated as the average of the statistics $Q_1 \ldots Q_i \ldots Q_m$, calculated for each single gene, where Q_i is the average of the squared covariance between the gene expression pattern and the clinical outcome. There were three main reasons: 1) recent publications have indicated that *globaltest* exhibited similar or superior performance when tested against several other algorithms (203-205); 2) *globaltest* is very flexible and supports binary, multi-class, and continuous phenotype labels; and 3) *globaltest* is computationally efficient as the p-values can be calculated based on the asymptotic distribution, which is correct for large sample sizes, but also gives a good indication for small sample sizes.

Conventional over representation analysis was implemented based on a cumulative hypergeometric distribution. Since many metabolite sets are tested

simultaneously, we also implemented methods to adjust for the multiple testing problems that occur during enrichment analysis. In addition to the original p-values, MSEA also reports *Bonferroni* corrected p-values and false discovery rate (FDR) according to Benjamini and Hochberg (206).

Web Server Characteristics

MSEA's web interface was implemented using the JSF or Java Server Faces (http://java.sun.com/javaee/javaserverfaces) framework. The enrichment analysis algorithms were implemented in the R (version 2.10.0) programming language (http://www.r-project.org). The communication between R and Java was established through the *Rserve* TCP/IP server (http://www.rforge.net/Rserve). The web application is hosted on GlassFish (version 3) using a Linux operating system (Fedora Core 12). MSEA's host server is equipped with two Intel Quad Core 2 processors (3.0 GHz each) and 8 GB of physical memory. The web application is platform independent and has been tested successfully on Internet Explorer 8.0, Mozilla Firefox 3.0, and Safari 4.0.

Program Description

MSEA's workflow is illustrated in **Figure 4.1**. Briefly, metabolite set enrichment analysis can be described in four steps - data input, data processing, data analysis, and results download. In addition to its analysis utilities, users can directly download, browse or search MSEA's metabolite set libraries or perform compound name and ID conversions. The details of each step are discussed below.

Step 1. Data Input

MSEA accepts data in three different formats: a) a list of compound names entered in a single-column format; b) a list of compound concentrations entered as two-column data with the first column corresponding to the compound names/labels and the second corresponding to the concentration values; c) a concentration table containing metabolite concentration data from multiple samples. The table must contain comma-separated values (.csv) with rows for sample**s** and columns for metabolites. The second column of the table is reserved for phenotype labels (binary, multi-class, or continuous). Examples of these input formats are provided on the MSEA homepage.

Step 2. Data Processing

In this step, both the compound labels and the concentration values are examined for their suitability for downstream analysis. It is critical that the compound labels be recognized by the program in order to be compared with MSEA's collection of compound names in metabolite sets. Therefore a consistency check is done with the input names or IDs against the names and IDs stored in MSEA's metabolite dictionary. Any nomenclature inconsistency is flagged and displayed to users for manual inspection and correction. For single sample profiling (SSP, discussed later), the concentrations must be provided in a standard concentration unit (umol for blood and CSF, and umol/mmol_creatinine for urine) in order for the input data to be properly compared with MSEA's reference concentrations database. For QEA, the concentration values can be normalized and negative/missing values are allowed. Two widely-used chemometric methods - principal component analysis

(PCA) and partial least square (PLS) analysis - are available in MSEA to allow for data visualization, pattern identification, and outlier detection. Note that MSEA does not perform data normalization. Users are advised to visit MSEA's companion web site MetaboAnalyst (207) to access a variety of data processing and normalization options.

Step 3. Enrichment Analysis

Depending on the type of user input, MSEA offers three kinds of enrichment analysis: over representation analysis (ORA), single sample profiling (SSP), and quantitative enrichment analysis (QEA). These analysis modules are described in more detail below.

a) Over Representation Analysis (ORA) - ORA is used to evaluate whether a particular set of metabolites is represented more than expected by chance within a given compound list. ORA is performed when the user provides only a list of compound names. Such a list can be obtained using standard feature selection methods that statistically rank all the compounds and select those scoring above a certain threshold. ORA is also very useful for analyzing a group of compounds exhibiting similar concentration changes or patterns. Such a list can be obtained from standard clustering analysis. Many commonly-used feature selection and feature clustering methods are available from our companion web application MetaboAnalyst. The p value from ORA indicates the probability of seeing at least a particular number of metabolites from a certain metabolite set in a given compound list. The *Bonferroni* corrected p-value and false discovery rate (FDR)

are also presented to account for problems associated with multiple comparisons. Users can click the "View" link in the **Details** column of any of MSEA's metabolite sets to see all its constituent metabolites with matched ones highlighted in red, as well as pathway images (when available).

b) Single Sample Profiling (SSP) - For common human biofluids such as blood, urine or cerebral spinal fluid (CSF), normal concentration ranges are known for many metabolites. In clinical metabolomic studies it is often desirable to know whether certain metabolite concentrations in a given sample are significantly higher or lower than their normal ranges. MSEA's SSP module is designed to provide this kind of analysis. In particular, SSP is performed when the user provides a two-column list of both compounds and concentrations. When called, the SSP module will compare the measured concentration values of each compound to its recorded normal reference ranges of the corresponding biofluid (**Figure 4.2A**). By default, only compounds with concentrations above or below *all* the reported normal ranges will be selected for further investigation. Users can manually select or deselect compounds to override this default selection by inspecting the concentration comparison plots generated by this module (**Figure 4.2B**). The selected compound list will be subjected to over-representation analysis as described in the previous section.

c) Quantitative Enrichment Analysis (QEA) **-** QEA is performed when the user uploads a concentration table containing metabolite concentration data from multiple samples. QEA is based on the *globaltest* algorithm to perform enrichment analysis directly from the raw concentration data and does not start

from a list of significant compounds. It can identify significant metabolite sets with compounds that have limited changes in concentrations. Enriched metabolite sets will be identified when only a few compounds are highly differentially changed or many compounds are only slightly (but consistently) changed. In addition to the Q-stat values, the QEA module also provide p-values, *Bonferroni* corrected p-values and estimates of false discovery rates (FDR). **Figure 4.2C** shows a screenshot of the result table from a typical quantitative enrichment analysis. Users can click the image icon of any matched metabolite set to view a detailed graphical summary of the contributions of individual metabolites, as shown in **Figure 4.2D**.

Step 4. Data Download

When users finish an enrichment analysis, a comprehensive report is generated with detailed descriptions of each step performed, embedded with graphical and tabular results. The processed data, images, R scripts, as well as the R command history are also available for download. Users familiar with R can easily reproduce the results on their local machine after installing the R packages and the corresponding metabolite set libraries (available on the **Resources Download** page).

Other Features

Compound Name and ID Mapping Tool

The MSEA web server also offers a number of other features to facilitate metabolomic data analysis, including 1) a compound name and identifier mapping

tool; 2) a browser for metabolite sets; and 3) a facility for custom metabolite set uploads. Given the fact that no consensus exists in labeling compounds in current metabolomic studies, we implemented a utility in MSEA to convert between common compound names, synonyms and the identifier codes used in nine major metabolite databases (see **Table 4.2** for details). This converter can also deal with spelling errors using an approximate text matching algorithm. In addition to this name/ID converter, MSEA also provides a browser to view MSEA's collection of metabolite set libraries. These libraries can provide a valuable source of information to investigate the biological implications of any metabolite sets identified after enrichment analysis. The browser implemented in the MSEA web server allows users to easily scan and search its metabolite set libraries. Each entry contains the metabolite set name, its constituent compounds, and links to original references. Given the incompleteness of MSEA's metabolite-set libraries, researchers may want to perform enrichment analysis using customized or self-defined metabolite sets other than the ones provided by the server. MSEA supports this option by allowing users to upload their own metabolite set library. The library file should be in a simple *.csv* file with the first column for metabolite set names and the second for compound members.

Limitations

Unlike genomics or transcriptomics, metabolomics has not yet achieved total metabolite coverage. Whereas Next-Gen DNA sequencers and modern microarrays routinely cover entire genomes, most metabolomic technologies only offer 5-10% coverage of a sample's metabolome. This makes many metabolomic

studies intrinsically biased. Since most of the metabolite sets in MSEA's libraries are also derived from experimental studies, they tend to suffer from the same sampling bias. Fortunately these biases tend to cancel each other out, as essentially the same metabolite population (the fraction of the metabolome that are "detectable" by current analytical technologies) is probed to generate both metabolite sets and user data. Nevertheless, users should always take note of their experimental conditions or technological limitations when interpreting the results from enrichment analysis.

Another key limitation to MSEA is its bias to human and/or mammalian metabolomics. The mouse/rat metabolite set libraries are currently under construction. We also plan to add other metabolite sets from plants and microbes. However, until these databases and data sets can be completed (likely in two years time) we would encourage researchers who are engaged in metabolomic studies of non-mammalian species to create their own customized metabolite sets for enrichment analysis and to contribute these sets to the MSEA server for public use.

Conclusions

Over the past few years a number of software tools have been developed to address the bioinformatics needs of metabolomics. However, most of these programs were designed for spectral data processing and compound identification. More recently, several freely available software tools for the statistical analysis of metabolomic data have started to appear, such as MetaboAnalyst and MeltDB

(208). As yet, no publicly available tools have been made available to assist in the functional interpretation of metabolomic data. To address this issue we have developed a web server, named MSEA (Metabolite Set Enrichment Analysis), designed to help researchers identify and interpret patterns of metabolite concentration changes in a biologically meaningful context. MSEA performs three kinds of enrichment analysis - over representation analysis (ORA), single sample profiling (SSP) and quantitative enrichment analysis (QEA). When only a list of compounds is available, ORA is performed. When both compound names and concentrations are available, the SSP module is called. When concentration data is available from multiple samples, MSEA performs QEA. The enrichment analyses performed by MSEA are based on five carefully compiled metabolite libraries consisting of ~6,300 entries. In addition to its enrichment analysis capabilities, MSEA allows custom metabolite sets to be uploaded for more specialized (non-mammalian) studies. MSEA also supports conversion between metabolite common names, synonyms, and major database identifiers. We believe that, over time, the MSEA approach will become more powerful as analytical technologies for metabolomics continue to improve their metabolite coverage and as the metabolomics community develops improved standards and ontologies (209). In the long run, we would like to turn the MSEA server into a resource for metabolomic annotation, visualization, and integrated discovery much as the DAVID server (210) has become just such a resource for microarray data analysis.

Tables

Table 4.1 Overview of MSEA's metabolite set libraries.

A total of 6292 biologically related metabolite sets were collected through text-mining and manual curation. These metabolite sets are divided into several categories based on their biological context.

Category	**Total #**
Biochemical Pathway	84
Disease - associated	851
*Blood	344
*Urine	290
*CSF	108
SNP-associated	4501
Predicted biomarker	912
Location - based	57

Table 4.2 Overview of compound labels currently supported by MSEA

Label Type	Examples
Common Name	Adenosine, Acetic acid, Adenine, Creatine
HMDB	HMDB00050, HMDB00042, HMDB00034
PubChem	60961, 176, 190, 586
ChEBI	16335, 15366, 16708, 16919
KEGG	C00212, C00033, C00147, C00300
BiGG	34273, 33590, 34039, 34543
METLINE	86, 3206, 85, 7
BioCyc	ADENOSINE, ACET, ADENINE, CREATINE
Reactome	114933, 114747, 114936, 114818
Wikipedia	Adenosine, Acetic acid, Adenine, Creatine

Figures

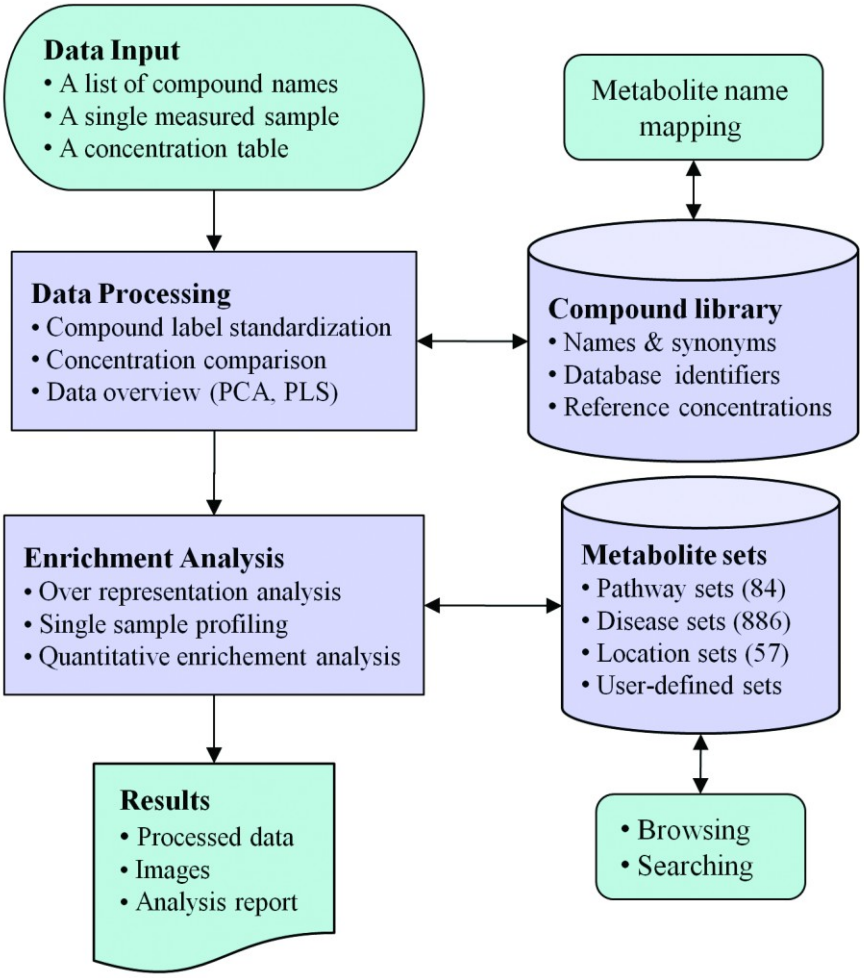

Figure 4.1 MSEA workflow.

Metabolite set enrichment analysis consists of four steps: data input, data processing, data analysis, and data download. Different analysis procedures are performed for different input types. MSEA allows users to directly browse and search its metabolite set libraries as well as to perform metabolite name mapping between different names and database identifiers.

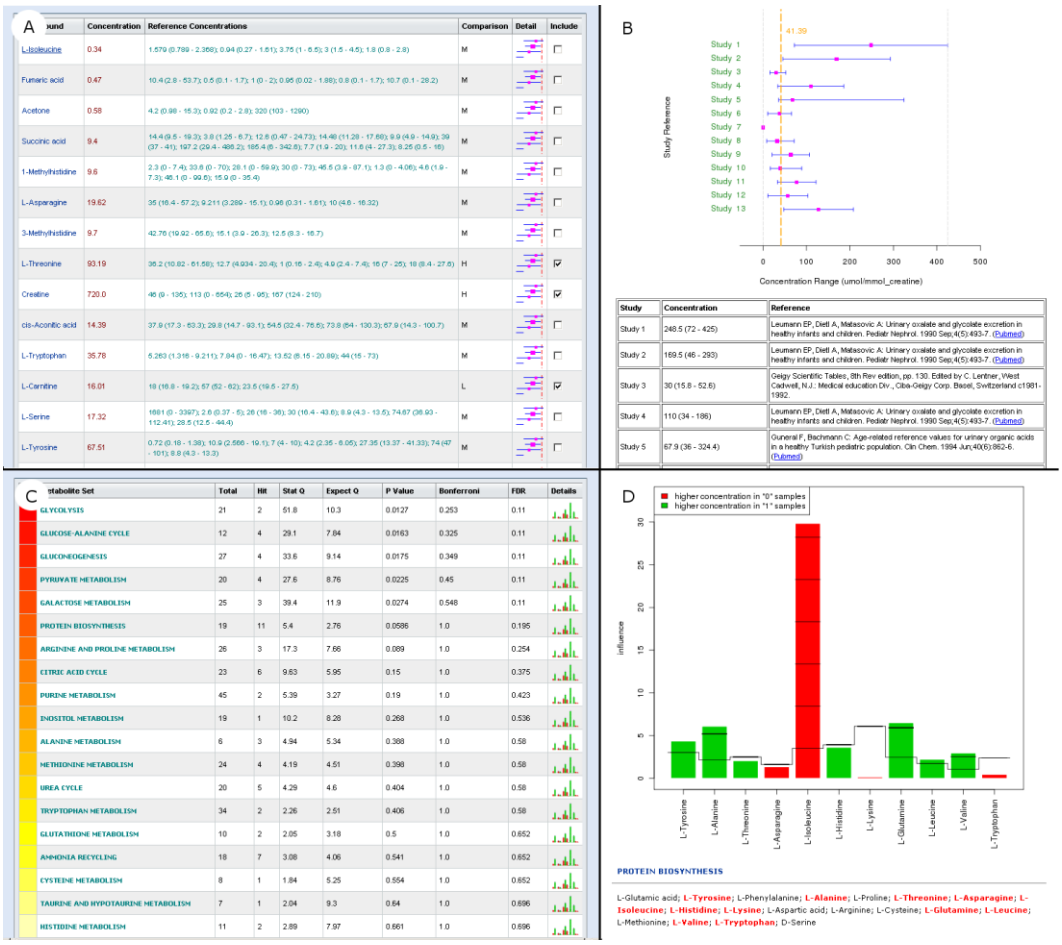

Figure 4.2 Enrichment analysis and visualization

Results from MSEA's enrichment analysis are presented both in tables as well as through graphical summaries. Figure A shows the comparison between the measured concentrations and reference concentrations using the single sample profiling (SSP) module. The top part of Figure B shows a graphical summary of the concentration comparison for a single compound when users click an image icon in Figure A. The bottom part of Figure B shows all the corresponding publications that reported these concentrations. Figure C shows the results generated by the quantitative enrichment analysis (QEA) module. The top part of Figure D is a metabolite-set plot indicating the influence of an individual compound on each of the selected metabolite sets. The bottom part of Figure D shows all its constituent metabolites with matched ones highlighted in red.

Chapter 5

Metabolic Pathway Analysis and Visualization[4]

[4] A version of this chapter has been published previously:

Xia, J. and Wishart, D.S. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, 26, 2342-2344

Introduction

Over the past decade, pathway analysis has emerged as an invaluable aid to understanding the data generated from various "omics" technologies. As a result, a number of robust software tools have been developed to support pathway analysis for genomics and proteomics studies (150-152,211). However, for metabolomics, pathway analysis is neither nearly as sophisticated nor as well-supported. Indeed, most pathway analysis in metabolomics is done through simple visual inspection of metabolic pathway databases such as KEGG (146), Reactome (149) or BioCyc (148). These databases provide, basically, little more than simple visual displays with modest support to highlight matched metabolites in the context of pathways. Given these limitations, it is clear that more capable pathway analysis tools are needed for metabolomics. For instance, the development of pathway databases/resources that can directly accept metabolite concentration tables (for more quantitative pathway analysis) would be a useful starting point.

Another limitation of today's pathway analysis tools or methods is they tend to focus on enrichment analysis and do not consider the inherent topological information in pathways. Over the past two decades, many algorithms have been developed in the field of graph theory to allow more quantitative interrogation of pathway and network structures. This kind of interrogation can provide useful insights into pathway function, stability and dynamic responses (212). Given that it can be very useful to quantify the importance of a molecule based on its position within a given pathway, it stands to reason that graph theory and network

topology analyses should be able to help support this kind of assessment. For instance, the importance of a compound within a given metabolic network can be estimated by its centrality, a notion that is routinely used in the study of social and computer communication networks to estimate the potential monitoring and control capabilities of a given node. There are two commonly used centrality measures – *degree centrality* and *betweenness centrality* (**Figure 5.1**). The degree centrality measures the number of connections the node of interest has to other nodes. Nodes with higher degree centrality act as hubs in a network. The betweenness centrality measures the number of shortest paths going through the node of interest. Nodes with high betweenness centrality are bottlenecks in a network.

