High-Throughput Analysis Tools for Mass Spectrometry Based Data Quantitation

Thesis Submitted to AcSIR For the Award of
the Degree of
DOCTOR OF PHILOSOPHY
In Chemical Sciences

By
Avinash D Ghanate
10CC13J26021

Under the guidance of
Dr. Venkateswarlu Panchagnula

CSIR-National Chemical Laboratory
Pune (India)

July 2018

I gratefully dedicate this thesis to my parents

for their endless love, support and encouragement

राष्ट्रीय रासायनिक प्रयोगशाला
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
डॉ. होमी भाभा मार्ग पुणे - 411 008. भारत
NATIONAL CHEMICAL LABORATORY
(Council of Scientific & Industrial Research)
Dr. Homi Bhabha Road, Pune - 411 008. India.

Thesis Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled **"High-Throughput Analysis Tools for Mass Spectrometry Based Data Quantitation"** submitted by **Mr. Avinash Ghanate** to Academy of Scientific and Innovative Research (AcSIR) in fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, embodies original research work under my guidance. I further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material obtained from other sources has been duly acknowledged in the thesis. Any text, illustration, table etc., used in the thesis from other sources, have been duly cited and acknowledged.

Ghanate

Avinash D Ghanate

(Research Student)

P. Venkateswarlu

Dr. Venkateswarlu Panchagnula

(Research Guide)

Declaration of authorship

I hereby declare that the original research work incorporated in this thesis entitled **“High-Throughput Analysis Tools for Mass Spectrometry Based Data Quantitation”** submitted to Academy of Scientific and Innovative Research (AcSIR) in fulfillment of the requirements for the award of the degree of Doctor of Philosophy (Ph.D.), is the outcome of original research and experimental investigations carried out by me under the supervision of **Dr. Venkateswarlu Panchagnula**, Sr. Scientist, Chemical Engineering Division, CSIR-National Chemical Laboratory, Pune. I affirm that the work is original and has not been submitted in part or full by me for award of any other degree or diploma in any other university.

Ghanate

Date: July 19, 2018 **Avinash D Ghanate**

Place: CSIR-National Chemical Laboratory

Pune (India) **(Research Student)**

Acknowledgements

The work in this dissertation was carried out in Chemical Engineering division (CEPD), CSIR-National Chemical Laboratory (NCL), Pune, India. Research funding from CSIR India and the Department of Biotechnology (Ramanlingaswamy fellowship) is gratefully acknowledged. I also owe a debt of gratitude to Director of NCL and CEPD head of division, for allowing me to complete my research work in the premises and avail facilities offered by NCL.

I am grateful to my supervisor, Dr. Venkateswarlu Panchagnula, who not only introduced me to this fascinating research field of mass spectrometry but always believed in my skills to adapt and evolve continuously, throughput this journey. His office door was always open for all of us research students whenever we ran into any trouble spot or had a question. But beyond being a supervisor, I believe he truly helped me shape my personality both professionally and at personal level.

I would also like to thank my Doctoral Advisory Committee, Drs. Anu Raghunathan and Chetan Gadgil for their continuous support and critical remarks that helped me address any shortcomings from different perspectives. I am specially indebted to Dr. Anu Raghunathan for providing me with collaboration and learning opportunities that helped me align my dissertation work and translate the analytical outcomes into system level biological interpretations that constitutes a chapter in this thesis. I am also grateful to Ms. Rupa (from Dr. Raghunathan's lab) for offering cell culture sample extracts and my lab colleagues, Ijaz & Dr. Nivedita for providing chromatography-free mass spectral analyses, that helped evaluate and support various tools developed as part of this thesis.

My past research experience, prior to dissertation work, with Dr Gadgil have actually instigated me in maintaining multi-disciplinary viewpoint while working with biological systems. And I have always admired his methodological approach in delineating research

problems that helped the most from each DAC meetings. I am fortunate to have Dr Mahesh Kulkarni, who also comes from analytical background, as a chair person for all my DAC meetings. His critical evaluation and positive attitude towards my dissertation have kept up my enthusiasm levels, in each passing year.

PhD is a long journey and every journey can be made more endearing and adventure filled depending on travel companion that one gathers along. And I am very happy to have shared this journey with many good friends and colleagues who actively contributed at different capacities as mentor, well-wisher, offering motivation or inspiration and also emotional support when one needed those most. I would like to start first with two of my great friends and colleagues, Dharmesh and Vishal. I have shared most of my PhD time with both of them and to summarize their contribution in few sentences would be infeasible. Although few of their peculiar qualities like, Dharmesh's looking through things with a clear and practical perspective while Vishal's dealing with any problems with most warm and light hearted fashion, have left me with a sense of admiration and inspiration towards them. I will never forget the late night scientific discussions over tea, which detoured the course of a project in a very interesting manner.

I would like to express my deepest gratitude for all past doctoral colleagues of my and collaborator's research group, Drs. Ajeet, Deepika, and Deepanwita along with Nirav, who helped me equally and offered their support throughput. I cannot thank enough few research colleagues from Dr. Gadgil's group with whom I worked in past- Nidhi, Shraddha and Sucheta. Their support and motivations offered all along these years is invaluable.

I reserve my special thanks for Anil, Bharat & Mayoor, from NCL along with Samartha & Pranay, from Guitar clases, who always kept me connected with the non-scientific world around, by means of discussions over many tea sessions and recreational activities like hiking and overnight plans. I would like to thank all my friends and fellow PhD students from NCL- Abhishek, Anurag, Dimpal, Madhura, Nilesh, Priyabhrata, Priyanka, Sachin, Saikat, Satej, Sayantan, Shantanu, Sutanu, Sourik, Vishwanath, and Yashwant. I am also grateful to many research trainees and project students - Vijay, Hussain, Kalyan, Pooja, Preshita, Yusuf, Saket, Onkar, Rutuja, Niveadita, Cheshta, Shanthini, Sangeetha and Kaushik. Varying amount of time shared on scientific and non-scientific aspects with each of them left me with memories to cherish. I would like to thank all administrative staff from NCL, especially from C&P and SAO section and Mr. Raheja from CEPD section who helped me with all diligence, for promptly completing any official formalities required, throughout my PhD.

I would also like to take this special opportunity and thank my mathematics teacher from school class 7^{th} standard, Mrs. Bagade. Her encouragement and passionate teaching allowed me to keep open mind for all these scientific endeavors so far, since that young age.

I would not have been able to complete my PhD without the unending love and support from my family. My family - parents (Deepa & Dattatraya), siblings (Rahul & Vinayak), and sister-in-laws (Veena & Sushma) who continuously believed in me and encouraged throughout my PhD tenure. And all the love received from newest addition to my family, my nephew, Atharva and niece, Ridhvi have made this closure most endearing.

Again, I express my deepest gratitude to each one of you and my sincere apologies to those whom I have not mentioned here by mistake.

– Avinash Ghanate

"I can see there's a connection between not following normal thinking and doing creative thinking. I wouldn't have had good scientific ideas if I had thought more normally."

– John Forbes Nash Jr.

Foreword

A multitude of developments in mass spectrometry (MS) methodology are at the helm of post genomic scientific pursuits such as proteomics, metabolomics and lipidomics. Significant advancements in MS instrumentation have also helped in enabling sensitive and accurate measurement of molecules from complex biological systems. It is now possible to comprehensively delineate metabolite profiles resulting from cellular processes and trace dynamics using high-resolution mass spectrometry (HRMS). Most often gas and liquid chromatography (GC / LC) coupled with HRMS are the methods of choice for qualitative metabolomics profiling. LC-triple quadrupole (LC QqQ MS with unit resolution) is subsequently used for quantitative metabolite analysis by selectively monitoring ion reactions. High resolution accurate mass (HR-AM) based workflows for HRMS data have of late shown significant promise in enabling simultaneous Qual/Quant metabolomics analysis utilizing the full spectrum of the HRMS data. However, using HR-AM quantitation with other non-chromatography based direct MS analysis has largely been unexplored in comparison. There is a significant need for algorithmic development, especially for non-chromatography based direct MS analysis, to fully realize the potential of HRMS metabolomics data in various applications. This is also vital for potential scalable translation of analyses to the 'market' as well as the 'clinic'.

Metabolic profiles generated using HRMS also hold potential in deducing functional insights of cellular dynamics. Although these metabolic profiles are essentially collective downstream effect of contribution from various catabolic and anabolic metabolic reactions, progressive analysis methods such as constraints based metabolic network analysis enable deciphering the mechanistic interplay. Furthermore, HRMS offers the advantage of analyzing targeted metabolic profiling data in retrospective manner extending its application domain to support any hypothesis resulting from the metabolic modeling pursuits. Metabolic modeling with genome scale models also enables spanning multiple molecular hierarchies in the cell through the classical central dogma for molecular biology (CD). Synergetic nature of analysis workflows of metabolic measurements and modeling offer a systems-level perspective to discover emergent properties that can be difficult to accomplish independently.

Existing LC-MS approaches require elaborate sample pre-processing and lengthy chromatographic time. A major resulting drawback is the extremely low throughput that has so far hindered the widespread adaptation of quantitative MS. Although direct MS methods such as laser desorption ionization HRMS offer increased throughput, their wider use is severely hindered by the lack of validated quantitative data processing tools. Currently available software tools either come bundled with MS instruments with support for only specific file formats / selective analytes or are geared only towards specific workflows (for example proteomics, metabolomics, or lipidomics profiling). Also, data handling in the majority of data processing tools for quantitative analysis is pegged with chromatography-based workflows leaving little room for those exploring direct HRMS measurements devoid of chromatography. The first part of this dissertation describes the development of an algorithmic tool that comprehensively supports high throughput data processing as well as targeted metabolite quantitation following direct HRMS workflows. In the second part of dissertation comprehensive, accurate and reliable extraction of differential metabolic features from LC HRMS data in an untargeted manner was showcased. Additional efforts to integrate HRMS metabolic profiles for system-level metabolic network analysis, to discover new biological insights and generate hypothesis, have also been explored and constitute the latter part of the dissertation.