One practical challenge in pathway analysis is pathway visualization. Visualization is very important for presentation and proper interpretation of pathway analysis results. Traditionally, web-based visualization of pathways has been implemented through static images (i.e. KEGG, SMPDB). Dynamic network presentation must normally be done through stand-alone programs (i.e. Cytoscape (213), GenMAPP (211)) due to the high computational cost of creating a network layout and the bandwidth constraint for transferring large images. However, because human eyes are able to track only a small portion of a complex network at any given time, it is not necessary to update the whole image every time. Only updating a viewport is much more efficient for image transfer, as demonstrated by Google Maps (http://maps.google.ca).

Therefore, by combining enrichment analysis methods with suitable algorithms originally developed from graph theory and presenting the result dynamically via web-interface using a Google-map style visualization system, it will be possible to develop an effective tool for metabolomic pathway analysis. Following this approach, I have developed a web server called MetPA (Metabolomic Pathway Analysis). MetPA is a user-friendly, web-based tool dedicated to the analysis and visualization of metabolomic data within the biological context of metabolic pathways. MetPA combines several advanced pathway enrichment analysis procedures along with the analysis of pathway topological characteristics to help identify the most relevant metabolic pathways involved in a given metabolomic study. The results are presented in a Google-map style network visualization system that supports intuitive and interactive data exploration through point-and-click, dragging, and lossless zooming. Additional features include a comprehensive compound library for metabolite name standardization, as well as the implementation of various univariate statistical procedures that can be accessed when users click on any metabolite node on a pathway map. MetPA currently enables analysis and visualization of 1173 metabolic pathways, covering 15 common model organisms. MetPA is freely available at http://metpa.metabolomics.ca. The remaining sections of this chapter provide information about MetPA's implementation followed by a description of its features as illustrated by an example data analysis.

Implementations

Analysis Algorithm

Pathway analyses in MetPA can be conducted through three routes. Pathway enrichment analysis supports both over-representation analysis as well as GSEA-based approaches. The available algorithms include Fisher's exact test, the hypergeometric test, globaltest (143), and GlobalAncova (214). MetPA's pathway topological analysis is based on the centrality measures of a metabolite in a given metabolic network. Centrality is a local quantitative measure of the position of a node relative to the other nodes, and is often used to estimate a node's relative importance or role in network organization (215). Since metabolic networks are directed graphs, MetPA uses relative betweenness centrality and out degree centrality measures to calculate compound importance. The pathway impact is calculated as the sum of the importance measures of the pathway metabolites that matched the query data normalized by the sum of the importance measures of all metabolites in the pathway. Finally, MetPA provides a number of univariate analyses performed at the compound level to provide a more detailed view of the distribution of individual metabolite concentrations with regard to phenotypes. They include the t-test, one-way analysis of variance (ANOVA), and linear regression.

Pathway Library Construction and Visualization

The pathway data used in MetPA were downloaded as KGML files from the KEGG database (146). Chemical compounds and pathway topology information were parsed into graph models using the *KEGGgraph* package (216). The current library contains 1173 metabolic pathways from 15 model organisms including humans, mouse, drosophila, Arabidopsis, E. coli, *etc*.

Metabolic pathways are presented as a network of chemical compounds with metabolites as nodes and reactions as edges. The graph generation and manipulation were implemented using Graphviz (http://www.graphviz.org) and ImageMagick (http://www.imagemagick.org). This visualization system supports lossless zooming, dragging, and linking operations based on Ajax (Asynchronous JavaScript with XML) technology (217). All relevant information can be obtained by clicking on the corresponding graphical elements.

Web Interface

MetPA's web interface was implemented using the Java Server Faces (JSF) (http://java.sun.com/javaee/javaserverfaces) framework. The pathway analysis algorithms were implemented in the R (version 2.10.0) programming language (http://www.r-project.org). The communication between R and Java was established through the *Rserve* TCP/IP server (http://www.rforge.net/Rserve). The web application is hosted on GlassFish (version 3) using a Linux operating system (Fedora Core 12). The server is equipped with two Intel Core 2 Quad processors (3.0 GHz each) and 8 GB of physical memory. The web application is platform

independent and has been successfully tested on Mozilla Firefox 3.0+, Safari 4.0+, Google-Chrome 5.0+, Opera 10.0+, and Internet Explorer 8.0.

Example Analysis

MetPA accepts either a list of significant compound names, or a compound concentration table with binary, multi-group, or continuous phenotype labels. In the latter case, it is advisable to first normalize the concentration data, i.e. using MetaboAnalyst (207). As an example, we present the analysis on urinary metabolite concentration data (log-normalized) from cancer patients experiencing either muscle gain (Y) or muscle loss (N) monitored over a three-month period. The purpose is to investigate if certain metabolic pathways are significantly different between the two groups of patients. The first step is to convert the compound names of the uploaded data to the compound names used in the pathway library. MetPA uses compound names, synonyms and database IDs data from the HMDB (25) to perform compound name mapping. The next step is to specify the parameters for the pathway analysis – i.e. the pathway library, the algorithm for pathway enrichment analysis, as well as the algorithm for topological analysis. In this case, we select the "Homo sapiens" library and use the default "Global Test" and "Relative Betweenness Centrality" for pathway enrichment analysis and pathway topological analysis, respectively. The result is presented in two parts - the graphical output (shown in **Figure 5.2A)** and a table containing all the analysis results. Users can intuitively explore the results by pointing and clicking on various hyperlinked nodes. For example, let's look at the

"Glycine, serine and threonine metabolism" pathway, which is the top pathway from the pathway topological analysis and is also significant in the pathway enrichment analysis (4.65E-5 after adjustment of multiple testing). Clicking the circle on the "metabolome view" (**Figure 5.2A**.) on the left panel launches the corresponding "pathway view" (**Figure 5.2B**) on the right. It is interesting to see that many of these significantly changed amino acids are in key positions for this pathway. Further checking (by clicking on each metabolite node) indicates that all the nine matched amino acids show higher concentration values in the muscle loss group, with Creatine being the most significant (**Figure 5.2C**). It is interesting to see that the most significant pathway identified from the enrichment analysis is "Galactose metabolism" (highlighted as the dark red circle on the top left corner of the "metabolome view"). Further checking indicates only three downstream peripheral compounds are involved, with "Myoinositol" being most significant. It is less likely that this pathway is strongly associated with muscle change.

Conclusions

The growing interest in metabolomics and systems biology has increased the need for computational and visual tools for pathway analysis. MetPA is a full-featured, easy-to-use pathway analysis and visualization environment that combines advanced statistical enrichment analysis with pathway topological characteristics to help researchers identify the most relevant pathways involved in the conditions under study.

Figures

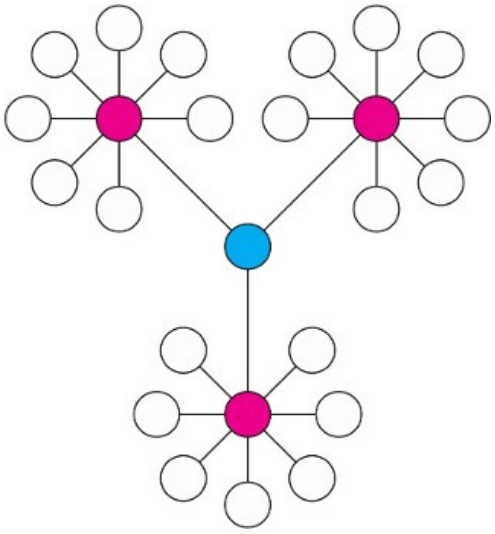

Figure 5.1 Illustration of centrality measures.

There are two commonly used centrality measures - degree centrality measures the number of connections the node of interest has to other nodes; betweenness centrality measures the number of shortest paths going through the node. The red nodes have the highest degree centrality and the blue node with highest betweenness centrality. The figure was adapted from *Junker et al* (218) .

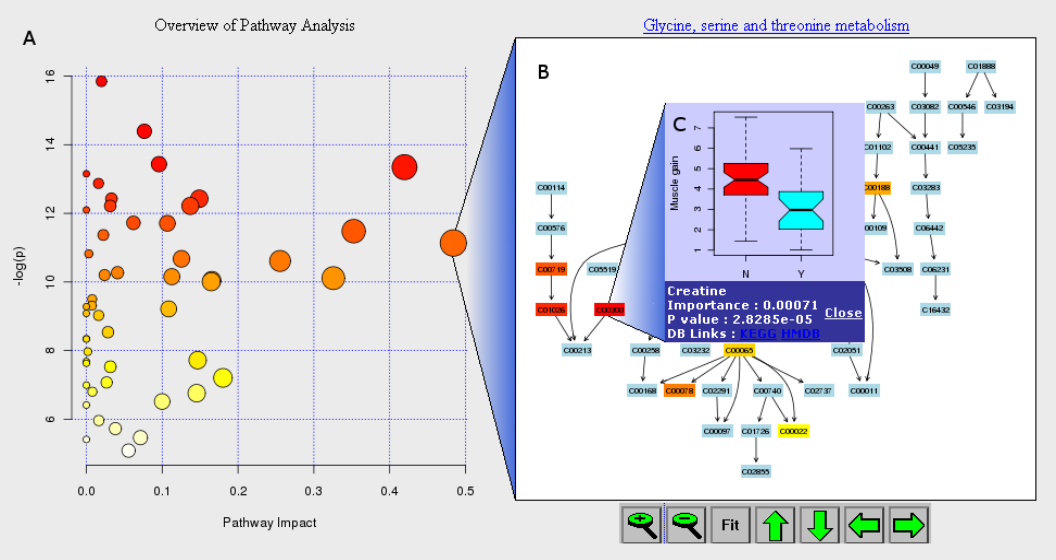

Figure 5.2 Screenshot illustration of MetPA's data visualization features

(A) metabolome view, (B) pathway view, and (C) compound view. Navigation to each view is by point-and-click on the corresponding node.

Chapter 6

Validation & Example Applications

Background

The primary objective of my thesis research has been to develop efficient and user-friendly tools to address several bioinformatics challenges that routinely arise in metabolomics studies. Many of the tools have been inspired by established concepts and established algorithms that have been developed to address the similar issues in other omics fields (i.e. transcriptomics). As a result of this approach, stringent evaluation and validation of the statistical methods have not been of particular concern. Instead, my metabolomic tools and their results have been illustrated using example test datasets. However, it is important to show that the concepts embodied by these programs and servers work as expected. In this chapter, I will perform somewhat more detailed validation of these servers and present some applications using both simulated and real-world metabolomic data.

Because MetaboMiner has already been subject to a great deal of thorough testing, validation, and comparison with other tools, I will not perform any further validation with this tool. Instead, this chapter will focus on the validation on the other three web applications - MetaboAnalyst, MSEA, and MetPA. The goal of this validation study is to test the utility of these web applications with respect to whether it can detect compounds that are significantly different, whether they can help interpret metabolomic data through enriched functional groups or affected pathways. In this chapter, I will show that these web-based bioinformatics tools can perform these intended tasks *correctly*, *efficiently*, and *in a user-friendly manner*.

Validation Design

In order to show that these tools can perform the intended tasks correctly, a gold standard test data set (i.e. a data set with a known and correct answer) is needed. This can be obtained from published data with a clear biologically interpretable result which has been subsequently confirmed in a follow-up study. Alternatively, it can also be a data set where the results have been widely-recognized as correct by the other researchers in the same field. For example, the famous Golub dataset (219) has been widely used by software developers to validate various tools for microarray gene expression analysis. In addition, there are many other benchmark microarray datasets available from the Gene Expression Omnibus (GEO) at NCBI (220) or the ArrayExpress at EBI (221). Unfortunately, no such widely-recognized dataset is available for metabolomics at this time. Indeed, currently there is no publicly available central repository for metabolomics data similar to the GEO or ArrayExpress. A practical solution to this limitation is to use simulated data. If the simulated data closely mimic the characteristics of the real-world data, the results can also be very useful. To show that these tools can perform their intended tasks quickly and correctly, a real-world data set from a study already published will be used. In this case, the goal is to show that similar results can be obtained using these tools in a very short time. To measure user-friendliness is very subjective. Instead, I will use the user statistics obtained from Google Analytics to show the popularity of these tools, which can serve as a proxy indicator of their popularity and user-friendliness.

Materials & Methods

Two test data sets were used to validate these web applications - a simulated data set and a real-world data set. The real-world data was obtained from a metabolomics study on dairy cows fed with different proportions of barley grain (222). Specifically it contains the concentrations of 47 metabolites measured on 39 rumen samples divided into four groups - 0, 15, 30, or 45 - indicating the percentage of barley grain in the diet. The simulated data is based on a real urinary metabolomics data obtained from a metabolomics study of cancer cachexia published recently (223). This data set contains 77 urine samples with each sample containing the concentrations for 63 compounds. There are two groups – a cachexic (muscle loss) group and a control group. The simulated data was generated based on two assumptions: i) the concentration values of each metabolite followed a normal distribution; and ii) most changes among metabolite concentrations were within normal variation, with a few changes being biologically significant. The procedures for generating the data are briefly described below:

1) Generate a concentration table with the same number of samples and compounds as the cachexia dataset. In particular, each compound must have the same mean and variance as the corresponding metabolite in the real data;

   For each compound,

a. Estimate its mean and standard deviation (SD) using robust measures - median and median absolute deviation (MAD);

b. Regenerate the concentration values by sampling a normal distribution with the calculated mean and SD; for negative values or very small positive values (< 0.05), replace them with 0.05;

c. Replace the original concentrations with the new values

2) Introduce changes to a few selected compounds in the cachexic group. Six metabolites involved in the Citric acid cycle were chosen: *2-Oxoglutarate, Citrate, Fumarate, Pyruvate, Succinate, and cis-Aconitate.* The concentration of these metabolites are re-sampled from a normal distribution with a two-fold change in the parameters, namely either (2*mean with 2*SD) or (1/2*mean with 1/2*SD). The up- or down-regulation of a particular compound was determined randomly.

Please note, using the above procedure, some characteristics of this metabolomic data set (concentration range and variations) will be preserved. However, the covariance structures between different metabolites will be lost. This is unavoidable unless we can model the distribution for all these metabolites.

The two data sets were subject to analysis using MetaboAnalyst, MSEA, and MetPA. As there are so many parameters and options available with these programs, here I will only show the results from a few common procedures - univariate tests, multivariate tests, enrichment analysis, and pathway analysis. A

more detailed step-by-step protocol on how to use these web-based tools can be found in **Appendix I**.

Results

Identification of Significantly Changed Compounds and Pathways

After the data processing and normalization steps, statistically significant compounds in these data sets can be identified using t-tests. The top 10 most significant compounds are given in **Table 6.1**. As expected, the five computationally "spiked-in" compounds occupy the top five of the list. However, *Succinnate* is ranked eighth (after *Quinolinate* and *Formate*), and is only marginally significant with a p value of 0.054. If we use the conventional p value cut-off 0.05, *Quinolinate* and *Formate* would both be selected as significant. As this is simulated data and we know that their concentrations in both cachexic and control groups are generated from the same normal distributions. Therefore, the observed differences in *Quinolinate* and *Formate* between the two groups are purely due to chance (false positives). The risk of finding false positives increases when multiple tests are performed. In this case, we should use FDR (false discovery rate correction). For example, with a common threshold of 20% (one in five is false positive), the top five will be selected.

The data can be further analyzed using PLS-DA. A three-dimensional score plot is shown in **Figure 6.1**. A good separation can be seen from the figure. The significant features identified by PLS-DA are almost the same as those

identified by the t-tests. The permutation test shows the difference between the two groups (cachexic and control) is marginally significant (permutation p value = 0.06). Note the variance explained of each component is almost identical (~4%). This is caused by the way the simulated data was generated, with each metabolite concentration being generated independently.

Metabolite set enrichment analysis (MSEA) using the pathway database from SMPDB (147) shows the top five candidates for the affected pathways (**Table 6.2**). As expected, the Citric acid cycle is identified as the most significantly altered. This result is further confirmed using pathway analysis. As shown in **Figure 6.3**, citric acid metabolism is located in the top right corner, indicating it is considered significant by both enrichment analysis and topology analysis. A detailed view on these matched metabolites in the context of the pathway structure is shown in **Figure 6.4**.

Reproduction of Results from Published Data

A metabolomic data analysis using MetaboAnalyst was performed following precisely the same procedures described in the original study by Ametaj, *et al*. (222), in which the authors used PCA and ANOVA to identify significant biomarkers as well as pattern of changes that are associated with the dietary change. The results can be easily reproduced with these tools.

After data processing and normalization, many compounds are significantly changed based on ANOVA tests with a p-value threshold 0.05 (**Figure 6.5)**. The details of these selected metabolites are shown in **Table 6.3**. In

addition, results from the post-hoc analyses are also performed to indicate which two groups are significantly different (using the same p-value cutoff). PCA 2D score plot (**Figure 6.6**) reveals four partially overlapping but distinct groups. The corresponding loading plot is shown in **Figure 6.7**. These results agree very well with the published data. Note that other analyses such as hierarchical clustering, PLS-DA can also be easily performed (results not shown). The whole process can be finished in 15~20 minutes.

User Profile Statistics

The design and implementation of these web-based tools for metabolomics was inspired by the existence of several successful counterparts developed for other omics fields. This was done to minimize the learning curve for end-users. It has also made these tools quite popular in the metabolomics community. According to Google Analytics (http://www.google.com/analytics), as of July 1, 2011, MetaboAnalyst has attracted over 23000 visits by more than 6000 distinct visitors from around 1200 cities worldwide since its publication (**Figure 6.8A**), with an average of ~2000 monthly visits. MSEA has attracted over 2000 visits since it was first published in May 2010, which translates to ~150/month. MetPA has attracted ~1700 visits since it was published in July, 2010 or about 120 monthly visits. Please note that MSEA and MetPA have been incorporated into MetaboAnalyst as of September, 2010, reducing the web traffic for these two websites (and concomitantly increasing the traffic for MetaboAnalyst).

It is important to note that the number of visits is not an accurate indicator of the people who actually used the web application. For example, the number of the users who actually uploaded their data may better reflect the utility of the websites. Unfortunately, the statistic is not available as MetaboAnalyst server uses a cron job to automatically remove user uploaded files after 72 hours. In one occasion, we forgot to turn on the cron job after a server update. The hard disk was full within one month's time. This could partially reflect the heavy use of MetaboAnalyst. In addition, as shown in **Figure 6.8B,** most visits are actually from returning visitors. This could also attest to the utility and user-friendliness of the web application.

Conclusions

In this chapter, I have performed a validation study on the three web applications using a simulated dataset and a real-world data set. The results from the simulated data indicate that these tools are able to correctly identify these "spike-in" metabolites either individually or as a group in the form of either metabolite sets or metabolic pathways. The results from the real-world data show that using these tools, users can quickly perform various data processing and analysis methods that are commonly seen in many published papers. Finally, as a qualitative measure of their user-friendliness, I have also shown that these tools have gained a great degree of popularity among researchers in the metabolomics community.

Tables

Table 6.1 Importance features identified from t-tests (simulated data)

	Compounds	P values	FDR
1	Fumarate	2.00E-05	0.00118
2	cis-Aconitate	4.00E-05	0.00118
3	2-Oxoglutarate	1.40E-04	0.00295
4	Citrate	3.70E-04	0.0058
5	Pyruvate	0.01431	0.18029
6	Quinolinate	0.02386	0.25049
7	Formate	0.03806	0.34253
8	Succinate	0.05453	0.42939
9	Glycine	0.09117	0.58266
10	Threonine	0.09915	0.58266

Table 6.2 Significant pathways identified from enrichment analysis (simulated data)

	Total	Hits	Stats.	P values	FDR
CITRIC ACID CYCLE	23	6	14.588	5.00E-12	2.30E-10
UREA CYCLE	20	4	11.8	1.25E-07	5.64E-06
MITOCHONDRIAL ELECTRON TRANSPORT CHAIN	15	2	13.03	1.83E-05	8.06E-04
ALANINE METABOLISM	6	2	12.71	3.18E-05	0.001369
MALATE-ASPARTATE SHUTTLE	8	1	17.683	1.40E-04	0.005891
PHENYLALANINE AND TYROSINE METABOLISM	13	2	10.688	1.42E-04	0.005891
GLUCOSE-ALANINE CYCLE	12	3	8.4949	1.60E-04	0.006399

Table 6.3 List of significant compounds identified by ANOVA (real data).

The last column shows the results from the post-hoc analysis using Fisher's LSD. Significantly different groups are presented as a pair linked by a hyphen.

	Compounds	p.value	-log10(p)	FDR	Fisher's LSD
1	Endotoxin	0.00	8.42	0.00	30 - 0; 45 - 0; 30 - 15; 45 - 15
2	Glucose	0.00	7.91	0.00	45 - 0; 45 - 15; 45 - 30
3	3-PP	0.00	7.80	0.00	0 - 15; 0 - 30; 0 - 45; 15 - 30; 15 - 45
4	Alanine	0.00	5.66	0.00	30 - 0; 45 - 0; 30 - 15; 45 - 15
5	Isobutyrate	0.00	4.76	0.00	0 - 30; 0 - 45; 15 - 30; 15 - 45
6	Methylamine	0.00	4.66	0.00	45 - 0; 45 - 15; 45 - 30
7	3-HP	0.00	4.28	0.00	15 - 0; 15 - 30; 15 - 45
8	Lactate	0.00	3.96	0.00	30 - 0; 30 - 15; 30 - 45
9	Uracil	0.00	3.63	0.00	15 - 0; 30 - 0; 45 - 0
10	Aspartate	0.00	3.39	0.00	0 - 45; 15 - 45; 30 - 45
11	Isoleucine	0.00	3.30	0.00	30 - 0; 0 - 45; 30 - 15; 30 - 45
12	Butyrate	0.00	3.00	0.00	0 - 15; 0 - 30; 15 - 30; 45 - 30
13	Acetate	0.00	2.53	0.01	0 - 30; 0 - 45; 15 - 30
14	NDMA	0.01	2.04	0.03	30 - 15; 45 - 15
15	Lysine	0.02	1.76	0.05	45 - 15; 45 - 30
16	Fumarate	0.02	1.63	0.07	15 - 30; 45 - 30
17	Ferulate	0.03	1.60	0.07	15 - 30; 45 - 30
18	Cadaverine	0.03	1.57	0.07	30 - 15; 45 - 15
19	Isovalerate	0.03	1.56	0.07	0 - 15; 0 - 30; 0 - 45
20	Benzoate	0.03	1.52	0.07	30 - 45
21	Phenylacetate	0.03	1.50	0.07	15 - 0; 15 - 30
22	Leucine	0.04	1.42	0.08	30 - 0; 30 - 15
23	Valine	0.04	1.37	0.09	30 - 0; 45 - 0; 30 - 15

Figures

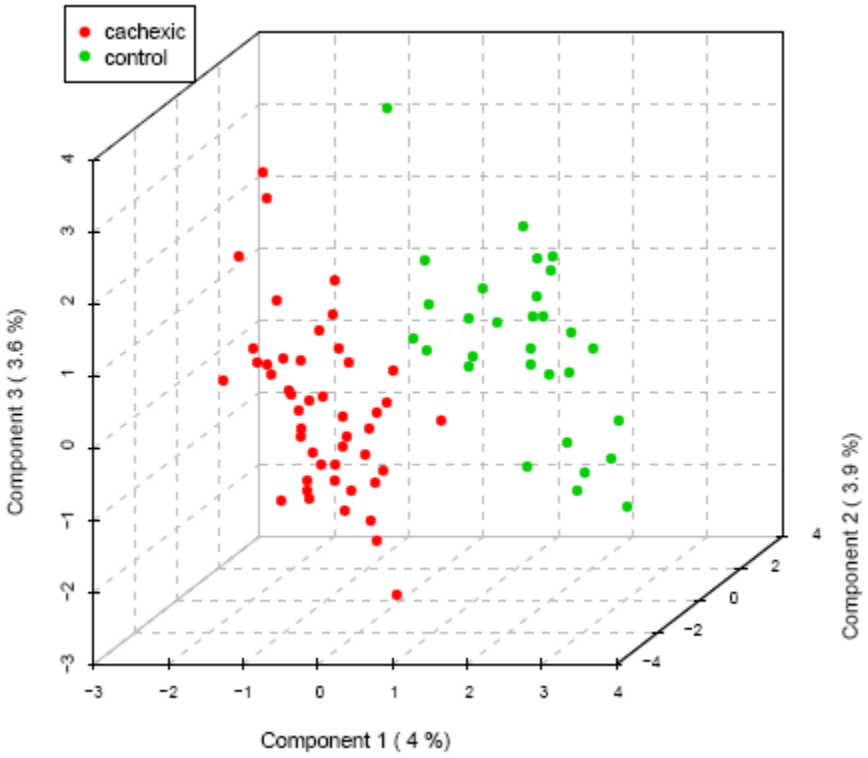

Figure 6.1 PLS-DA score plot with top three components (simulated data).

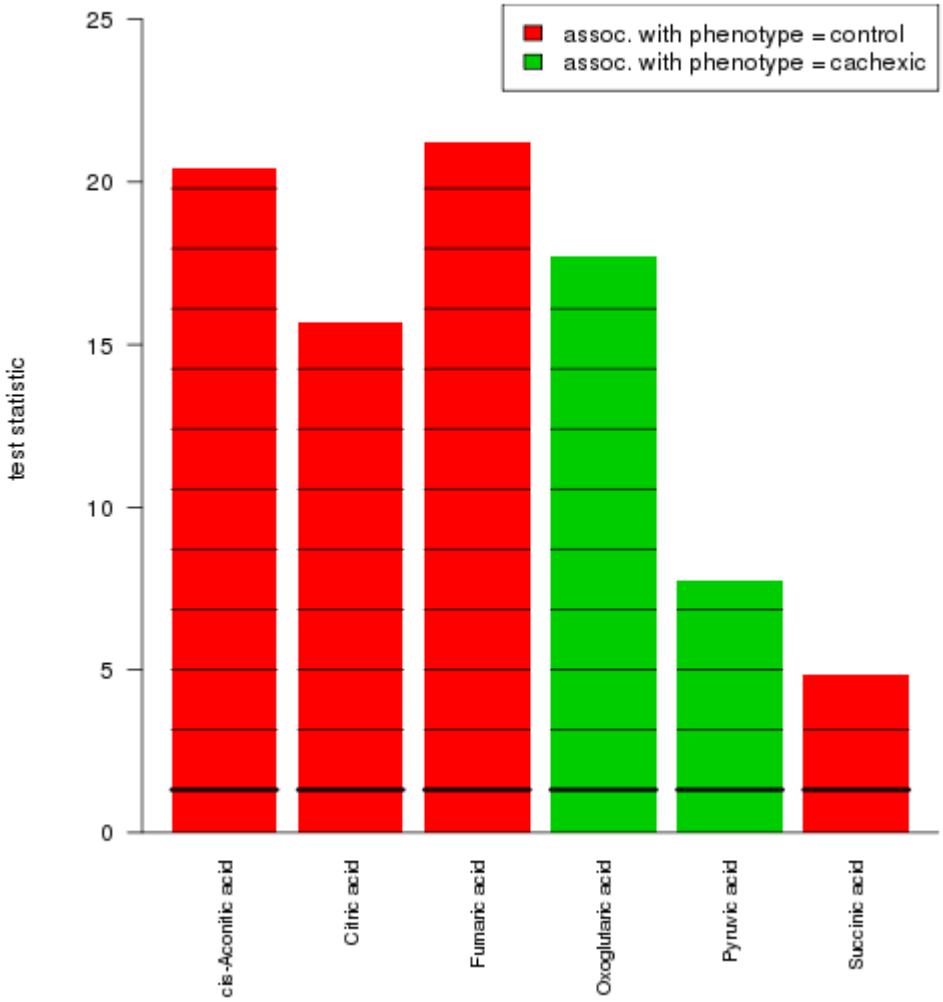

Figure 6.2 Matched metabolites in Citric acid metabolism (simulated data).

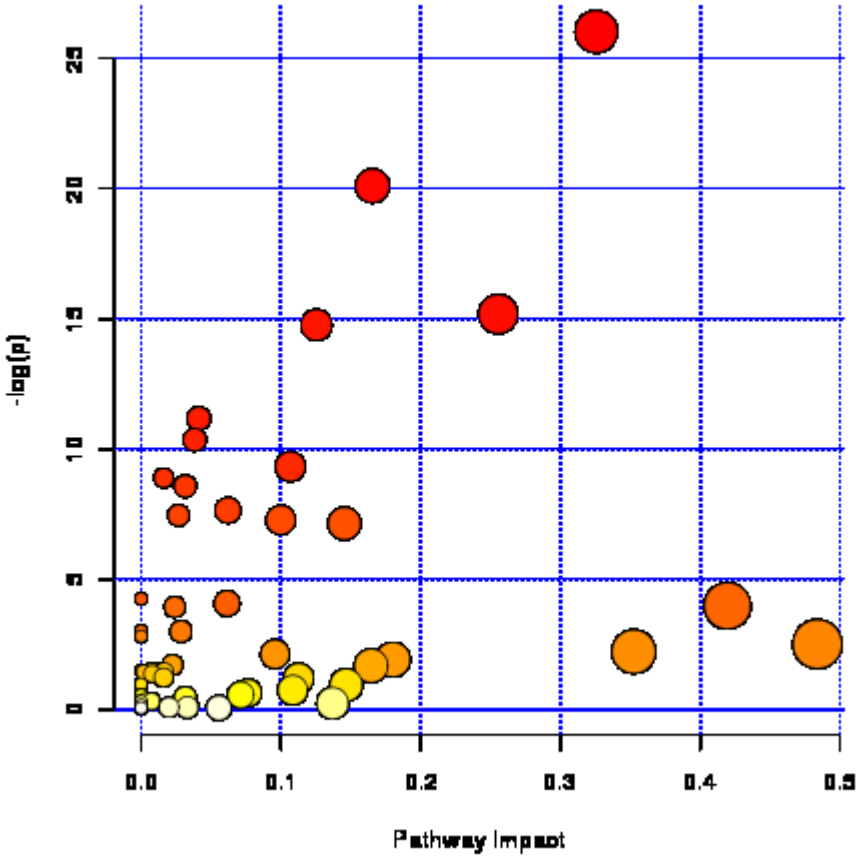

Figure 6.3 Overview of the affected pathways (simulated data).

The x-axis in this graph indicates the impact of the pathways based on node importance measures; the y-axis indicates the significance of each pathway based on its p-value from the enrichment analysis. Each pathway is represented by a circle rendered in a color according to its p-value with a radius proportional to its impact value. The most affected pathways should be lying along the diagonal (lower left to top right). The top-right circle located around (0.33, 26) represents the TCA pathway.

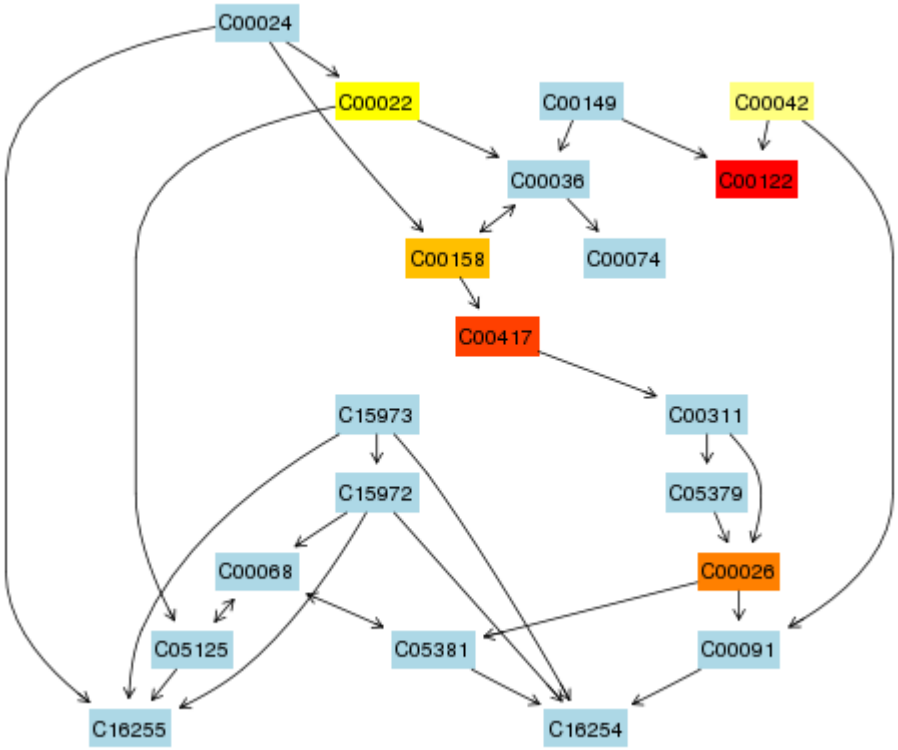

Figure 6.4 Pathway view of Citric acid cycle (simulated data)

Unmatched metabolites are rendered in blue background and matched metabolites are highlighted in different colors from yellow to red based on their p values. The pathway, as displayed by MetPA, is zoomable and all nodes are clickable and hyperlinked.

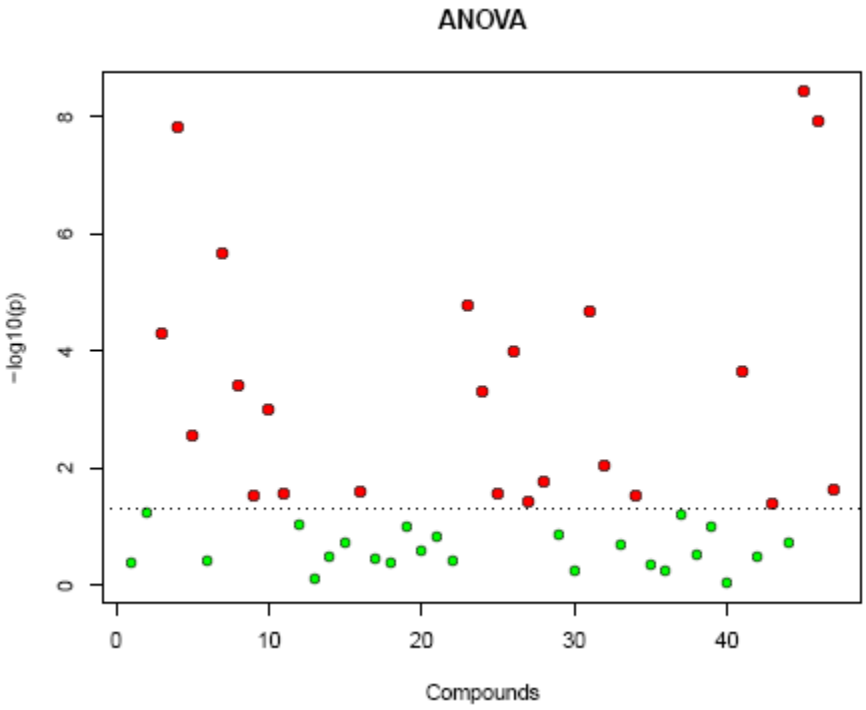

Figure 6.5 Significant compounds identified using ANOVA (real data).

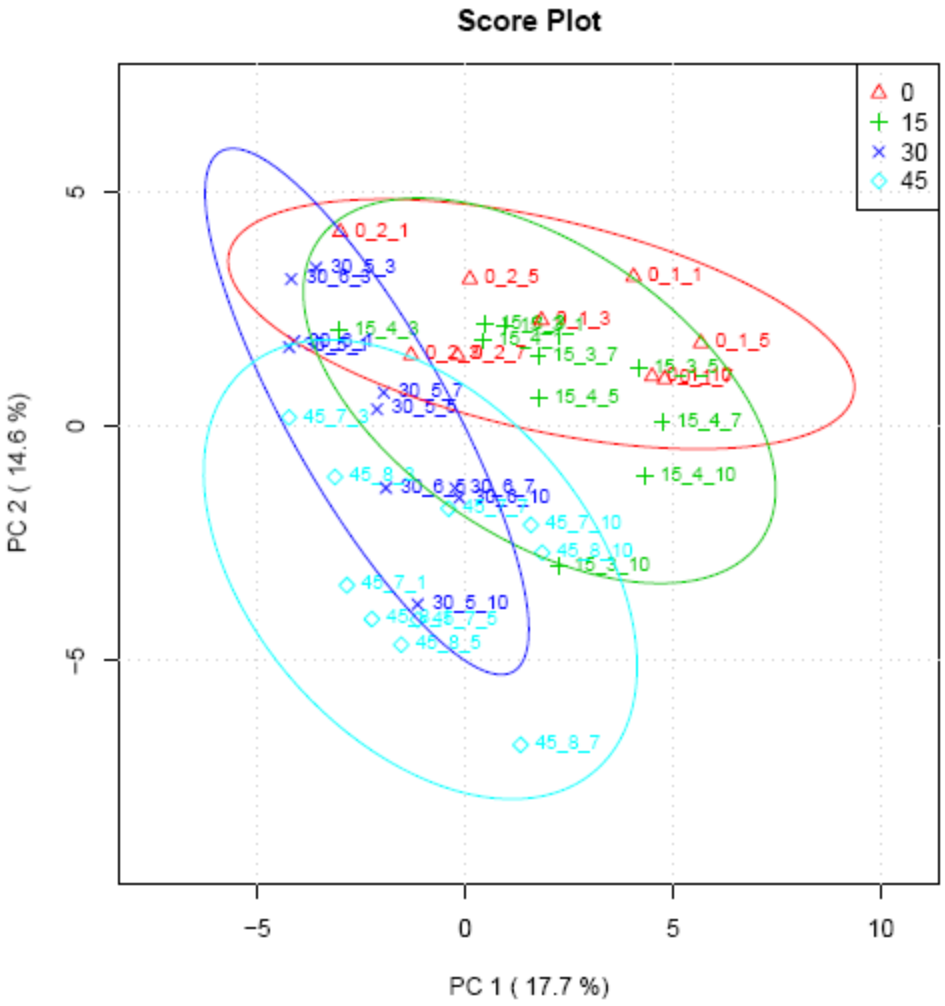

Figure 6.6 PCA 2D score plot (real data).

The ellipses indicate the 95% confidence area.