Abbreviations

AP atmospheric pressure

API atmospheric ionization

CBM Constraints based modeling

CD classical central dogma for molecular biology

CE-MS capillary electrophoresis

CI chemical ionization

COBRA Constraints Based Reconstruction and Analysis

COW correlation optimized warping

CV cross validation

CWT continuous wavelet transform

DART direct analysis in real time

DESI desorption electrospray ionization

DI direct infusion

DNA deoxyribonucleic acid

DWT discrete wavelet transform

EI electron ionization

EMG exponentially modified gaussian

ESI electrospray ionization

FBA flux balance analysis

FT-ICR Fourier transform - ion cyclotron resonance

FT-IR Fourier transform infrared spectroscopy

FT-MS Fourier transform mass spectrometry

FVA flux variability analysis

FWHM full-width at half maximum

GC gas chromatography

GC-MS gas chromatography coupled to mass spectrometer

GPR gene protein reaction relation

GSM genome scale metabolic

HR-AM High resolution accurate mass

HRMS high-resolution mass spectrometry

HR TOF-MS high resolution time-of-flight mass spectrometry

IS internal standard

LC liquid chromatography

LC-MS liquid chromatography coupled to mass spectrometer

LP linear programming

LW loading or factor weights

MALDI matrix-assisted laser desorption ionization

MALDI MS matrix-assisted laser desorption ionization mass spectrometry

MEW mass extraction window

MFA metabolic flux analysis

MOMA Minimization of Metabolic Adjustment

MS mass spectrometry

MSI mass spectral imaging

MS/MS tandem mass spectrometry

m/z mass-to-charge ratio

NMR nuclear magnetic resonance

PCA principal component analysis

PLS partial least square

PLS-DA partial least square discriminant analysis

ppm parts per million

QC quality control

Q-TOF quadrupole time of flight

RC regression coefficient

RNA ribonucleic acid

RSD relative standard deviation

RT retention time

SAM S-adenosylmethionine

SAH S-adenosine-L-homocysteine

S/N signal to noise ratio

TIC total ion chromatogram

TOF time of flight mass analyzer

TOF/TOF time of flight mass analyzer in reflector mode

VIP variable importance on projection

XIC extracted ion chromatogram

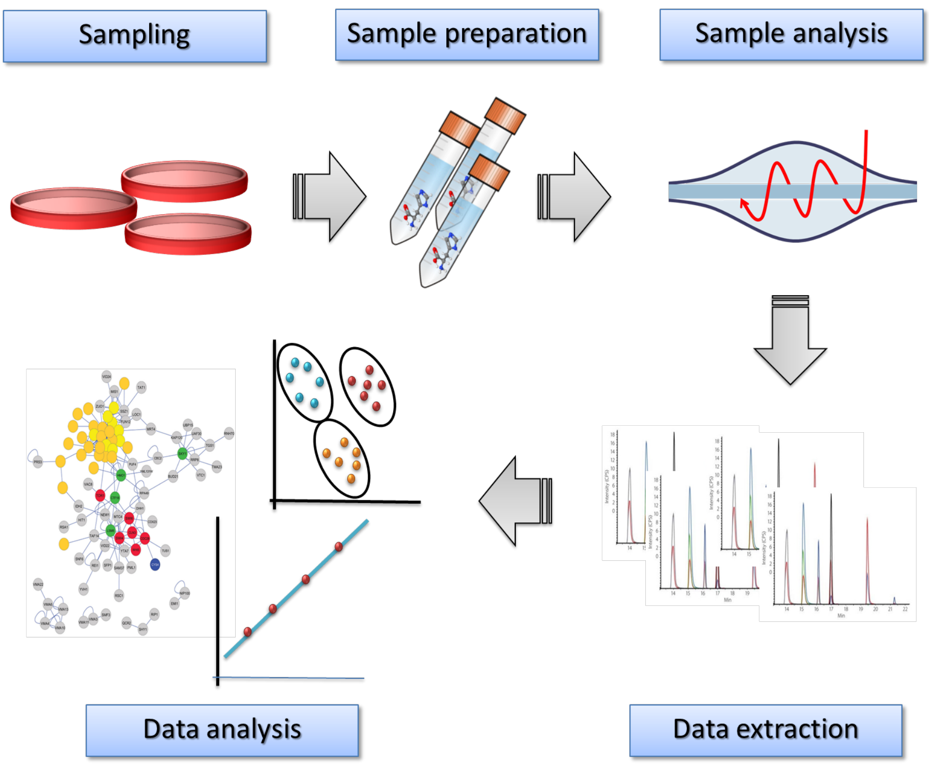
Contents

List of Figures

List of Tables

Chapter 1

Introduction



Schematic of metabolomic analysis using HRMS

1.1 Metabolomics, available analytical methods and challenges for data analysis

Metabolites are small chemical molecules that characterize the functional state of cellular biochemical activity. They are involved in various metabolic reactions that are essential for growth, maintenance and normal functioning of cell. Application for metabolites as biomarkers of disease diagnosis has a long history (1500-2000 BC).[1] Rapid technological development in early 20th century provided modern researchers with tools to investigate the hierarchy of metabolic involvement in biochemical reaction network and also to elucidate their role in diseased states.[2] The information flow from DNA to protein through ribonucleic acid (RNA), presented as classical central dogma for molecular biology (CD), is more rigorously controlled through epigenetic regulations and post-translation modifications. In contrast, dynamics of biochemical reactions and their end products are most predictive of cellular phenotype.[3,4] This also forced to conceptually change CD by introducing metabolites while respecting the information relationship from DNA to metabolic profiles, which is also termed as 'omics cascade'.[5,6](Figure 1.1)