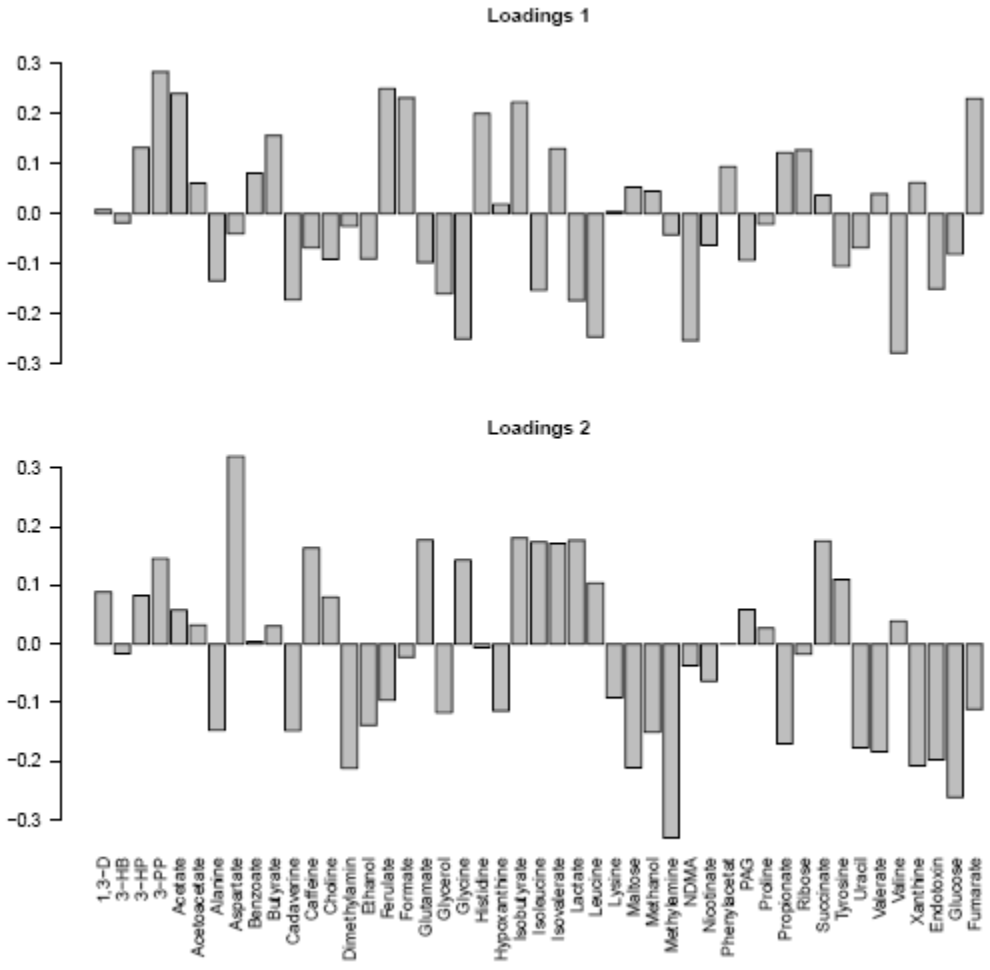

Figure 6.7 PCA loading plot for PC1 and PC2 (real data).

Figure 6.8 MetaboAnalyst user profile

A) MetaboAnalyst has attracted over 23,000 visits by over 6,000 distinct visitors from over 1200 cities worldwide, with over 80 visits per day (as of July 1, 2011); B) Weekly user traffic of the last two years. The drop indicated by the arrow corresponds to the Christmas holiday.

Chapter 7

General Conclusions & Future Work

7.1 General Conclusions

My thesis project has been centered on three computational challenges in metabolomic data analysis – 1) compound identification in complex biofluids; 2) metabolomic data processing and analysis; and 3) functional interpretation of metabolomic data. To address these issues, I have implemented a series of four freely available and user-friendly software tools – MetaboMiner, MetaboAnalyst, MSEA, and MetPA. Each of these tools has been designed to address a particular aspect of the data analysis issues mentioned above.

7.1.1 Compound Identification from 2D NMR with MetaboMiner

MetaboMiner was designed for efficient compound identification from 2D NMR spectra of complex biofluid mixtures. As there are many analytical platforms currently employed in metabolomics studies, I chose to focus on NMR, specifically 2D NMR. This was done for two main reasons - NMR spectra are highly reproducible and 2D (TOCSY and HSQC) spectra contain substantially more compound structural information than 1D proton NMR spectra. A major drawback associated with this choice is that the sensitivity of 2D NMR when acquired at natural abundance is very low, thus many signals may not be detected in normal biofluid mixtures.

MetaboMiner is a desktop application, written in the *Java* programming language. It supports both automated peak matching as well as spectral visualization for manual annotation. Tests using both synthetic and real spectra of

compound mixtures showed that MetaboMiner is able to automatically identify ~80% of metabolites from good quality 2D NMR spectra. This performance is substantially better (>20%) than any other currently available programs for 2D NMR mixture analysis.

The key idea behind MetaboMiner is the use of a knowledgebase - a pre-defined spectral library of compounds corresponding to the known metabolite composition of the biofluids of interests. This approach greatly reduces the size of the search space and therefore, reduces the number of false positives. Another important concept is the implementation of a self-adaptive threshold search algorithm, which automatically expands or shrinks the search space to maximize peak matches without introducing many false positives. The library and algorithm have since been incorporated in the NMR spectral search engine of the latest HMDB release (25). The most recent version of MetaboMiner can be downloaded from the project home page http://wishart.biology.ualberta.ca/metabominer/.

MetaboMiner is primarily concerned with compound identification based on 2D NMR spectra. For 1D proton NMR, the spectra are somewhat more crowded and manual fitting is usually required in order to get a reliable result. However, some promising approaches have been reported for analyzing 1D spectra that employ the cross entropy method exploiting (partial) decomposability (CEED) (224).

7.1.2 General Data Processing and Analysis with MetaboAnalyst

MetaboAnalyst is a web-based application that provides a broad array of options for metabolomic data processing, analysis, and annotation. MetaboAnalyst was originally designed for quantitative metabolomics, and later expanded to support the analysis of raw spectral data as well as peak lists data as might be generated using chemometric approaches. The first release of MetaboAnalyst only supported two-group discrimination analysis. Multi-group data analysis support was implemented with the most recent release. The latest release also incorporated a new functional module for two-factor and time-series metabolomic data analysis.

The analysis procedures in MetaboAnalyst have been implemented as a series of functional modules – the data upload module, the data preprocessing module, the feature selection module, the report generation module, *etc.* The design of MetaboAnalyst was inspired and influenced by the popular open-source genomics data analysis pipeline - GenePattern (182). The three main elements found in GenePattern - *Analysis and Visualization*, *Data Pipelines,* and *Servers,* were closely followed throughout the design and implementation of MetaboAnalyst. This idea has proven to be very useful for bench biologists as GenePattern is widely regarded as being easy to understand and easy to use. It is also very useful for the tool developer, as each functional module can be developed and debugged independently. Furthermore, new functions can be easily introduced by simply adding a new module.

MetaboAnalyst supports a comprehensive array of analytical methods that includes biomarker identification, pattern discovery, and classification. For biomarker identification, both the standard univariate tests (t-tests, ANOVA) and the moderated approaches (SAM, Limma eBayes) have been implemented. The choice of these standard analytical tools was based on the fact that metabolomics experiments usually have more samples than what is available for microarray experiments. Standard tests often work well and are easy to use for many biologists who are unfamiliar with the statistical concepts specifically developed for microarray data. For pattern discovery, MetaboAnalyst supports PCA and several other clustering methods (i.e. hierarchical clustering and SOM). Bicluster analysis is currently very computationally intensive, and is not suitable for web-based application. For classification, MetaboAnalyst offers the three widely used approaches that are known to work well for high-dimensional data including dimensional reduction (PLS-DA), ensemble methods (random forest), and shrinkage methods (soft-margin SVM).

One of the more useful features in MetaboAnalyst is its support for interactive data visualization. Users can adjust key parameters for most methods and visualize the results immediately. These analysis and visualization capabilities were enabled by using the powerful R statistical environment while the data pipeline and servers were realized by using the Java Server Faces (JSF) framework. This allowed me to seamlessly encapsulate different modules into a streamlined workflow. MetaboAnalyst is hosted on our local server with regular updates and maintenance (http://www.metaboanalyst.ca).

7.1.3 High-level Data Interpretation with MSEA and MetPA

MSEA is designed to perform metabolite set enrichment analysis using seven pre-defined libraries containing a total of 6292 functionally-related metabolite sets. MSEA can help identify biologically meaningful patterns that are obvious as well as subtle but coordinated changes occurring within quantitative metabolomic data sets.

In some cases, particularly for well-studied model organisms, high-quality metabolite sets also contain extra information about the functional relationship among different member compounds in the form of metabolic pathway information. This extra information can be used to further refine metabolomic results so that they are more aligned with a domain expert's manual interpretation. MetPA is designed to address this aspect of metabolomic data interpretation by taking into account of both the enrichment score and topological information (in the corresponding pathways) into account. In addition, MetPA also attempts to address another important issue – the visualization of a large amount of data in a web-based server. This is usually only possible with desktop applications as the interactive nature and the large size of the graphic images. To overcome this problem, I have developed a Google-map style visualization system using Ajax technology.

7.2 Summary & Future Perspectives

Metabolomics is still a relatively new member of the omics family. The field of metabolomics is rapidly growing and it is finding applications in many different

fields, from disease diagnosis, drug toxicity assessment, to environmental monitoring. My PhD research has allowed me to explore different analytical, biological, statistical, and computational aspects of metabolomics. During this time I developed several widely used software tools for compound identification (MetaboMiner), statistical analysis (MetaboAnalyst), metabolite set enrichment analysis (MSEA), and metabolic pathway analysis (MetPA). Together, they constitute a coherent and comprehensive solution for many bioinformatics issues encountered with many of today's metabolomic studies.

However, metabolomics on its own does not allow one to explore all aspects of a biological system. Genomics, transcriptomics, and proteomics also play key roles in providing detailed information on the changes in the genes and proteins arising from various perturbations, diseases or environmental stimuli. Combining and analyzing omics data from multiple omics platforms is the basis to systems biology. However, the analysis tools for metabolomics, transcriptomics, genomics and proteomics have largely been developed independently from each other. As a result, one of the major challenges facing bench biologists is finding ways to effectively integrate data from these different omics platforms to obtain a comprehensive understanding of biological systems. The next major focus in the field of bioinformatics is to develop powerful computational tools to support integrative analysis across different omics platforms. Different kinds of omics data need to be integrated at a high-level, where they can be statistically characterized, intuitively visualized, and presented in an appropriate biological

context to facilitate knowledge discovery and functional interpretation. There should be at least three essential components:

a) Specially-designed algorithms which can handle multiple heterogeneous omics datasets. These data sets usually consist of multiple data matrices of very high-dimensional nature, making most conventional analysis methods insufficient. Recently, several promising multivariate statistical algorithms have been proposed for data summarization, variable selection, correlation, and association analysis with multi-omics data (225,226).

b) A comprehensive knowledge base on genes, proteins, and metabolites. I believe metabolic networks, most of which are well established, provide an effective scaffold for organizing systems biology data. For instance, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and the human metabolome database (HMDB) are good starting points for this purpose.

c) An efficient data visualization system to present biological networks and statistical results. There are several promising tools for network visualization. For instance, the Google-map style network visualization system I have implemented for metabolic pathways visualization can be extended to much more general applications. The recently available web-version of Cytoscape (227) is another viable option. For statistical visualization, advanced on-line interactive 3D visualization using techniques based on virtual reality markup language (VRML: http://www.w3.org/MarkUp/VRML/) or Java LiveGraphics3D (http://www.vis.uni-stuttgart.de/~kraus/LiveGraphics3D) could be a very effective

and powerful presentation tool. The arrival of HTML5 (228) is expected to greatly facilitate the development of web-based data visualization programs.

As modern biology is increasingly becoming dependent on using different omics technologies to explore biological responses or to measure biological phenotypes, the interpretation of these omics data sets relies crucially on using sophisticated, computerized approaches for compiling, analyzing, and visualizing these data. I expect that the development of bioinformatics tools for integrative data analysis and visualization will be the next major effort in bioinformatics and systems biology.

REFERENCES

1. Fiehn, O. (2002) Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol*, **48**, 155-171.

2. Wishart, D.S. (2008) Quantitative metabolomics using NMR. *Trac-Trends in Analytical Chemistry*, **27**, 228-237.

3. Dunn, W.B. and Ellis, D.I. (2005) Metabolomics: Current analytical platforms and methodologies. *Trac-Trends in Analytical Chemistry*, **24**, 285-294.

4. Trygg, J., Holmes, E. and Lundstedt, T. (2007) Chemometrics in metabonomics. *J Proteome Res*, **6**, 469-479.

5. Weckwerth, W. and Morgenthal, K. (2005) Metabolomics: from pattern recognition to biological interpretation. *Drug Discov Today*, **10**, 1551-1558.

6. Holmes, E., Wilson, I.D. and Nicholson, J.K. (2008) Metabolic phenotyping in health and disease. *Cell*, **134**, 714-717.

7. Wishart, D.S. (2005) Metabolomics: the principles and potential applications to transplantation. *Am J Transplant*, **5**, 2814-2820.

8. Wishart, D.S. (2008) Applications of metabolomics in drug discovery and development. *Drugs R D*, **9**, 307-322.

9. Kaddurah-Daouk, R., Kristal, B.S. and Weinshilboum, R.M. (2008) Metabolomics: a global biochemical approach to drug response and disease. *Annu Rev Pharmacol Toxicol*, **48**, 653-683.

10. Pauling, L., Robinson, A.B., Teranishi, R. and Cary, P. (1971) Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc Natl Acad Sci U S A*, **68**, 2374-2376.

11. Teranishi, R., Mon, T.R., Robinson, A.B., Cary, P. and Pauling, L. (1972) Gas chromatography of volatiles from breath and urine. *Anal Chem*, **44**, 18-20.

12. Hoult, D.I., Busby, S.J., Gadian, D.G., Radda, G.K., Richards, R.E. and Seeley, P.J. (1974) Observation of tissue metabolites using 31P nuclear magnetic resonance. *Nature*, **252**, 285-287.

13. Ackerman, J.J., Grove, T.H., Wong, G.G., Gadian, D.G. and Radda, G.K. (1980) Mapping of metabolites in whole animals by 31P NMR using surface coils. *Nature*, **283**, 167-170.

14. Bottomley, P.A. (1987) Spatial localization in NMR spectroscopy in vivo. *Ann N Y Acad Sci*, **508**, 333-348.

15. Spraul, M., Neidig, P., Klauck, U., Kessler, P., Holmes, E., Nicholson, J.K., Sweatman, B.C., Salman, S.R. *et al.* (1994) Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *J Pharm Biomed Anal*, **12**, 1215-1225.

16. Holmes, E., Foxall, P.J., Nicholson, J.K., Neild, G.H., Brown, S.M., Beddell, C.R., Sweatman, B.C., Rahr, E. *et al.* (1994) Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal Biochem*, **220**, 284-296.

17. Holmes, E., Nicholls, A.W., Lindon, J.C., Ramos, S., Spraul, M., Neidig, P., Connor, S.C., Connelly, J. *et al.* (1998) Development of a model for classification of toxin-induced lesions using 1H NMR spectroscopy of urine combined with pattern recognition. *NMR Biomed*, **11**, 235-244.

18. Moka, D., Vorreuther, R., Schicha, H., Spraul, M., Humpfer, E., Lipinski, M., Foxall, P.J., Nicholson, J.K. *et al.* (1998) Biochemical classification of kidney carcinoma biopsy samples using magic-angle-spinning 1H nuclear magnetic resonance spectroscopy. *J Pharm Biomed Anal*, **17**, 125-132.

19. Beckwith-Hall, B.M., Nicholson, J.K., Nicholls, A.W., Foxall, P.J., Lindon, J.C., Connor, S.C., Abdi, M., Connelly, J. *et al.* (1998) Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chem Res Toxicol*, **11**, 260-272.

20. Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol*, **16**, 373-378.

21. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Research*, **35**, D521-D526.

22. Lundberg, P., Vogel, T., Malusek, A., P.-O., L., Cohen, L. and Dahlqvist, O. (2005) MDL - The Magnetic Resonance Metabolomics Database (mdl.imv.liu.se). *ESMRMB, Basel, Switzerland*.

23. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R. *et al.* (2005) METLIN - A metabolite mass spectral database. *Therapeutic Drug Monitoring*, **27**, 747-751.

24. Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R. *et al.* (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology*, **26**, 162-164.

25. Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res*, **37**, D603-610.

26. Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N.S., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, **19**, 45-50.

27. Kuhara, T. (2005) Gas chromatographic-mass spectrometric urinary metabolome analysis to study mutations of inborn errors of metabolism. *Mass Spectrom Rev*, **24**, 814-827.

28. Boverhof, D.R. and Zacharewski, T.R. (2006) Toxicogenomics in risk assessment: applications and needs. *Toxicol Sci*, **89**, 352-360.

29. Lindon, J.C., Nicholson, J.K., Holmes, E., Antti, H., Bollard, M.E., Keun, H., Beckonert, O., Ebbels, T.M. *et al.* (2003) Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol Appl Pharmacol*, **187**, 137-146.

30. German, J.B., Watkins, S.M. and Fay, L.B. (2005) Metabolomics in practice: emerging knowledge to guide future dietetic advice toward individualized health. *J Am Diet Assoc*, **105**, 1425-1432.

31. Weckwerth, W. (2003) Metabolomics in systems biology. *Annu Rev Plant Biol*, **54**, 669-689.

32. Jaffer, F.A. and Weissleder, R. (2005) Molecular imaging in the clinical arena. *Jama*, **293**, 855-862.

33. Hieter, P. and Boguski, M. (1997) Functional genomics: it's all how you read it. *Science*, **278**, 601-602.

34. Fu, J., Keurentjes, J.J., Bouwmeester, H., America, T., Verstappen, F.W., Ward, J.L., Beale, M.H., de Vos, R.C. *et al.* (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet*, **41**, 166-167.

35. Fernie, A.R. and Schauer, N. (2009) Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet*, **25**, 39-48.

36. Gieger, C., Geistlinger, L., Altmaier, E., Hrabe de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.W., Wichmann, H.E. *et al.* (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet*, **4**, e1000282.

37. Illig, T., Gieger, C., Zhai, G., Romisch-Margl, W., Wang-Sattler, R., Prehn, C., Altmaier, E., Kastenmuller, G. *et al.* (2010) A genome-wide perspective of genetic variation in human metabolism. *Nat Genet*, **42**, 137-141.

38. Fell, D.A. (2005) Enzymes, metabolites and fluxes. *J Exp Bot*, **56**, 267-272.

39. Kell, D.B. and Westerhoff, H.V. (1986) Towards a Rational Approach to the Optimization of Flux in Microbial Biotransformations. *Trends in Biotechnology*, **4**, 137-142.

40. ter Kuile, B.H. and Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett*, **500**, 169-171.

41. Eldjarn, L., Jellum, E. and Stokke, O. (1974) Application of gas chromatography-mass spectrometry in routine and research in clinical chemistry. *J Chromatogr*, **91**, 353-366.

42. Nicholson, J.K., O'Flynn, M.P., Sadler, P.J., Macleod, A.F., Juul, S.M. and Sonksen, P.H. (1984) Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochem J*, **217**, 365-375.

43. Salek, R.M., Maguire, M.L., Bentley, E., Rubtsov, D.V., Hough, T., Cheeseman, M., Nunez, D., Sweatman, B.C. *et al.* (2007) A metabolomic comparison of urinary changes in type 2 diabetes in mouse, rat, and human. *Physiol Genomics*, **29**, 99-108.