Increasing interests in metabolic profiling gave rise to various terminologies for 'omics' technologies, such as, metabolome - A complete set of metabolites in an organism[7], metabonomics - quantitation of metabolic responses to pathophysiological stimuli or genetic modifications in a cell[8], metabolic foot-printing - extracellular metabolite profiling to investigate their secretion or uptake by cell, and so on. In general metabolomics is the generic acronym used, which encompasses all these studies and defines comprehensive characterization of metabolites in a biological system.

Global profiling of metabolites is challenging owing to their varied chemical classes, sampling complexities, diverse abundancies and dynamic biotransformations. Dramatic development in analytical technologies over the past few decades however began enabling researchers to measure individual biomolecules from biofluids involved in biochemical reactions. Analytical platforms such as nuclear magnetic resonance (NMR)[8], Fourier transform infrared spectroscopy (FT-IR)[9,10] and mass spectrometry (MS)[11–13] have been deployed for metabolic profiling with varying degrees of success. Of these, MS-based technologies, often with gas or liquid chromatography (GC or LC) as a 'front-end', have seen a significant increase in adaptation due to their versatile, sensitive, and selective detection along with analytical robustness.[14] Mass spectrometry offers significant advantages over other analytical platforms and enables efficient qualitative and quantitative metabolomics analysis.

Figure 1.2 shows a generic workflow of MS based metabolic investigation.[14] There are several important considerations in metabolomics investigations in addition to the choice of a specific MS platform or analyzer. There are various published reports that discuss elaborate details of sample harvesting, handling, storage, processing, derivatization and extractions that are vital prior to MS analysis.[3,14,15] Chromatographic resolution of metabolites based on their physio-chemical characteristics following introduction into an analytical separation column is an important aspect in most MS-based analytical protocols (other than direct MS workflows). Metabolites generally elute at different time points based on their retention time (RT) that depend on differential partitioning

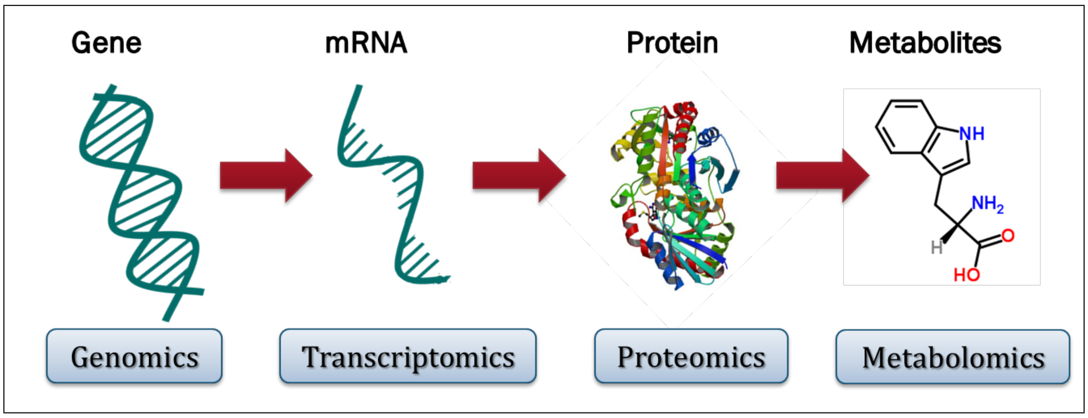


Figure 1.1 Omic cascade and improved central dogma of biology.

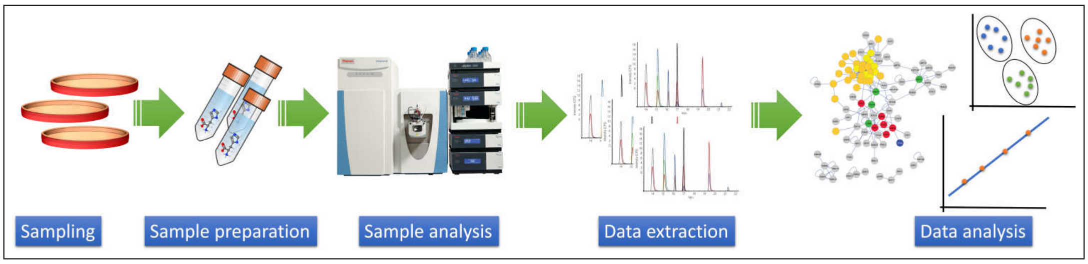


Figure 1.2 Experimental workflow for MS based metabolic analysis.

between the stationary and mobile phases. Chromatographic method development that includes selection of optimal column and separation chemistries usually precedes actual analysis. The methods are generally validated using reference standards and isotopically labelled metabolite analogs, especially in targeted metabolomics approaches, prior to adaptation with metabolites extracted from biological samples. Subsequently metabolites are ionized into the gas phase with the help of online injection as small aliquots and resolved along their respective mass-to-charge ratio (m/z) in the mass analyzer. At this stage, evaluating sample matrix effects to attain optimal ionization efficiencies and minimization of ion suppression / enhancement is crucial. At end of this elaborate workflow, mass spectral profiles are acquired and strategies to qualify and quantify analytes are adapted. However, most biological samples could potentially contain a large pool of metabolites, intermediates and / or degradation or biotransformation product peaks that could challenge the capabilities of a lower resolution mass analyzer.

In a typical global metabolomics study, constituting both quantitative and qualitative workflow, LC-MS is most widely used mass spectrometry technology on account of its ability to separate and detect broad range of compounds. Other hyphenated technologies such as GC-MS[13], CE-MS[16] have also contributed significantly to the field of metabolomics with continued method advancements. Technologies that use alternative ionization to ESI such as, MALDI[17], DART[18], DI and DESI[19] in tandem with newer generation HRMS have evolved and are increasingly finding their applicability to unique metabolite analysis contexts - mass spectral imaging being a case in point. From a global metabolic profiling perspective, these technologies suffer from limitations such as ion interferences from isobaric and closely matched metabolites, signal interference from matrix sources and in some cases insufficient reproducibility.

Development of high-resolution mass spectrometry (HRMS) with magnetic sector or Fourier transform - ion cyclotron resonance (FT-ICR) mass analyzers helped address resolution challenges of complex samples including those from biological sources. With mass resolving power of >50000 FWHM, it is now feasible to resolve biomolecules at mass spectral level with highly improved mass accuracy where isobaric interferences are generally at a minimum.[20] However applications of HRMS until recently, were limited to qualitative workflows, on account of its lower scan rate that affected scan resolution of chromatographic runs. Prohibitive costs of HRMS instruments were yet another factor. Modern HRMS instruments (with TOF and Orbitrap mass analyzers and relatively 'affordable' benchtop models) with their fast scan rates allows acquisition of enough data points across a chromatographic peak and use extracted ion chromatogram (XIC) for quantitation.[21–23] A growing list of articles and review publications illustrate such analysis workflows made feasible with HRMS. Applications of HRMS in Qual / Quant approach have been demonstrated in food safety analysis[24], organic contaminant analysis in environmental sciences[25], clinical biomarker studies[26] and metabolic fingerprinting[22,23] to name a few. It is now theoretically possible to separate and resolve several hundreds of metabolites from each individual sample aliquot that can be reliably measured and quantitatively determined. But as a consequence, data complexity resulting from both chromatographic separation and HRMS based spectral profile generation increase. Coupling HRMS platforms with direct MS approaches such as direct analysis in real time (DART) or matrix-assisted laser desorption ionization (MALDI) interfaces offers high-throughput and enables newer frontiers of analysis, but yet require support with data analysis.[18,27] A representative example is of mass spectral imaging (MSI), using direct

MS platform, that offers opportunity to investigate molecular interactions from intact sample surfaces, which would be challenging to achieve using traditional chromatographic approaches.

An ever increasing list of application portfolio that utilizes Qual/Quan HRMS toolbox, either direct MS or in conjunction with chromatography, brings about significant 'data' analysis and interpretation challenges. Development and adaptation of appropriate data handling and analysis methods is indeed crucial across all metabolomics workflows. Lack of streamlined data analysis tools to support such diverse analytical approaches including direct MS, hinders the pace of development using HRMS based strategies. The work described in this dissertation is an attempt to address some of these challenges through the development of complimentary tools that offer cross platform data analysis to fully exploit the Qual/Quan capabilities of HRMS data. The first part of the dissertation describes the development and benchmarking of 'MQ' - a modular high throughput platform tool for comprehensive qualitative and quantitative analysis using direct HRMS analysis. MQ has successfully been used for several metabolomics analysis contexts in the recent past and a fully evolved (and thoroughly tested). The academic version of MQ can be accessed at www.ldi-ms.com.

The breadth of analyte coverage achievable from individual HRMS scans also make it amenable for implementing advanced meta-analysis, to gain invaluable fundamental insights of cellular mechanisms. There is a growing list of published literature reporting global metabolic profiling that offer cellular functional insights using HRMS based workflows.[6,15,28–32] Such functional interpretations can be commonly achieved using multivariate analysis tools.[15,33–35] Interrogation using metabolic network modeling that can benefit not only from metabolic coverage but quantitative accuracies of HRMS, offers the next paradigm for a systems-level understanding and integration across multi-omics data. The second part of this dissertation showcases the use of HRMS based metabolic profiles