44. Makinen, V.P., Soininen, P., Forsblom, C., Parkkonen, M., Ingman, P., Kaski, K., Groop, P.H. and Ala-Korpela, M. (2006) Diagnosing diabetic nephropathy by 1H NMR metabonomics of serum. *Magma*, **19**, 281-296.

45. Makinen, V.P., Soininen, P., Forsblom, C., Parkkonen, M., Ingman, P., Kaski, K., Groop, P.H. and Ala-Korpela, M. (2008) 1H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol Syst Biol*, **4**, 167.

46. Sabatine, M.S., Liu, E., Morrow, D.A., Heller, E., McCarroll, R., Wiegand, R., Berriz, G.F., Roth, F.P. *et al.* (2005) Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, **112**, 3868-3875.

47. Shah, S.H., Bain, J.R., Muehlbauer, M.J., Stevens, R.D., Crosslin, D.R., Haynes, C., Dungan, J., Newby, L.K. *et al.* (2010) Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. *Circ Cardiovasc Genet*, **3**, 207-214.

48. Gulston, M.K., Titman, C.M. and Griffin, J.L. (2007) Applications of metabolomics to understanding obesity in mouse and man. *Biomark Med*, **1**, 575-582.

49. Kristal, B.S., Shurubor, Y.I., Kaddurah-Daouk, R. and Matson, W.R. (2007) Metabolomics in the study of aging and caloric restriction. *Methods Mol Biol*, **371**, 393-409.

50. Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R. *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, **457**, 910-914.

51. Lindon, J.C., Keun, H.C., Ebbels, T.M., Pearce, J.M., Holmes, E. and Nicholson, J.K. (2005) The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics*, **6**, 691-699.

52. Simpson, M.J. and McKelvie, J.R. (2009) Environmental metabolomics: new insights into earthworm ecotoxicity and contaminant bioavailability in soil. *Anal Bioanal Chem*, **394**, 137-149.

53. Hines, A., Staff, F.J., Widdows, J., Compton, R.M., Falciani, F. and Viant, M.R. (2010) Discovery of metabolic signatures for predicting whole organism toxicology. *Toxicol Sci*, **115**, 369-378.

54. Scalbert, A., Manach, C., Morand, C., Remesy, C. and Jimenez, L. (2005) Dietary polyphenols and the prevention of diseases. *Crit Rev Food Sci Nutr*, **45**, 287-306.

55. Khachik, F., Carvalho, L., Bernstein, P.S., Muir, G.J., Zhao, D.Y. and Katz, N.B. (2002) Chemistry, distribution, and metabolism of tomato carotenoids and their impact on human health. *Exp Biol Med (Maywood)*, **227**, 845-851.

56. Manach, C., Hubert, J., Llorach, R. and Scalbert, A. (2009) The complex links between dietary phytochemicals and human health deciphered by metabolomics. *Mol Nutr Food Res*, **53**, 1303-1315.

57. Solanky, K.S., Bailey, N.J., Beckwith-Hall, B.M., Bingham, S., Davis, A., Holmes, E., Nicholson, J.K. and Cassidy, A. (2005) Biofluid 1H NMR-based metabonomic techniques in nutrition research - metabolic effects of dietary isoflavones in humans. *J Nutr Biochem*, **16**, 236-244.

58. Wang, Y., Tang, H., Nicholson, J.K., Hylands, P.J., Sampson, J. and Holmes, E. (2005) A metabonomic strategy for the detection of the metabolic effects of chamomile (Matricaria recutita L.) ingestion. *J Agric Food Chem*, **53**, 191-196.

59. Mennen, L.I., Sapinho, D., Ito, H., Bertrais, S., Galan, P., Hercberg, S. and Scalbert, A. (2006) Urinary flavonoids and phenolic acids as biomarkers of intake for polyphenol-rich foods. *Br J Nutr*, **96**, 191-198.

60. Fardet, A., Llorach, R., Orsoni, A., Martin, J.F., Pujos-Guillot, E., Lapierre, C. and Scalbert, A. (2008) Metabolomics provide new insight on the metabolism of dietary phytochemicals in rats. *J Nutr*, **138**, 1282-1287.

61. Simmons-Boyce, J.L., Purcell, S.L., Nelson, C.M. and MacKinnon, S.L. (2009) Dietary Ascophyllum nodosum increases urinary excretion of tricarboxylic acid cycle intermediates in male Sprague-dawley rats. *J Nutr*, **139**, 1487-1494.

62. German, J.B., Bauman, D.E., Burrin, D.G., Failla, M.L., Freake, H.C., King, J.C., Klein, S., Milner, J.A. *et al.* (2004) Metabolomics in the opening decade of the 21st century: building the roads to individualized health. *J Nutr*, **134**, 2729-2732.

63. Edwards, J.S. and Palsson, B.O. (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, **97**, 5528-5533.

64. Forster, J., Famili, I., Fu, P., Palsson, B.O. and Nielsen, J. (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res*, **13**, 244-253.

65. Griffin, J.L. (2006) The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos Trans R Soc Lond B Biol Sci*, **361**, 147-161.

66. Sundararaj, S., Guo, A., Habibi-Nazhad, B., Rouani, M., Stothard, P., Ellison, M. and Wishart, D.S. (2004) The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli. *Nucleic Acids Res*, **32**, D293-295.

67. Nobata, C., Dobson, P.D., Iqbal, S.A., Mendes, P., Tsujii, J., Kell, D.B. and Ananiadou, S. (2010) Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, **4**, 393-405.

68. Kind, T., Scholz, M. and Fiehn, O. (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One*, **4**, e5440.

69. Wishart, D.S., Lewis, M.J., Morrissey, J.A., Flegel, M.D., Jeroncic, K., Xiong, Y.P., Cheng, D., Eisner, R. *et al.* (2008) The human cerebrospinal fluid metabolome. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences*, **871**, 164-173.

70. Psychogios, N., Hau, D., Peng, J., Guo, A., Mandal, R., Bouatra, S., Sinelnikov, I., Krishnamurthy, R. *et al.* (2010) The Human Serum Metabolome. *PLoS One*, **6(2)**.

71. Oliver, S.G. (2002) Functional genomics: lessons from yeast. *Philos Trans R Soc Lond B Biol Sci*, **357**, 17-23.

72. Slupsky, C.M., Rankin, K.N., Wagner, J., Fu, H., Chang, D., Weljie, A.M., Saude, E.J., Lix, B. *et al.* (2007) Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Anal Chem*, **79**, 6995-7004.

73. Assfalg, M., Bertini, I., Colangiuli, D., Luchinat, C., Schafer, H., Schutz, B. and Spraul, M. (2008) Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U S A*, **105**, 1420-1424.

74. Holmes, E., Loo, R.L., Stamler, J., Bictash, M., Yap, I.K., Chan, Q., Ebbels, T., De Iorio, M. *et al.* (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*.

75. Hollywood, K., Brison, D.R. and Goodacre, R. (2006) Metabolomics: current technologies and future trends. *Proteomics*, **6**, 4716-4723.

76. Yang, Y., Li, C., Nie, X., Feng, X., Chen, W., Yue, Y., Tang, H. and Deng, F. (2007) Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning 1H NMR spectroscopy in conjunction with multivariate data analysis. *J Proteome Res*, **6**, 2605-2614.

77. Pfeuffer, J., Tkac, I., Provencher, S.W. and Gruetter, R. (1999) Toward an in vivo neurochemical profile: Quantification of 18 metabolites in short-echo-time H-1 NMR spectra of the rat brain. *Journal of Magnetic Resonance*, **141**, 104-120.

78. Van, Q.N., Issaq, H.J., Jiang, Q., Li, Q., Muschik, G.M., Waybright, T.J., Lou, H., Dean, M. *et al.* (2008) Comparison of 1D and 2D NMR spectroscopy for metabolic profiling. *J Proteome Res*, **7**, 630-639.

79. Keun, H.C., Ebbels, T.M., Antti, H., Bollard, M.E., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U. *et al.* (2002) Analytical reproducibility in (1)H NMR-based metabonomic urinalysis. *Chem Res Toxicol*, **15**, 1380-1386.

80. Dumas, M.E., Maibaum, E.C., Teague, C., Ueshima, H., Zhou, B.F., Lindon, J.C., Nicholson, J.K., Stamler, J. *et al.* (2006) Assessment of analytical reproducibility of H-1 NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry*, **78**, 2199-2208.

81. Dunn, W.B., Bailey, N.J. and Johnson, H.E. (2005) Measuring the metabolome: current analytical technologies. *Analyst*, **130**, 606-625.

82. Ceglarek, U., Leichtle, A., Brugel, M., Kortz, L., Brauer, R., Bresler, K., Thiery, J. and Fiedler, G.M. (2009) Challenges and developments in tandem mass spectrometry based clinical metabolomics. *Mol Cell Endocrinol*, **301**, 266-271.

83. Martin, D.B., Holzman, T., May, D., Peterson, A., Eastham, A., Eng, J. and McIntosh, M. (2008) MRMer, an interactive open source and cross-platform system for data extraction and visualization of multiple reaction monitoring experiments. *Mol Cell Proteomics*, **7**, 2270-2278.

84. Guo, K. and Li, L. (2010) High-Performance Isotope Labeling for Profiling Carboxylic Acid-Containing Metabolites in Biofluids by Mass Spectrometry. *Anal Chem*.

85. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662-666.

86. Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279-284.

87. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403-410.

88. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank. *Nucleic Acids Research*, **36**, D25-D30.

89. Green, P. (1999) Phrap, version 0.990329. *http://phrap.org*.

90. Green, P. and Ewing, B. (2002) Phred, version 0.020425c. *http://phrap.org*.

91. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, **12**, 1611-1618.

92. Tao, W.A. and Aebersold, R. (2003) Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr Opin Biotechnol*, **14**, 110-118.

93. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.

94. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.

95. Xia, J. (2006) Expressed Sequence Tags (EST) analysis, annotation and immune gene identification from a spleen cDNA library in the duck (Anas platyrhynchos). *Master's Thesis, University of Alberta, Edmonton*, 129 p.

96. McPherson, J.D. (2009) Next-generation gap. *Nat Methods*, **6**, S2-5.

97. Zhao, Q., Stoyanova, R., Du, S., Sajda, P. and Brown, T.R. (2006) HiRes--a tool for comprehensive assessment and interpretation of metabolomic data. *Bioinformatics*, **22**, 2562-2564.

98. Wang, T., Shao, K., Chu, Q., Ren, Y., Mu, Y., Qu, L., He, J., Jin, C. *et al.* (2009) Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinformatics*, **10**, 83.

99. Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR*, **6**, 277-293.

100. Goddard, T.D. and Kneller, D.G. (2006) SPARKY 3. *University of California, San Francisco*.

101. Lewis, I.A., Schommer, S.C. and Markley, J.L. (2009) rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem*, **47 Suppl 1**, S123-126.

102. Zheng, M., Lu, P., Liu, Y., Pease, J., Usuka, J., Liao, G. and Peltz, G. (2007) 2D NMR metabonomic analysis: a novel method for automated peak alignment. *Bioinformatics*, **23**, 2926-2933.

103. Sandusky, P. and Raftery, D. (2005) Use of selective TOCSY NMR experiments for quantifying minor components in complex mixtures: application to the metabonomics of amino acids in honey. *Anal Chem*, **77**, 2455-2463.

104. Massou, S., Nicolas, C., Letisse, F. and Portais, J.C. (2007) Application of 2D-TOCSY NMR to the measurement of specific(13C-enrichments in complex mixtures of 13C-labeled metabolites. *Metab Eng*, **9**, 252-257.

105. Xi, Y., de Ropp, J.S., Viant, M.R., Woodruff, D.L. and Yu, P. (2008) Improved identification of metabolites in complex mixtures using HSQC NMR spectroscopy. *Anal Chim Acta*, **614**, 127-133.

106. Stein, S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, **10**, 770-781.

107. Fiehn, O., Wohlgemuth, G. and Scholz, M. (2005) Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. . *Proc. Lect. Notes Bioinformatics*, **3615**, 224-239.

108. Lommen, A. (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem*, **81**, 3079-3086.

109. Broeckling, C.D., Reddy, I.R., Duran, A.L., Zhao, X. and Sumner, L.W. (2006) MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Anal Chem*, **78**, 4334-4341.

110. Duran, A.L., Yang, J., Wang, L.J. and Sumner, L.W. (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, **19**, 2283-2293.

111. Luedemann, A., Strassburg, K., Erban, A. and Kopka, J. (2008) TagFinder for the quantitative analysis of gas chromatography--mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics*, **24**, 732-737.

112. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem*, **78**, 779-787.

113. Pluskal, T., Castillo, S., Villar-Briones, A. and Oresic, M. (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

114. Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem*, **78**, 4281-4290.

115. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. and van der Werf, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.

116. Baldi, P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509-519.

117. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98**, 5116-5121.

118. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**, Article3.

119. Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G. and Guedj, M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS One*, **5**, e12336.

120. Lu, Y., Liu, P.Y., Xiao, P. and Deng, H.W. (2005) Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21**, 3105-3113.

121. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat Rev Genet*, **2**, 418-427.

122. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863-14868.

123. Madeira, S.C. and Oliveira, A.L. (2004) Biclustering algorithms for biological data analysis: a survey *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**, 24-45.

124. Wu, C.J. and Kasif, S. (2005) GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res*, **33**, W596-599.

125. Luscher, A., Csardi, G., de Lachapelle, A.M., Kutalik, Z., Peter, B. and Bergmann, S. (2010) ExpressionView--an interactive viewer for modules identified in gene expression data. *Bioinformatics*, **26**, 2062-2063.

126. Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Statist. Sinica*, **12**, 61-86.

127. Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal Chem*, **78**, 567-574.

128. Westerhuis, C.A., Hoefsloot, C.J.H., Smit, S., Vis, J.D., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M. and van Dorsten, F.A. (2007) Assessment of PLSDA cross validation. *Metabolomics*, **4**, 81-89.

129. Diaz-Uriarte, R. and Alvarez de Andres, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

130. Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5-32.

131. Lunetta, K.L., Hayward, L.B., Segal, J. and Van Eerdewegh, P. (2004) Screening large-scale association study data: exploiting interactions using random forests. *Bmc Genetics*, **5**, -.

132. Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

133. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.

134. Mahadevan, S., Shah, S.L., Marrie, T.J. and Slupsky, C.M. (2008) Analysis of metabolomic data using support vector machines. *Anal Chem*, **80**, 7562-7570.

135. Noble, W.S. (2006) What is a support vector machine? *Nat Biotechnol*, **24**, 1565-1567.

136. Zhang, X., Lu, X., Shi, Q., Xu, X.Q., Leung, H.C., Harris, L.N., Iglehart, J.D., Miron, A. *et al.* (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**, 197.

137. Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389-422.

138. Becker, N., Werft, W., Toedt, G., Lichter, P. and Benner, A. (2009) penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics*, **25**, 1711-1712.

139. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. and Krawetz, S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98-104.

140. Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C. and Romero, R. (2007) A systems biology approach for pathway level analysis. *Genome Res*, **17**, 1537-1545.

141. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E. *et al.* (2003) PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267-273.

142. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

143. Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93-99.

144. Goeman, J.J. and Buhlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-987.

145. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587-3595.

146. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, **36**, D480-484.

147. Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B. *et al.* (2010) SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res*, **38**, D480-487.

148. Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, **36**, D623-631.

149. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, **37**, D619-622.

150. Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75-82.

151. Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A. (2010) TopoGSA: network topological gene set analysis. *Bioinformatics*, **26**, 1271-1272.

152. Goffard, N. and Weiller, G. (2007) PathExpress: a web-based tool to identify relevant pathways in gene expression data. *Nucleic Acids Res*, **35**, W176-181.

153. Benson, D.A., Boguski, M.S., Lipman, D.J. and Ostell, J. (1997) GenBank. *Nucleic Acids Res*, **25**, 1-6.

154. Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) A relational database for sequence-specific protein NMR data. *J Biomol NMR*, **1**, 217-236.

155. Austin, C.P., Brady, L.S., Insel, T.R. and Collins, F.S. (2004) NIH Molecular Libraries Initiative. *Science*, **306**, 1138-1139.

156. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M. *et al.* (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, **36**, D344-350.

157. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**, D277-280.

158. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. and Palsson, B.O. (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, **7**, 129-143.

159. Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, **22**, 245-252.

160. Nicholson, J.K., Lindon, J.C. and Holmes, E. (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181-1189.

161. Crockford, D.J., Keun, H.C., Smith, L.M., Holmes, E. and Nicholson, J.K. (2005) Curve-fitting method for direct quantitation of compounds in complex biological mixtures using 1H NMR: application in metabonomic toxicology studies. *Anal Chem*, **77**, 4556-4562.

162. Viant, M.R., Rosenblum, E.S. and Tieerdema, R.S. (2003) NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. *Environ Sci Technol*, **37**, 4982-4989.

163. Cerdan, S., Kunnecke, B. and Seelig, J. (1990) Cerebral metabolism of [1,2-13C2]acetate as detected by in vivo and in vitro 13C NMR. *J Biol Chem*, **265**, 12916-12926.

164. Artemov, D., Bhujwalla, Z.M., Pilatus, U. and Glickson, J.D. (1998) Two-compartment model for determination of glycolytic rates of solid tumors by in vivo 13C NMR spectroscopy. *NMR Biomed*, **11**, 395-404.

165. Miccheli, A., Tomassini, A., Puccetti, C., Valerio, M., Peluso, G., Tuccillo, F., Calvani, M., Manetti, C. *et al.* (2006) Metabolic profiling by 13C-NMR spectroscopy: [1,2-13C2]glucose reveals a heterogeneous metabolism in human leukemia T cells. *Biochimie*, **88**, 437-448.

166. Viant, M.R., Pincetich, C.A., Hinton, D.E. and Tjeerdema, R.S. (2006) Toxic actions of dinoseb in medaka (Oryzias latipes) embryos as determined by in vivo 31P NMR, HPLC-UV and 1H NMR metabolomics. *Aquat Toxicol*, **76**, 329-342.

167. Griffin, J.L. (2003) Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Curr Opin Chem Biol*, **7**, 648-654.

168. Weljie, A.M., Newton, J., Mercier, P., Carlson, E. and Slupsky, C.M. (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem*, **78**, 4430-4442.

169. Holmes, E. and Antti, H. (2002) Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. *Analyst*, **127**, 1549-1557.

170. Schroeder, F.C., Gibson, D.M., Churchill, A.C., Sojikul, P., Wursthorn, E.J., Krasnoff, S.B. and Clardy, J. (2007) Differential analysis of 2D NMR spectra: new natural products from a pilot-scale fungal extract library. *Angew Chem Int Ed Engl*, **46**, 901-904.

171. Barrere, B., Peres, M., Gillet, B., Mergui, S., Beloeil, J.C. and Seylaz, J. (1990) 2D COSY 1H NMR: a new tool for studying in situ brain metabolism in the living animal. *FEBS Lett*, **264**, 198-202.

172. Kikuchi, J., Shinozaki, K. and Hirayama, T. (2004) Stable isotope labeling of Arabidopsis thaliana for an NMR-based metabolomics approach. *Plant Cell Physiol*, **45**, 1099-1104.

173. Ward, J.L., Baker, J.M. and Beale, M.H. (2007) Recent applications of NMR spectroscopy in plant metabolomics. *Febs J*, **274**, 1126-1131.