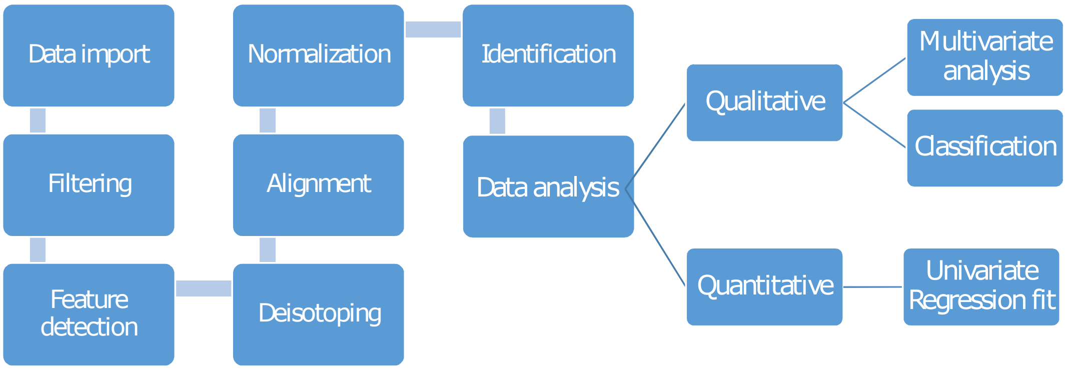


Figure 1.3 Data analysis workflow for processing of mass spectrometry data.

to formalize metabolic reconstruction models of cancer cell lines. Consistency of the *in silico* predictions with cellular metabolic phenotypes along with statistical interpretations made using multivariate tools, helped evaluate the model's performance.

The remainder of this introductory chapter provides an overview of relevant data processing and analysis methods that form the background for the work described in this dissertation. Various open source data analysis tools exist that employ different strategies for data treatment and are indeed useful for HRMS based metabolomics analysis. Applicability of such alternative algorithms is also discussed while keeping different analytical platforms in context. Available routes for data interpretations following data processing discussed subsequently.

1.2 Overview of data processing methods

Various data pre-processing methods are generally useful for processing HRMS data prior to qualitative interpretation and quantitative analysis. Specific steps for the data pre-processing essentially involve filtering, feature detection, deisotoping, alignment and normalization followed by data interpretation (Figure 1.3). Filtering methods help remove data interference coming from noise or baseline signal. Feature detection extracts the signal counts for each probable peak from data. In virtue of differing isotopic compositions

Table 1.1 List of open-source tools for data processing of MS data for metabolomics

Tool name	Platform support	Features	Availability
apLCMS[45]	LC-MS	Feature extraction, peak alignment, retention time correction, recover weak signal	http://web1.sph.emory.edu/apLCMS/
COMSPARI[46]	LC-MS, GC-MS	Visual comparison	http://sourceforge.net/projects/comspari/
MAVEN[38]	LC-MS	Alignment, Peak detection, Isoptope/Adduct calculator, Pathway visualizer	http://genomics-pubs.princeton.edu/mzroll/index.php
MetaboAnalyst[39]	LC-MS, GC-MS	Analysis like statistical, functional enrichment, pathway, ROC curve based biomarker	http://www.metaboanalyst.ca
MetAlign[47]	LC-MS, GC-MS	Filtering, peak detection and Alignment	http://www.wageningenur.nl/nl/Expertises-Dienstverlening/Onderzoeksinstituten/rikilt/show/MetAlign.htm
MET-COFEA and MET-XALIGN[48]	LC-MS	Feature extraction, alignment and annotation	http://bioinfo.noble.org/manuscript-support/met-xalign/index.php
MET-IDEA[49]	LC-MS, GC-MS, CEMS	Feature extraction of selected metabolites	http://www.noble.org/plantbio/sumner/met-idea.html
MetSign[50]	LC-MS, DI-MS	Alignment, Metabolite annotation, normalization, statistical tests, pattern recognition	http://metaopen.sourceforge.net/metsign.html
MSFACTs[51]	LC-MS, GC-MS	alignment, comparison across samples	http://www.noble.org/plantbio/sumner/msfacts/index.html
MSight[52]	LC-MS	Visual comparison	http://web.expasy.org/MSight/
msInspect[53]	LC-MS	Peak detection, alignment, normalization	http://proteomics.fhcrc.org/CPAS
MultiAlign[54]	LC-MS and LC-MS/MS	Alignment and database matching	http://omics.pnl.gov/software/multialign
MZedDB[44]	LC-MS	Metabolite annotation based on ionisation behavior and biological source	http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html
MZMine/MZMINE2[42]	LC-MS, GC-MS	Filtering, peak detection, alignment, normalization	http://mzmine.sourceforge.net/
OpenMS[37]	LC-MS	Normalization, alignment, feature extraction and database search	http://open-ms.sourceforge.net
SIRIUS[43]	Any HRMS	Metabolite annotation based on isotope pattern	http://bio.informatik.uni-jena.de/sirius/
SuperHirn[55]	LC-MS	Peak detection, alignment, normalization, sample clustering and classification	http://proteomics.ethz.ch/muellelu/web/SuperHirn.php
XCMS/XCMS2[40]	LC-MS, GC-MS	Filtering, peak detection, alignment	http://masspec.scripps.edu/xcms/xcms.php

of compounds, cluster of peaks are observed in spectra. Deisotoping methods group such cluster of peaks with the specific peak for monoisotopic counterpart. Alignment involves identification of peaks specific to metabolites across various samples processed. Normalization deals with signal count correction for systematic variation, in order for their comparison across other samples from the study. A detailed overview of these data processing steps implemented in various tools is described below.

An appropriate workflow for data pre-processing should account for different sources of variations arising from experimental or technical inputs and should be adaptive on parameters, which depends on, (i) analytical method for chromatographic separation, (ii) analytical platform used, (iii) experimental design, (iv) subsequent data interpretation methods. Several tools and software packages that offer most of the data pre-processing cascade have been made available by various research groups in recent past. A list of open-source software tools designed to address different steps from data processing and analysis workflow, is provided in Table 1.1. Some tools are designed for handling proteomic data exclusively although a few of these could potentially be applicable to metabolomics analysis as well, as they share various pre-processing steps.[36,37] Several of the tools are specially designed for metabolomics data processing and interpretations.[38,39]. Whereas, there are other tools, which can be applied to any sort of LC-MS data.[40–42] Besides these, there are few tools that specifically cater to a specific step of analysis showcased in Figure 1.3.[43,44]

Available tools for data processing include both commercial and open-source alternatives. Though both types of these software tools follows similar strategy for data processing, as illustrated in Figure 1.3, the open-source tools provides transparency about the algorithm used, flexibility for developers to supplement new methods along with support for orthogonal list of data sources from different vendors.

1.2.1 Data conversion in platform independent format

With a range of available MS platforms, a number of proprietary data formats exist. For processing of data with freeware/open-source alternatives, data will need to be converted into platform independent formats, such as netCDF, mzXML, mzML *etc.* Most of the vendor specific software tools provide inbuilt support for conversion of proprietary formats into a compact binary data format, netCDF, widely used for storing experimental data. (http://www.unidata.ucar.edu/software/netcdf/) In addition, freely available interface libraries for netCDF are provided for common programming languages such as C, Fortran, Java, Python, IDL, MATLAB, R, C++, Ruby, and Perl. This facilitates further development of tools providing analysis solution in a platform independent manner.

An attempt to standardize a common data format, mzXML, for all types of MS data was made by Seattle Proteome Center (SPC) in Trans Proteomic Pipeline (TPP) software solution.[56] Besides the proteomics related set of analysis tools, an array of data converters were made available by the project for transforming different MS instrument specific formats into mzXML. A more recent project, ProteoWizard[57], provides a unified solution for data conversion from most of the vendor specific formats (for AB Sciex, Agilent, Bruker, Shimadzu, Thermo-Scientific and Waters) to mzXML and mzML. This conversion tool is also integrated in TPP software package.

Most of the open-source data processing tools support aforementioned MS data formats. Adapting and implementing such common data formats offers an opportunity to perform cross-platform comparison of any study and help identify better suited platform required for a complex study such as metabolomics.

1.2.2 Data filtering

In any data analysis workflow, the downstream resultant conclusions on processed data critically depend on the quality of input data. Like most of the instrumental signals, chromatographic MS data is constituted by signal, noise and background (Figure 1.4). Various strategies are followed for reducing the effect of these non-informative features that facilitates peak detection, in reducing false positive predictions and would help in improving quantitative information of data.[58]

While performing a batch acquisition of samples, baseline drift caused by chemical noise, is commonly observed distortion in MS analysis. Collective sources of such signals are clusters of matrix molecules from samples, interferences from solvents or buffers or impurities build up on the separation column.[59,60] A simple two-step process is typically followed for baseline removal: (i) fitting baseline profile, (ii) subtraction of this fitted response from raw signal. Different approaches for baseline estimation includes, regression fit through the peak bases of smoothed spectral segments,[61] non-linear filter like top-hat operator with small window size,[62] Savitzky-Golay filter with lower order polynomial[63] or iterative asymmetric least-square regression.[64] Estimated baseline also serves as the threshold level for calculation of noise from MS data.

In addition to the chemical noise, the MS data is also affected by random noise that arises mainly from detector. Like any electrical instrument there are various sources of noise in data apart from the contributions coming from physico-chemical nature of ions. The influence of these different sources vary with types of MS detectors used for study warranting careful evaluation of method used for de-noising. Traditional signal processing methods such as moving average window[65], moving median filter[66,67] and polynomial fitting following Savitzky-Golay method[63] transforms spectral signal by reducing noise interference. A mathematical representation of filtering process, commonly referred as