174. Wang, Y., Bollard, M.E., Keun, H., Antti, H., Beckonert, O., Ebbels, T.M., Lindon, J.C., Holmes, E. *et al.* (2003) Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning 1H nuclear magnetic resonance spectroscopy of liver tissues. *Anal Biochem*, **323**, 26-32.

175. Viant, M.R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem Biophys Res Commun*, **310**, 943-948.

176. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res*, **35**, D521-526.

177. Cui, Q., Lewis, I.A., Hegeman, A.D., Anderson, M.E., Li, J., Schulte, C.F., Westler, W.M., Eghbalnia, H.R. *et al.* (2008) Metabolite identification via the Madison Metabolomics Consortium Database. *Nat Biotechnol*, **26**, 162-164.

178. Chikayama, E., Sekiyama, Y., Hirayama, T., Saito, K., Shinozaki, K. and Kikuchi, J. (2006) SpinAssign : an Integrated Spectrum Analysis System for a Hetero-nuclear NMR-based Metabomics. *47th ENC Conference, Pacific Grove*.

179. Rosse, G., Neidig, P. and Schroder, H. (2002) Automated structure verification of small molecules libraries using 1D and 2D NMR techniques. *Methods Mol Biol*, **201**, 123-139.

180. Lewis, I.A., Schommer, S.C., Hodis, B., Robb, K.A., Tonelli, M., Westler, W.M., Sussman, M.R. and Markley, J.L. (2007) Method for determining molar concentrations of metabolites in complex solutions from two-dimensional 1H-13C NMR spectra. *Anal Chem*, **79**, 9385-9390.

181. Kupce, E. and Freeman, R. (2007) Fast multidimensional NMR by polarization sharing. *Magnetic Resonance in Chemistry*, **45**, 2-4.

182. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P. (2006) GenePattern 2.0. *Nat Genet*, **38**, 500-501.

183. Xia, J., Bjorndahl, T.C., Tang, P. and Wishart, D.S. (2008) MetaboMiner--semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507.

184. Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J.M., Santoyo, J. and Dopazo, J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research*, **31**, 3461-3467.

185. Rainer, J., Sanchez-Cabo, F., Stocker, G., Sturn, A. and Trajanoski, Z. (2006) CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Research*, **34**, W498-W503.

186. Steinfath, M., Groth, D., Lisec, J. and Selbig, J. (2008) Metabolite profile analysis: from raw data to regression and classification. *Physiol Plant*, **132**, 150-161.

187. Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. (2007) pcaMethods--a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164-1167.

188. Rubingh, C.M., Bijlsma, S., Derks, E.P.P.A., Bobeldijk, I., Verheij, E.R., Kochhar, S. and Smilde, A.K. (2006) Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics*, **2**, 53-61.

189. Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151-1160.

190. Burges, C.J.C. (1998) A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121-167.

191. Neuweger, H., Albaum, S.P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J. and Goesmann, A. (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, **24**, 2726-2732.

192. Lee, H.K., Braynen, W., Keshav, K. and Pavlidis, P. (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.

193. Barry, W.T., Nobel, A.B. and Wright, F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943-1949.

194. Dinu, I., Potter, J.D., Mueller, T., Liu, Q., Adewale, A.J., Jhangri, G.S., Einecke, G., Famulski, K.S. *et al.* (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

195. Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107-129.

196. Backes, C., Keller, A., Kuentzer, J., Kneissl, B., Comtesse, N., Elnakady, Y.A., Muller, R., Meese, E. *et al.* (2007) GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res*, **35**, W186-192.

197. Zheng, Q. and Wang, X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res*, **36**, W358-363.

198. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.

199. Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B. *et al.* SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res*, **38**, D480-487.

200. Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N. *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, **33**, 6083-6089.

201. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. and Mesirov, J.P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, **23**, 3251-3253.

202. Kim, S.Y. and Volsky, D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.

203. Hulsegge, I., Kommadath, A. and Smits, M.A. (2009) Globaltest and GOEAST: two different approaches for Gene Ontology analysis. *BMC Proc*, **3 Suppl 4**, S10.

204. Song, S. and Black, M.A. (2008) Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, **9**, 502.

205. Liu, Q., Dinu, I., Adewale, A.J., Potter, J.D. and Yasui, Y. (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.

206. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289-300.

207. Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, **37**, W652-660.

208. Neuweger, H., Albaum, S.P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J. and Goesmann, A. (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, **24**, 2726-2732.

209. Sansone, S.A., Fan, T., Goodacre, R., Griffin, J.L., Hardy, N.W., Kaddurah-Daouk, R., Kristal, B.S., Lindon, J. *et al.* (2007) The metabolomics standards initiative. *Nat Biotechnol*, **25**, 846-848.

210. Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, **4**, P3.

211. Salomonis, N., Hanspers, K., Zambon, A.C., Vranizan, K., Lawlor, S.C., Dahlquist, K.D., Doniger, S.W., Stuart, J. *et al.* (2007) GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, **8**, 217.

212. Albert, R. (2005) Scale-free networks in cell biology. *J Cell Sci*, **118**, 4947-4957.

213. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, **13**, 2498-2504.

214. Hummel, M., Meister, R. and Mansmann, U. (2008) GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78-85.

215. Aittokallio, T. and Schwikowski, B. (2006) Graph-based methods for analysing networks in cell biology. *Brief Bioinform*, **7**, 243-255.

216. Zhang, J.D. and Wiemann, S. (2009) KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, **25**, 1470-1471.

217. Berger, S.I., Iyengar, R. and Ma'ayan, A. (2007) AVIS: AJAX viewer of interactive signaling networks. *Bioinformatics*, **23**, 2803-2805.

218. Junker, B.H., Koschutzki, D. and Schreiber, F. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.

219. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

220. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, **37**, D885-890.

221. Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A. *et al.* (2010) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*.

222. Ametaj, B.N., Zebeli, Q., Saleem, F., Psychogios, N., Lewis, J.L., Dunn, S.M., Xia, J. and Wishart, D.S. (2010) Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows. *Metabolomics*, **5**, 375-386.

223. Eisner, R., Stretch, C., Eastman, T., Xia, J., Hau, D., Damaraju, S., Greiner, R., Wishart, D.S. *et al.* (2010) Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics*, **3**, 207-214.

224. Ravanbakhsh, S., Poczos, B. and Greiner, R. (2010) A Cross-Entropy Method that Optimizes Partially Decomposable Problems: A New Way to Interpret NMR Spectra. *National Conference on Artificial Intelligence (AAAI)*.

225. Le Cao, K.A., Gonzalez, I. and Dejean, S. (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855-2856.

226. de Tayrac, M., Le, S., Aubry, M., Mosser, J. and Husson, F. (2009) Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*, **10**, 32.

227. Lopes, C.T., Franz, M., Kazi, F., Donaldson, S.L., Morris, Q. and Bader, G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347-2348.

228. Boulos, M.N., Warren, J., Gong, J. and Yue, P. (2010) Web GIS in practice VIII: HTML5 and the canvas element for interactive online mapping. *Int J Health Geogr*, **9**, 14.

Appendix I

Using Web-based Tools for Metabolomic Data Analysis and Interpretation[5]

Analysis Overview

The procedure described here provides a step-by-step protocol for using MetaboAnalyst to fully analyze quantitative metabolomic data. It begins with a general overview of the program, followed by a detailed description on how to format and upload data, how to "cleanse" the data, how to normalize it and how to identify significant features or generate lists of "important metabolites". It concludes with a description on how to perform metabolite set enrichment analysis and how to perform metabolic pathway analysis. While the protocol is specific to MetaboAnalyst, many of the early-stage statistical steps can be readily adapted to other statistical analysis packages (such s SIMCA-P+ and SAS). As noted earlier, not all of MetaboAnalyst's options or data analysis paths will be discussed in detail. However, the protocol described here should be applicable to many common data analysis scenarios in metabolomics.

---

[5] The appendix is an excerpt from the following paper:

Xia, J. and Wishart, D.S. (2011) Web-based Inference of Biological Patterns, Functions and Pathways from Metabolomic Data using MetaboAnalyst. *Nature Protocols* 6, 743-760.

MetaboAnalyst consists of three main modules: 1) a data processing module; 2) a statistics module; and 3) a high-level functional interpretation module. The data processing module is responsible for data input, data processing and data normalization. The statistics module supports a number of statistical (univariate, multivariate) and machine learning methods for feature selection, clustering and classification. The high-level functional interpretation module includes enrichment analysis and pathway analysis. The enrichment analysis provides metabolite set enrichment analysis (MSEA) using several comprehensive metabolite-set libraries. The pathway analysis offers pathway enrichment analysis and pathway topology analysis via a Google-map style interactive pathway visualization system. As illustrated in **Fig. A1**, the data processing module is the entry-way to access the other two modules. The statistics module, which is perhaps the most important module in MetaboAnalyst, is designed for general-purpose metabolomic data analysis and can be used to analyze a number of different data types including compound concentration data, peak lists, or binned spectral data (i.e. both targeted and non-targeted data). For high-level functional interpretation, only quantitative metabolomic data (i.e. compound concentration data or a list of metabolite names) can be accepted. It is important to note that MetaboAnalyst's high-level functional analysis is organism-specific as dictated by the underlying knowledgebase. For enrichment analysis, the collection of ~6300 metabolite sets was compiled primarily from human studies. Therefore users need to provide their own custom metabolite sets if they wish to perform enrichment analysis for other organisms. MetaboAnalyst's pathway analysis

currently supports 15 model organisms with ~1200 pre-compiled KEGG pathways. Prior to using this option, users need to decide whether these predefined libraries are applicable to their organism(s) under study. To perform high-level functional analysis, one critical step is to match compound names between users' data and MetaboAnalyst's knowledgebase. As there are currently no universally accepted set of metabolite names or IDs, we have implemented an automated compound "disambiguator" to convert various compound IDs and synonyms to HMDB compound names for metabolite set enrichment analysis and to KEGG compound names for pathway analysis. In some cases, there will be redundancies and conflicts due to different naming schema adopted by different databases. Those compounds with name conflicts will be highlighted for subsequent manual inspection. We recommend that users try the recently released Chemical Translation Service (http://cts.fiehnlab.ucdavis.edu) to clarify these ambiguities before performing any kind of high-level analysis.

MetaboAnalyst uses a navigation tree to guide users through its different analysis procedures (**Fig. A2**). All the available functions are represented as tree nodes and these nodes are organized into different branches or functional categories. Users may click the corresponding nodes to navigate among different MetaboAnalyst functions. Depending on the context, some tree nodes may be disabled when the required preliminary steps have not been performed by the user. The current node is always highlighted during the analysis as shown in **Fig. A2**.

This protocol is organized into five sections: 1) data formatting, uploading and processing; 2) identifying important features using univariate analysis; 3) multivariate statistical analysis; 4) metabolite set enrichment analysis; and 5) metabolic pathway analysis. Two compound concentration datasets are provided to demonstrate these procedures. The first dataset contains metabolite concentrations of 39 bovine rumen samples measured by ^{1}H NMR. The rumen samples were collected from dairy cows fed with different proportions of barley grain. The samples are labeled in four groups - 0, 15, 30, and 45 - indicating different percentages of barley in the diet. The second dataset contains metabolite concentrations of 77 urine samples from cancer patients, also measured by ^{1}H NMR. The samples are divided into two groups -- control or cachexic (significant muscle loss).

Materials

Equipment Setup

- A PC with an internet connection;
- Browser requirements: MetaboAnalyst has been tested on all modern web browsers with JavaScript enabled, including Mozilla Firefox 3.0+, Safari 4.0+, Google-Chrome 5.0+, Opera 10.0+, and Internet Explorer 8.0.
- Data files: MetaboAnalyst has a number of example data sets for format illustration purposes as well as for testing purposes. Users can

directly select a testing data set in MetaboAnalyst's data upload page without actually downloading it. For this protocol, we will download a concentration data set and then re-upload it to better illustrate how local or user-generated data files may be handled. First go to the MetaboAnalyst home page (http://www.metaboanalyst.ca), and then click the "Data Formats" link on the left menu bar. In the Data Formats page, under the "Comma Separated Value (.csv) format", click and download the first concentration file - "Compound concentration data set - cow, four groups" and save it as "cow_diet.csv". The second concentration file to retrieve is "Compound concentration data set - human, two groups". Save this file as "human_cachexia.csv".

Procedures

Data Upload, Processing and Normalization *(Time: 5-10 minutes)*

1| *Starting up (Time: 10 sec).* Go to the MetaboAnalyst home page (http://www.metaboanalyst.ca) and click the "click here to start" link to enter the data upload page.

Critical Step: As most browsers support multiple tabs, do not access MetaboAnalyst from more than one tab during an analysis. Opening up multiple connections to MetaboAnalyst within the same browser will cause problems due to having the session data overwritten.

? TROUBLESHOOTING (SEE TABLE A2)

2| *Data upload (Time 1-2 min).* Depending on the type of analysis that a user wishes to

perform, they can upload their data via any of the three available tab options - Statistical Analysis, Enrichment Analysis, or Pathway Analysis (**Fig. A2**). Here, we show how to upload data from the "Statistical Analysis" tab which is selected by default (data upload instructions for Enrichment Analysis is provided at Steps 21-24, and data upload directions for Pathway Analysis is given at Step 32). In the "Upload your data" section, users can upload either a comma separated values (CSV) file or a compressed (ZIP) file. For the example data we use for this protocol, choose the "Concentrations" as the data type, and "Samples in rows (unpaired)" as the data format. Click the "**Browse**" button to locate the "cow_diet.csv" file and click the "**Submit**" button.

Critical Step: user must specify the correct data type and data format that match their data. Failure to do so will result in MetaboAnalyst launching the wrong data processing procedure.

Critical Step: Users can also easily perform paired analysis in MetaboAnalyst. For any kind of paired data comparison, there must be an even (2n) number of samples. For data in CSV format, the pair-wise information must be given by the class labels as integer values between -1 and -n/2 and between 1 and n/2. Samples with class labels having the same absolute integer values are considered to be pairs (i.e. -18 is paired with +18); For ZIP formatted data, users need to upload a separate text file (.txt) to give the pair information. Each pair is specified as two sample names (without a suffix) separated by a colon with one pair per row.

? TROUBLESHOOTING (SEE TABLE A2)

3| *Data integrity checking (Time: 20 sec to 5 min).* If the data has been uploaded successfully, a data integrity check is performed. After this check is completed, MetaboAnalyst will provide a summary of the data characteristics. Two common

issues that often arise with metabolomic data are missing values and outliers. To handle missing values, users can click the "Missing value imputation" button to use a variety of options to either exclude or replace these values. Outlier identification and removal is an iterative process and is usually performed in combination with preliminary data exploratory analysis. See Step 28 for an example. For this particular data set, we accept the data "as is" and so we will click the "Skip" button to go to the normalization step.

4| *Data normalization (Time 30 – 60 sec).* There are two normalization procedures - row-wise normalization and column-wise normalization. In the data normalization page, choose "**normalization by a reference sample**" then select the first sample name "0-1-1" for row-wise normalization. **Critical Step:** The choice for a reference sample is generally the sample in the control group with the fewest missing values. Alternatively, users can choose to use a pseudo-reference sample created by averaging all samples in the control group. For high quality data in which samples in the same groups are very homogenous, the effects of either procedure should be very similar.

5| Select "**auto-scaling**" for column-wise normalization.

6| After the normalization steps have been completed, click "**next**" to view a graphic summary of the normalization effects on the data (**Fig. A3**.).

7| (*Optional step*) *Compound name standardization (Time: 1-2 min).* This step is only applicable for compound concentration data. Click the "**Name check**" node under the "Processing" branch. The results of the name conversion process will be shown as a table. Compounds without an exact match in MetaboAnalyst's name library will be highlighted in either yellow (approximate match found) or red (no match found).

Users should manually examine the compounds with approximate matches and choose the correct one. Otherwise the first match in the candidate name list will be used. Click "**Submit**" button to finish the name checking. Note that after this step, all three major nodes on the navigation tree - "Statistics", "Enrichment" and "Pathway" should be enabled. Note if the data is uploaded under the "Enrichment Analysis" or "Pathway Analysis" tab, the compound name mapping will be performed by default. The data are now processed, normalized and ready for a variety of downstream analysis procedures.

Identification of Significant Features with Univariate Methods *(Time: ~10 minutes)*

8| *Identification of significantly different features (Time: 2-3 min).* MetaboAnalyst directly supports significant feature identification using several methods including t-tests, ANOVA, volcano plots, SAM, *etc*. As the example data contains four groups, we use ANOVA (Option A) and SAM based-method (Option B) to select important features.

a. ANOVA-based feature selection

i. Click the **ANOVA** node on the navigation tree to enter the "One-way ANOVA and post-hoc analysis" page.

ii. Significant features are identified with the default p-value threshold of 0.05. As the ANOVA F-test only tells that at least two of the groups differ, the post-hoc analysis further tests which ones differ from each other. MetaboAnalyst offers two commonly used methods - Fisher's least significant difference (LSD) and Tukey's honestly significant difference (HSD). Tukey's HSD is generally more

conservative than Fisher's LSD.

iii. Click the "**view details**" link to see a data table from the ANOVA and post-hoc tests using Fisher's LSD (the default). Users can click any compound name to view a box plots summary of its concentrations in different groups.

b. SAM-based feature selection

i. SAM is designed to control the false positives when running multiple tests on high-dimensional data. To use the SAM method, click the "**SAM**" node on the MetaboAnalyst navigation tree.

ii. The default view is the Step 1 tab which contains two plots to help users decide a suitable Delta value. The left plot shows the FDR change with different Delta values, and the right plot shows the number of significant compounds identified given different Delta values. For example, using the default Delta value 0.6 will identify ~25 compounds with an FDR ~ 0.3; using a Delta value of 1.0 will identify ~ 20 significant compounds with the FDR less than 0.1. Enter 1.0 as the new Delta value and click "**Submit**".

iii. The Step 2 tab shows a typical SAM plot with the Delta equals 1.0. Click the "**View details …**" link to see the SAM results table. A total of 21 compounds were identified above the chosen threshold. Notice the top ten compounds are almost exactly the same as those identified using ANOVA.

9| *Identification of other features with patterns of interest (Time: 2-3 min).* This step

allows users to investigate trends or patterns in metabolite concentration changes. Click the "**Correlations**" node on the navigation tree to enter the "Correlation Analysis" page. There are two types of correlation analysis that can be performed in MetaboAnalyst - correlation with a defined pattern (Option A) or correlation with a specific feature (Option B).

A. Correlation with a defined pattern of change

i. Here we will attempt to identify those metabolites that increase concentrations with the percentage of grain in the diet. Choose a pre-defined pattern "1-2-3-4" from the "**select a predefined pattern**" drop-down list, which corresponds to a linear concentration increase in groups 0, 15, 30, and 45, accordingly. Alternatively, users can specify their own patterns in the "**define your own pattern**" text field.

ii. Click the '**Submit'** button beside the drop-down list used in the previous step. The result is shown in **Fig. A4a**. The light blue shows those metabolites exhibiting a negative correlation and the light pink shows those with a positive correlation with the given pattern of change.

iii. Click the "**view details**" link to see a table of all the compounds listed as well as their correlation coefficients. Clicking any compound name will generate a graphic summary of its concentration distribution within each group (**Fig. A4b**).

B. Correlation with a specific feature

i. Based on the above analysis and a review of the literature, we know that elevated levels of *Endotoxin* are important for initiating certain inflammatory responses. We are interested in identifying other metabolites with patterns of change similar to *Endotoxin.* We will use the default "**Pearson r**" as the distance measure and then select "Endotoxin" from the "**Select a feature**" drop-down list.

ii. Click the "**Submit**" button. The resulting image shows a number of other features that are either positively or negatively correlated to *Endotoxin* levels. The details can be obtained by following the "**view details**" link.

10| *Report generation and result download (Time: 20 sec).* Click the "**Download**" node on the navigation tree. MetaboAnalyst will generate a detailed analysis report based on the steps that the user has previously executed. The report contains a brief description of each method used followed by the graphical and textual results based on the last parameter set. The normalized data, as well as any graphs generated during the analysis are also available for download.

Multivariate Data Analysis *(Time: ~10 minutes)*

11| *Data exploration and visualization with PCA (Time: 2-3 min).* PCA summarizes data into a few components that explains most of data variance. Click the "**PCA"** node on the navigation tree to enter the PCA page. This page shows six main output panels from MetaboAnalyst's PCA analysis. The default view is a pair-wise score plot from the top five PCs with the diagonal panels showing the explained variance.

12| Click the "2D score plot" tab to see a detailed scores plot using PC1 and PC2. The samples are labeled and colored according to their group memberships. In this view, users should look for: (a) Outliers - if there are obvious outliers, use the "**DataEditor**" under the "Processing" navigation tree to exclude outliers. Outlier removal should be done with considerable care and outliers should only be removed only if there is some clear justification (sample stability problems, sample collection issues, instrument problems, typographical errors, etc.); (b) Sample dispersion - if the data points in the score plot are not well dispersed or exhibit a high degree of skewing, this may be due to insufficient normalization. Click the "**Normalization**" node under the "Processing" branch to choose a different normalization procedure. In particular, Autoscaling or Range Scaling is very effective for correcting severely skewed data.

13| In our case, no obvious outliers or skewed distribution can be detected. Furthermore, some modest separation or clustering is noticed among different groups. There are also some clusters that appear to overlap with each other. Users can click the "**3D score plot**" to see if a better separation can be identified with an extra dimension.

14| *Identification of influential or important features (Time: 15-30 sec).* If good separation patterns are seen in a scores plot, users should go to the "**Loading plot**" as well as the "**Biplot**" views to identify those features that are most responsible for the separation. The loading plot can be viewed either as a scatter plot or a bar plot as specified by the user. In this particular case, since there are no clear separations, it is very difficult to identify which features are important. We will use a supervised method - PLS-DA for this purpose.

15| *Data exploration and visualization with PLS-DA (Time: 1- 2 min).* PLS-DA can

perform both classification and feature selection. Click the "**PLS-DA"** node on the navigation tree to start this analysis. The default view is pair-wise summary of the score plots of the top 5 components.

16| Click the "**2D Score plot**" for a detailed view of the separation patterns. A much better separation is obtained with PLS-DA compared to the PCA result obtained in Step 10. The 3D Score plot shows an almost perfect separation with the first three components (**Fig. A5a**).

17| *Choosing the optimal number of components (Time: 1-2 min).* MetaboAnalyst calculates R^2, Q^2 and prediction accuracies through cross-validation. Click the "**Cross Validation**" tab to start the process. Users can choose "10-fold cross validation" or "Leave-one-out cross validation (LOOCV)". In this case, we will choose "LOOCV" and click the "**Submit**" button. The result indicates that using the top two components gives the best performance based on Q^2 (**Fig. A5b**). Click the "**view details …**" link to get a detailed table of the calculated values.

? TROUBLESHOOTING (SEE TABLE A2)

18| *Result validation (Time: 2-3 min).* As noted earlier, PLS-DA tends to overfit the data and this can often lead to false separations or incorrect classification. As a result PLS-DA models need to be validated to see if the separation is statistically significant or due to random noise. This can be done using permutation tests. In each permutation, a PLS-DA model is built between the data (X) and the permuted class labels (Y) using the optimal number of components determined in the previous step. MetaboAnalyst provides two kinds of performance measures: (a) The separation distance which is defined as the ratio of the between-group sum of the squares and the within-group sum of squares (B/W-ratio) as suggested by Bijlsma *et al* (127). (b)

The prediction accuracy. This is the default approach used by MetaboAnalyst. Click the "**Permutation**" button to view the results. The resulting histogram summarizes the distribution of the permutation test scores with the red arrow indicating the performance based on the original labels. The further the arrow it is to the right of the distribution, the more significant the separation between the two groups. **Fig. A5c** shows a typical permutation result based on separation distance. As seen in this figure, the original class assignment is very significant and not part of the distribution we obtained using the permuted data. A p-value < 0.0005 is reported based on 2000 permutations.

? TROUBLESHOOTING (SEE TABLE A2)

19| *Identification of important features (Time: 1-2 min).* Click the "**Var. Importance**" tab to see a list of important features identified based on the VIP score (**Fig. A5d**). For multiple group analysis, the VIP score is calculated for each component. The overall VIP score shown in the figure is the average across all the selected components. Users can also use the coefficient - based importance measure. For multiple-group discriminant analysis, the same number of predictors will be built with one for each group. The overall coefficient-based importance is the average of feature coefficients in all predictors. Click the "**View details …"** link to see the individual VIP scores in each selected component, or the coefficients in each group predictor if coefficient-based importance is used.

? TROUBLESHOOTING (SEE TABLE A2)

20| *Report generation and result download (Time: 20-30 sec).* Click the "**Download**" node to download all the data, tables, figures produced from this particular analysis.

Metabolite Set Enrichment Analysis *(Time: 5-10 minutes)*

21| In the Upload page, click the "**Enrichment Analysis**" tab.

22| There are three drop-down panels for three different types of enrichment analysis. Each method accepts a different data type: (a) A list of compound names entered in a single-column format for over-representation analysis (ORA); (b) A list of compound concentrations entered as two-column table for single sample profiling (SSP); (c) A concentration table (.csv) with samples in rows and metabolites in columns for quantitative enrichment analysis (QEA). The phenotype information must be placed in the second column and can be binary, multi-class, or continuous. Click the third drop-down pane "A concentration table (quantitative enrichment analysis)".

23| In the open page, click "**Browse**" to locate the "human_cachexia.csv" data file.

24| Make sure the selected compound label type" is "compound names" and the phenotype label is "Discrete (Classification)", and then click "**Submit**".

? TROUBLESHOOTING (SEE TABLE A2)

25| *Compound name conversion (Time: 1- 2 min).* The purpose of this step is to compare and convert the compound names to common compound names used in the HMDB. The compound identities can be specified by common names or major database IDs (i.e. KEGG, PubChem, HMDB, MetLin, BiGG, *etc*). MetaboAnalyst's compound name/ID conversion is based on a name-mapping table from the HMDB. Each HMDB compound ID is associated with a common name, a set of synonyms, as well as compound IDs used in other major metabolomic databases. Any naming inconsistency is flagged and displayed to users for manual inspection and correction (see Step 7 for more details).

Critical Step: users must label their compounds with either common compound names or common database IDs. Abbreviated names usually cannot be recognized. Unmatched or unidentified compounds will be excluded from downstream analyses.

26| *(Optional) Concentration comparison (Time: 1-2 min).* This step is only applicable when the uploaded data is a list of compound concentrations used for SSP. The basic idea behind SSP is to compare the measured concentration values of each compound to its normal reference ranges in the corresponding biofluid. For common human biofluids such as blood, urine or CSF, normal concentration ranges are known for many metabolites. In clinical metabolomic studies it is often desirable to know whether certain metabolite concentrations in a given sample are higher or lower than their normal ranges. This procedure is designed to provide this kind of analysis. Click the "**Conc. check**" to start concentration comparison. By default, only compounds with concentrations above or below all the known or reported normal ranges will be selected for further investigation. Users should manually select or deselect compounds to override this default selection by inspecting the concentration comparison plots as well as the original reports by clicking the image icon in the "Details" column.

27| *(Optional) Data normalization (Time: 10-20 sec).* This step is only applicable when the uploaded data is a concentration table. In this case, we select "Normalization by a reference sample", and then choose "create a pooled average sample from the 'control' group. Choose "Autoscaling" for column-wise normalization.

28| *(Optional) Data visualization and outlier detection (Time: 1-2 min).* The purpose of this step is to check if the data values are relatively homogenous and for outlier detection. Click the "PCA" node to open the PCA page. On the 2D score plot, a clear outlier "PIF_115" is noticeable as it sits far away from all other data points. This

particular outlier is due to sample deterioration/contamination; Follow the route "Processing → DataEditor" and select "PIF_115" under the "Sample Editor" tab, click "Remove" and then click "Finish" to go back to the normalization page. Perform the data normalization as Step 27. Re-check the PCA score plot. This time, no obvious outlier should be detected. Follow the "Enrichment → Set param." and enter the page to specify parameters for enrichment analysis.

29| *Set parameters for enrichment analysis (Time: 30 sec – 1 min).* In this step, users must specify a metabolite set library (or upload a custom metabolite set library) to start the analysis. Users can also indicate whether a filter should be applied to exclude metabolite sets containing very few compounds. In this case, we use the default "Pathway-associated metabolite sets" and click the "**Next**" button.

30| *View the MSEA results (Time: 3-5 min).* The MSEA result is presented both graphically and in a detailed table (**Fig. A6a**). The horizontal bar graph summarizes the most significant metabolite sets identified during the analysis. The bars are colored based on their p-values and the bar length is based on the fold enrichment calculated as the *actual matched # / expected # of match* (for over-representation analysis) and *calculated statistic / expected statistic* (for quantitative enrichment analysis), respectively. The Bonferroni corrected p-value and FDR are also provided. Users can click the image icon in the "**Details**" column of each matched metabolite set to view all its constituent metabolites with matched ones highlighted in red (**Fig. A6b**), as well as SMPDB pathway images(147) (when available).

31| *Report generation and result download (Time: 10-20 sec).* Click the '**Download**' node to download the analysis report, images as well as the processed data.

Metabolic Pathway Analysis *(Time: ~ 10 minutes)*

32| *Data upload and processing (Time: 2- 3 min).* In the Upload page, click the "**Pathway Analysis**" tab to start upload and process the "human_cachexia.csv" data. The steps are similar to those involved in the enrichment analysis (see Steps 21-25 for more details). Note that users can also provide a list of compound names for pathway analysis.

? TROUBLESHOOTING (SEE TABLE A2)

33| *Set parameters for pathway analysis (Time: 30 sec-1 min).* Three parameters must be specified for pathway analysis including the pathway library, the algorithm for pathway enrichment analysis, and the algorithm for topology analysis. Users can also supply a reference metabolome to correct for any potential bias in the enrichment analysis. The reference metabolome is specified as a list of KEGG compound IDs. In this case, we select the "Homo sapiens" library and use the default "Global Test" and "Relative Betweenness Centrality" for pathway enrichment analysis and pathway topology analysis, respectively.

34| *Result visualization (Time: 3-5 min).* The results from the pathway analysis are presented in two parts - a graphical output on the top section and a table containing all the numerical results at the bottom. Users can intuitively explore the results by pointing and clicking on various graphic elements. There are three types of view (**Fig. A7**). The left panel is the "metabolome view" which displays all the matched pathways as circles (**Fig. A7a**). The color and size of each circle is based on their p-values and pathway impact values, respectively. Pointing the mouse over different nodes will show the corresponding pathway names. Clicking the node of interests will launch the corresponding "pathway view" on the right panel (**Fig. A7b**). Users

can zoom or drag to focus on a particular section the pathway. Clicking on any matched compound node (with highlighted background) will show the corresponding "compound view" which contains detailed summary of the compound concentrations, importance measure, as well as the p value (**Fig. A7c**).

35| *Report generation and result download (Time: 10-20 sec).* Click the "**Download**" node to get the complete analysis report as well as the processed data and images produced during the analysis.

Timing

The times required to perform the steps described in the protocol depend on the data set size as well as the number of active users connected to the web server. For the test datasets used for these protocols, most results should be returned in a few seconds after a user has selected the appropriate parameters. The most time consuming computational step is probably the permutation test used by PLS-DA (15-20 sec for 1000 permutations). The most time-consuming non-computational test is typically the data visualization or data inspection step. Data upload, processing and normalization (Steps 2-7) should take about 5-10 min; feature selection using univariate analysis (Steps 8-10) usually take around 3-5 min; multivariate analysis (Steps 11-20) take around 10 min, respectively. For high-level functional analysis, metabolite set enrichment analysis (Steps 21-31) should take 5-10 min, while metabolic pathway analysis (Steps 32-35) should occupy ~10 min, respectively. Once the data has been uploaded, an experienced user should be able to execute the full protocol in 30-40 min.

Anticipated Results

Graphical Output:

The graphical outputs produced during the analysis procedures are given in **Figures A1-7**. Some of MetaboAnalyst's algorithms use time-dependent random number generators to calculate certain statistical values and the results may vary slightly among runs.

Data Processing:

The **data integrity check** for the data in "cow_diet.csv" will detect four groups with a total of 51 zero values and no missing values. The **data integrity check** for "human_cachexia" will yield two groups with no zero or missing values.

Feature Selection Using Univariate Methods:

In MetaboAnalyst's **ANOVA** analysis of the "cow_diet.csv" data, the top five compounds identified with the default threshold should be: *Endotoxin, 3-PP, Glucose, Isobutyrate*, and *Methylamine*. The top five compounds identified using **SAM** method will be the same. In **correlation analysis** using the pre-defined "1-2-3-4" pattern, *Endotoxin* and *Alanine* are the top two compounds that will be positively correlated with this pattern, while *3-PP* and *Aspartate* are the top two compounds that will be negatively correlated with this pattern. The same compounds should be identified as being correlated/anti-correlated with "Endotoxin" using the "Pearson r".

Multivariate Data Analysis:

The score plot from the **PCA** analysis of the "cow_diet.csv" data should not show a clear separation, with group 1 and 2 overlapping significantly and group 3 slightly overlapping with group 2 and 4. A much better group separation will be achieved via **PLS-DA**. From PLS-DA, the five most important compounds identified by VIP will be: *Endotoxin, 3-PP, Alanine, Methylamine,* and *Glucose*. The best PLS-DA model will use just top two

components based on the Q^2 score estimated from LOOCV (0.814). The p-value based on 2000 permutations should yield a value of p < 5e-04, which is very significant.

Metabolite Set Enrichment Analysis:

All compound names from the "human_cachexia.csv" data set should be found to have an exact match during the name conversion step. The PCA score plot should not show a clear separation, although it should show PIF_115 as being a clear outlier. In the enrichment analysis using the pathway-based metabolite sets, the top five metabolic pathways that appear to be associated with cachexia will be – *Pyrimidine metabolism, Beta-alanine metabolism, Ketone body metabolism, Purine metabolism,* and *Glutamate metabolism*.

Metabolic Pathway Analysis:

The top five pathways from the "human_cachexia.csv" data set that should be identified by pathway enrichment analysis alone are: *Pyrimidine metabolism, Pantothenate and CoA biosynthesis, Beta-alanine metabolism, Synthesis and degradation of ketone bodies,* and *Propanoate metabolism*. Note that three of these pathways are similar to those previously identified by MSEA. The top three pathways identified by topology analysis alone should be: *Glycine, serine and threonine metabolism, Pyruvate metabolism*, and *Taurine and hypotaurine metabolism*. Overall, three pathways - *Pantothenate and CoA biosynthesis*, *Citrate cycle (TCA cycle)*, *Alanine, aspartate and glutamate metabolism* appear to be perturbed as a consequence of cachexia as these will be located in the diagonal area of the plot with relatively good scores from both analyses.

Tables

Table A1 Comparison of different metabolomic data analysis tools.

The level of support is rated by the number of '+', with '+++' as the highest.

Tool	MetaboAnalyst	MeltDB	metaP-Server	SIMCA-P	SAS
Software type	Web-based	Web-based	Web-based	Stand-alone	Stand-alone
License	Free	Free (registry required)	Free	Commercial	Commercial
Data input	Data table, NMR, MS, GC-MS data, compound/peak lists	Raw mass spectral files	Data table	Data table	Data table
Graphical interface	+++	++	++	+++	+/-
Normalization	+++	+	+	++	++
Univariate analysis	+++	++	+++		+++
Multivariate analysis	+++	++	+	+++	+++
Clustering	+++	++			++
Classification	++				++
Enrichment analysis	++				
Pathway analysis	+++		+		
Pathway visualization	++				
Integration with other omics data		+			
Peak annotation	++	+++			

Table A2 Troubleshooting guide.

Step	Problem	Possible reason	Possible solution
1	The content of the home page does not show up	JavaScript is disabled in your browser	For Mozilla Firefox 3.0+, go to the Tools → Options → Content, then select the checkbox beside "Enable JavaScript"; For Internet Explorer 8.0, go to the "Tools → Internet Options → Security, then select "Internet" from the Zone icons. Click the "Custom level …" button. From the list of available options, make sure the "Disable" radio button is not selected under "Active scripting" item; For Safari 4.0 +, go to the Edit → Preferences → Security, then select the checkbox beside "Enable JavaScript"; Please check the documentation for other browsers on how to enable JavaScript.
2, 24, 32	Fail to upload data	Non-unique or unusual names; small sample size; wrong data formats; unrecognized zip format.	Make sure sample or feature (peak/ compound) names are unique and consist of a combination of English letters, underscores, or numbers for naming purposes; The names should contain no space or other special characters; Make sure at least three samples per group; Make sure the selected data format matches your data; For Microsoft Excel users, choose "CSV (Macintosh)" to generate .csv file; For WinZip (v12.0) users, choose the "Legacy compression (Zip 2.0 compatible)" for compression.
17-19	No image is generated	The sample size is too small	These procedures require a minimum of five samples per group
10, 20, 31, 35	No PDF report is generated	Some expected data are not generated	Set appropriate parameter values to make sure the resulting images are generated; Make sure a minimum of five samples per group for PLS-DA analysis.

Figures

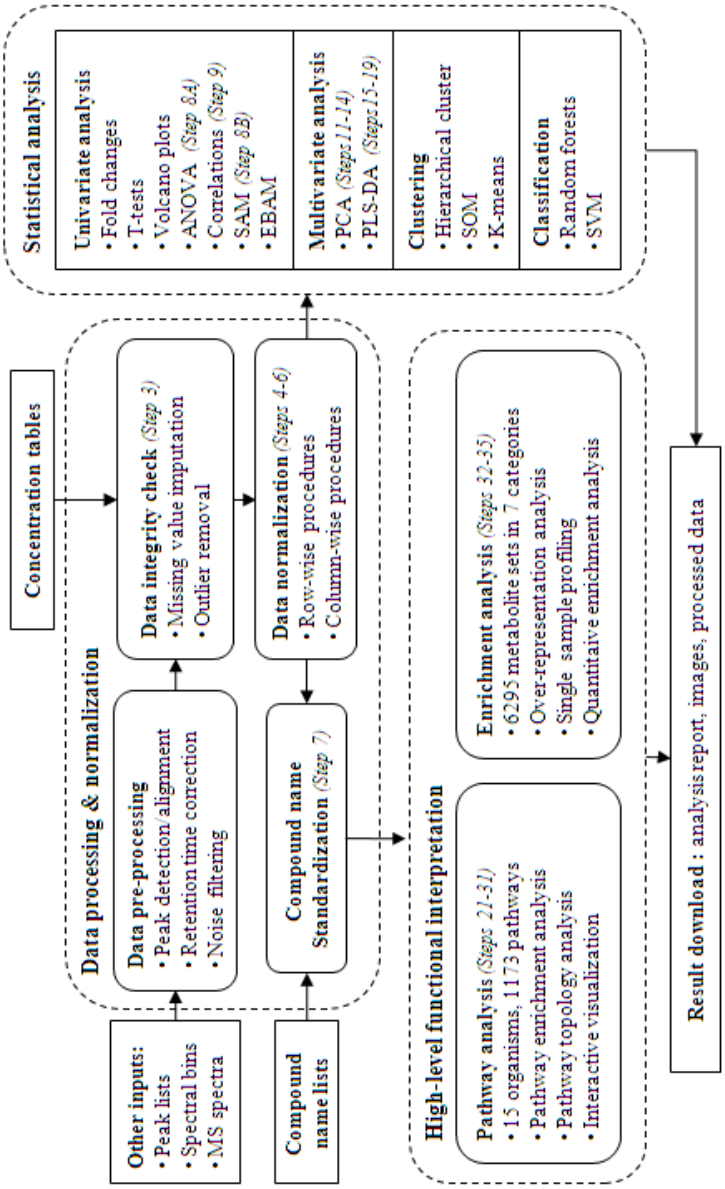

Figure A1. MetaboAnalyst's flowchart.

The procedures that have been described in the protocol are indicated by the corresponding steps.

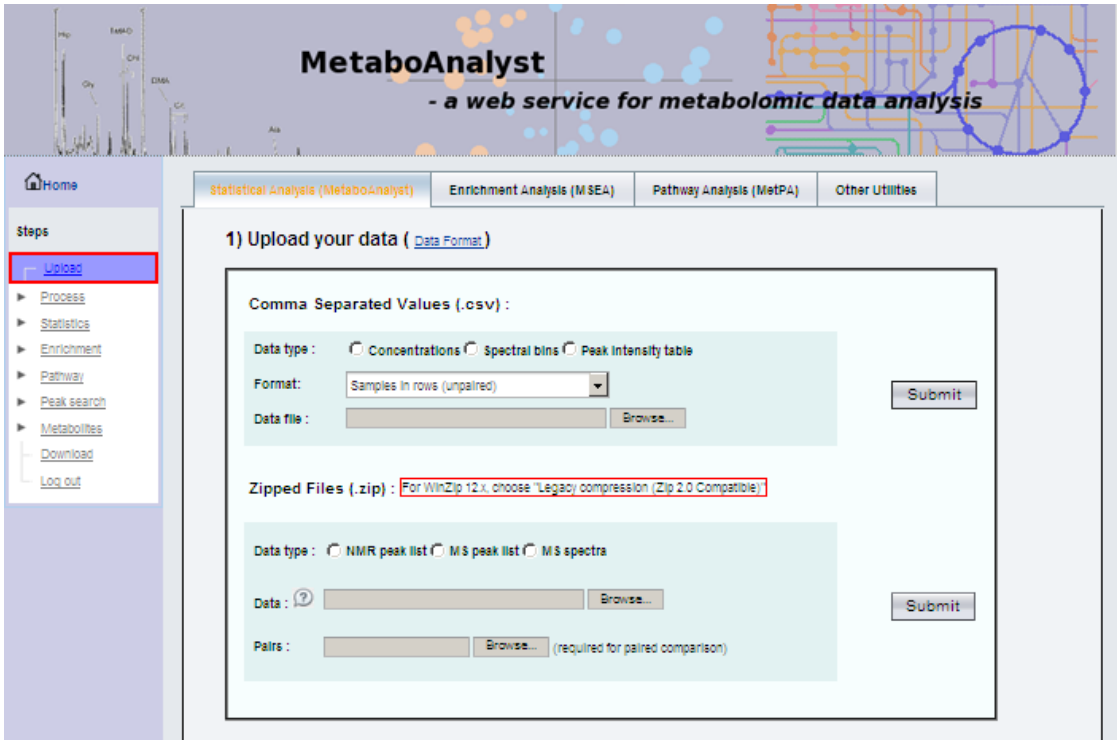

Figure A2. Data upload view

This screenshot shows MetaboAnalyst's available data analysis modules with the "Statistical Analysis" module being selected for data upload. Clicking the tab labeled "Enrichment Analysis" or "Pathway Analysis" will allow users to upload data for the corresponding data analysis. The navigation tree is located on the left panel with the current step "Upload" highlighted.

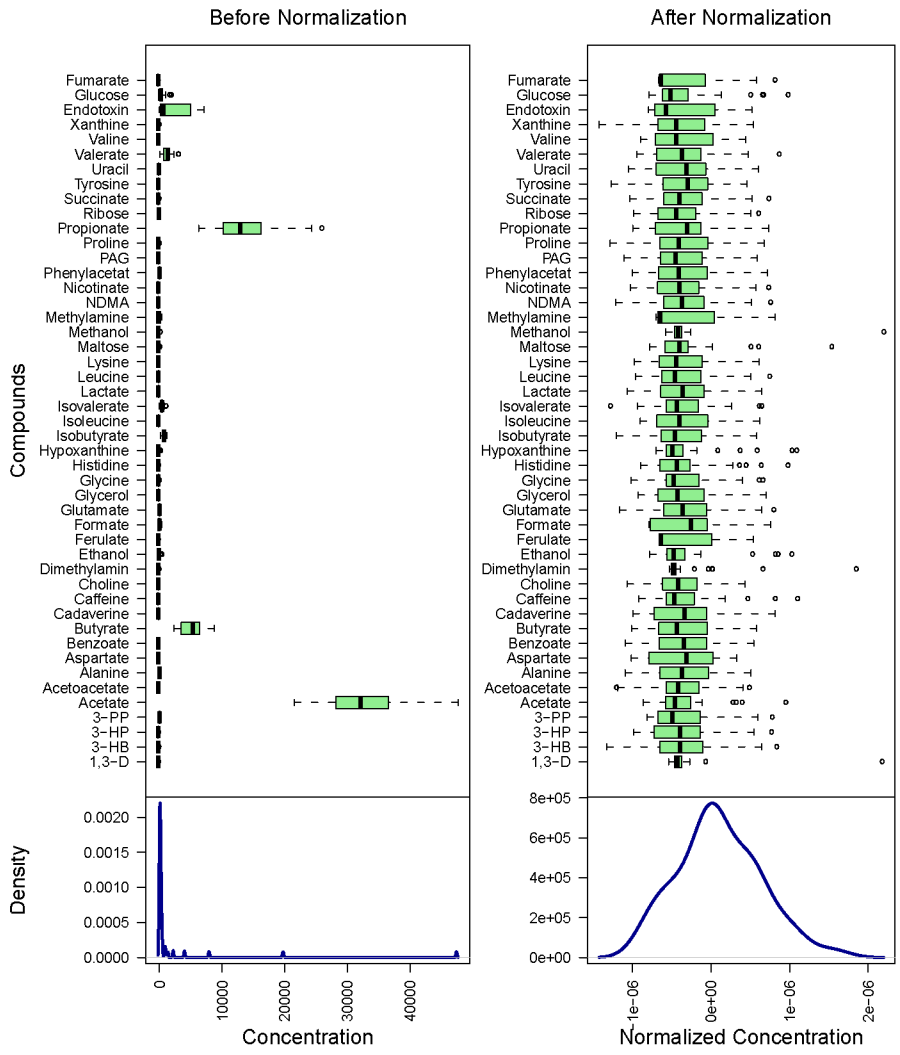

Figure A3. Data normalization view

The graph summarizes the distribution of input data values before and after normalization. The box plots on the top show the concentration distributions of individual compounds, while the bottom plots show the overall concentration distribution based on kernel density estimation.

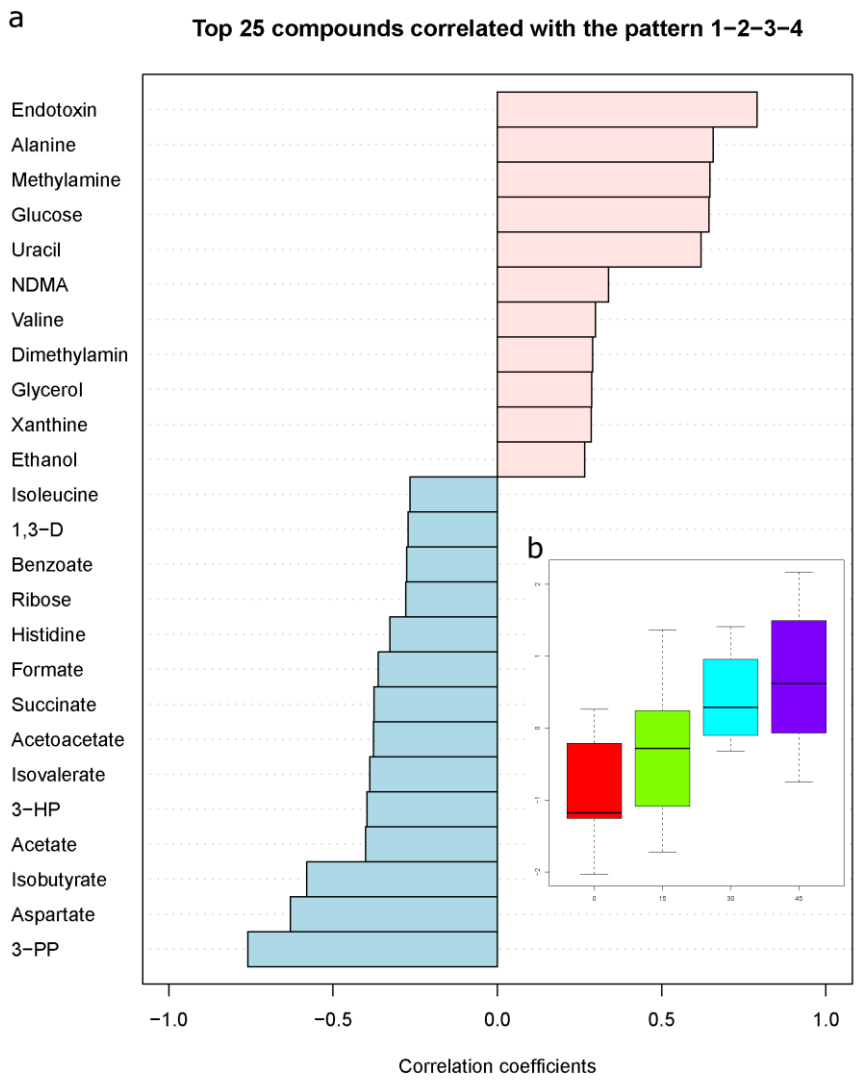

Figure A4. Multivariate analysis using PLS-DA

(**a**) Correlation plot showing the compounds that are significantly associated with a given patterns "1-2-3-4" (a linear concentration increase under different conditions). The compounds are represented as horizontal bars with colors in light pink indicating positive correlations and light blue for negative correlations. Users can click the "view details" link to see a detailed table. (**b**) Box plots summarizing the concentration distributions of a selected compound.

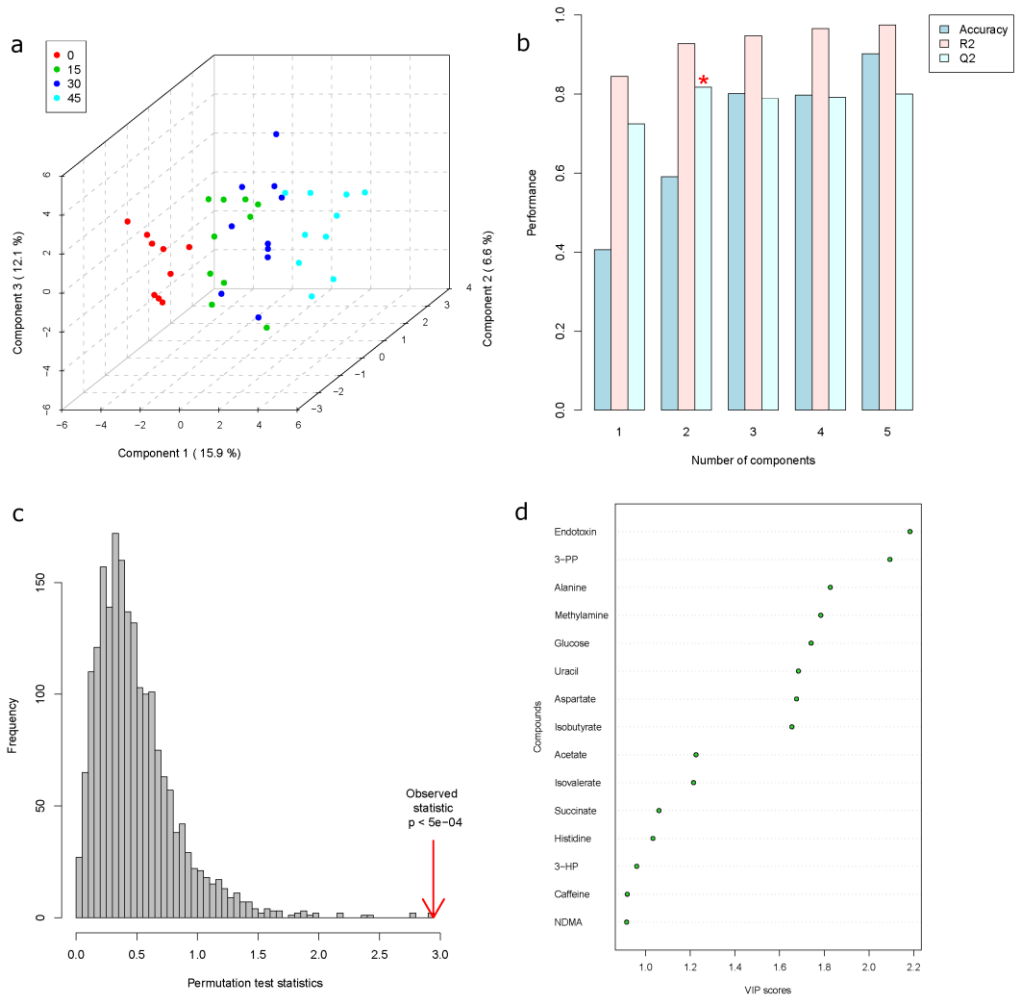

Figure A5. Correlation analysis to identify features with specific patterns

(**a**) PLS-DA 3D score plot. (**b**) Bar plots showing the three performance measures using different number of components. The red "*" indicates the best values of the currently selected measures (Q^2). (**c**) The result of permutation tests summarized by a histogram. (**d**) The top 15 compounds ranked by VIP scores.

a

Metabolite Set	Total	Hit	Statistic	Expected	P Value	Holm P	FDR	Details
PYRIMIDINE METABOLISM	36	3	5.9815	1.3158	0.0026037	0.11977	0.065958	
BETA-ALANINE METABOLISM	13	2	7.5707	1.3158	0.0028677	0.12905	0.065958	
KETONE BODY METABOLISM	10	1	9.089	1.3158	0.0077101	0.33925	0.10832	
PURINE METABOLISM	45	2	5.9459	1.3158	0.0094194	0.40503	0.10832	
GLUTAMATE METABOLISM	18	3	4.2634	1.3158	0.019852	0.8338	0.18264	
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	36	3	3.6972	1.3158	0.060012	1.0	0.4159	
INOSITOL METABOLISM	19	1	4.5238	1.3158	0.063289	1.0	0.4159	
MITOCHONDRIAL ELECTRON TRANSPORT CHAIN	15	2	3.3111	1.3158	0.082012	1.0	0.42415	
PROTEIN BIOSYNTHESIS	19	10	2.2262	1.3158	0.093859	1.0	0.42415	
BETAINE METABOLISM	10	2	2.9882	1.3158				
CITRIC ACID CYCLE	23	6	2.4262	1.3158				
PANTOTHENATE AND COA BIOSYNTHESIS	10	1	3.1919	1.3158				
GALACTOSE METABOLISM	25	3	2.4909	1.3158				
PROPANOATE METABOLISM	18	1	2.9087	1.3158				
TYROSINE METABOLISM	38	3	2.3462	1.3158				
PHOSPHOLIPID BIOSYNTHESIS	19	1	2.7762	1.3158				
UREA CYCLE	20	4	2.1039	1.3158				
METHIONINE METABOLISM	24	4	1.8657	1.3158				

b

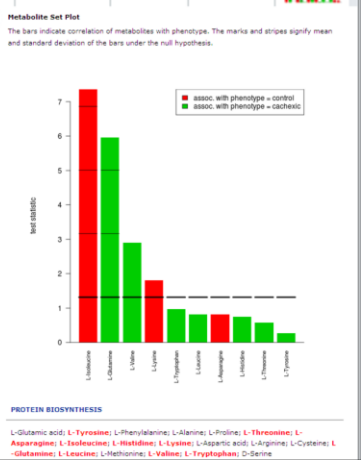

Figure A6. Results from metabolite set enrichment analysis

(**a**) The result table summarizing the matched metabolite sets ranked by their p-values. (**b**) The detailed view of a matched metabolite set (accessed by clicking the corresponding bar icon on the last table column).

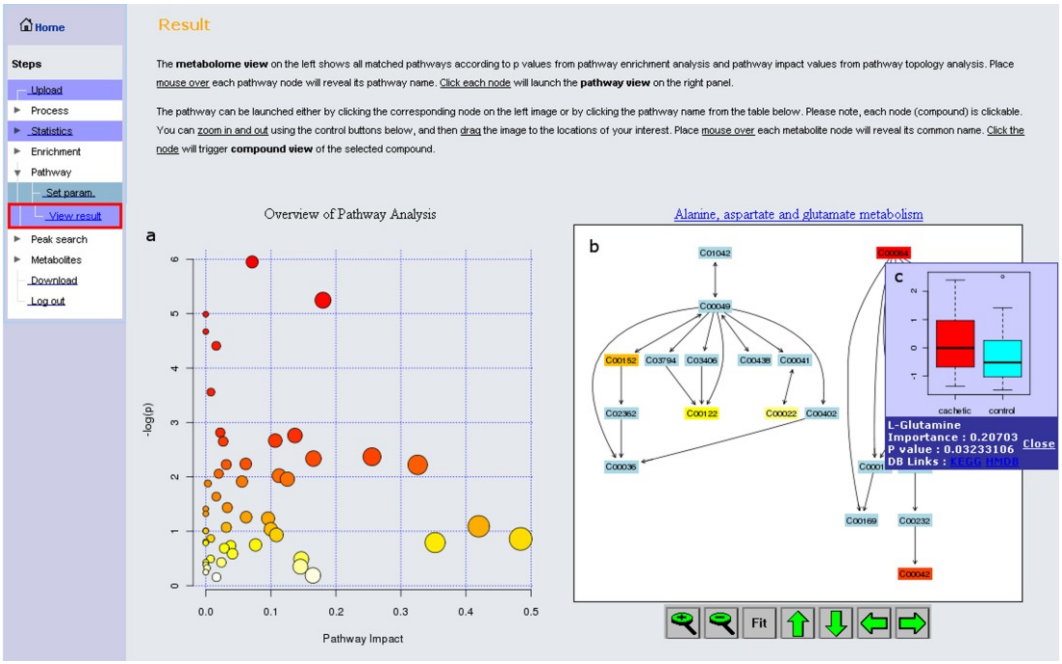

Figure A7. Metabolic pathway analysis and visualization

(a) The "metabolome view" showing all metabolic pathways arranged according to the scores from enrichment analysis (y-axis) and from topology analysis (x-axis). (b) The "pathway view" showing the corresponding metabolic pathway after clicking any node in the "metabolome view". The matched metabolites are highlighted according to their p-values. Users can zoom or drag the pathway map to view a subset of the compounds. (c) The "compound view" showing the concentration distribution of the corresponding metabolite after clicking any matched compound node. The p-value and the node importance are indicated below.