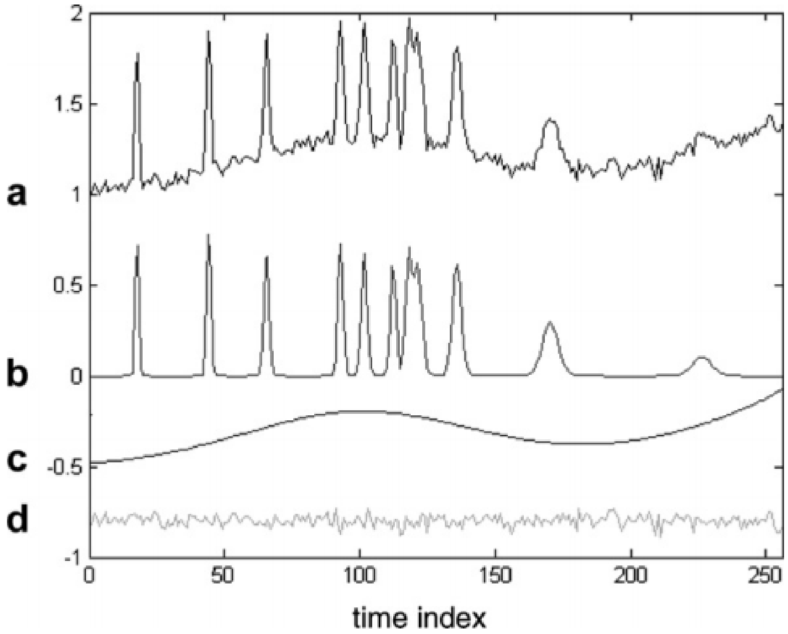
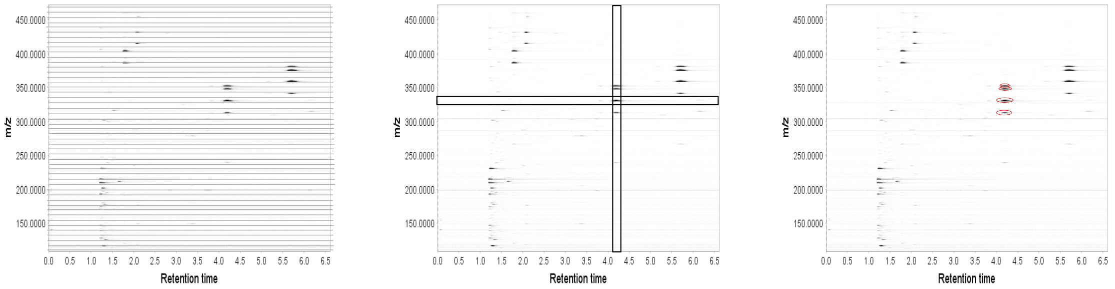


Figure 1.4 Different components of typical analytical signal. (a) observed signal response; (b) expected signal for analyte elution; (c) background signal profile; and (d) noise. (Image source: Ref 58)

'convolution', is shown in the following equation:

$$s(t) = \sum_{w=-\frac{L}{2}}^{\frac{L}{2}} F(t)y(t+w) \tag{1.1}$$

where s denotes processed signal value for data position t, F is the filter function with processing window length of L and y represents the raw data vector. The primary drawback of following such transformations is distortion of signal response value for the peak, which affects quantitative performance of the data. In order to circumvent this, other approaches that deal with noise by selectively filtering noise component were made available. Such methods either differentiate significant peaks based on user parameters like, peak width, slope threshold and use remaining data point for noise estimation[67] or follow an inverse approach to estimate noise signal levels first following methods such as, average/median response of low abundant signal values from spectra,[68] median absolute



(a) XIC scan of m/z bins

(b) Peak detection from both dimensions

(c) Fitting isotope model

Figure 1.5 Different strategies of feature extraction for separation based HRMS data. (a) Chromatograms generated from 2D MS data is represented. Horizontal lines schematically showcases bins created along m/z direction, (b) Features identified by considering peak profiles along both the time and m/z dimensions, (c) An isotopic model is fit to the data along m/z direction following RT values within threshold. (Image adapted from review article by Katajamaa *et.al.* that discusses recent literature related to LC-MS data handling[72])

deviation within specified window,[40] and linear regression model fitted to signal counts of noise peaks.[69] Owing to the heteroscedastic nature, *i.e.* unequal variance across the range of values, of the noise observed in MS data more sophisticated algorithms such as, wavelet transformation[70] and variance-stabilizing normalization methods[35] have shown to handle noise characteristics more efficiently.

Open-source tools MetAlign[47], MZmine[42] and OpenMS[37] provide filtering options as a processing step for correcting noise interference such as Moving mean, Savitzky-Golay, Binomial and Gaussian low-pass filter. Among other tools listed in Table 1.1, MAVEN[38], XCMS[40] integrate feature detection and filtering in a single step. For baseline correction MZmine provides various method implementations available under Bioconductor packages[71] through interface with R platform for statistical analysis.

1.2.3 Feature detection

Data structure for chromatographic and direct ionization MS

Feature detection is a crucial step in data processing for any data analysis workflow. It involves identification of signals associated with true metabolite ions, avoid false positive detection in the presence of noise or non specific peaks and provide quantifiable information for respective metabolites. Different workflows for metabolomics involving direct ionization or chromatography separation based MS analysis lead to differential treatment of raw data for feature extraction. Raw data for direct ionization MS analysis represents mass spectral profile with ion signal response of constituent metabolites. Feature extraction from such data, extracts lists of *m/z* for metabolite ions and their signal responses. Whereas in case of LC-MS or GC-MS, a series of mass spectra acquired at successive RT are stored in raw data. Considering the added dimension of separation, the output from feature extraction data processing consists of an additional vector of data information with RT value for each pair of *m/z* and signal response of metabolite ions. Different strategies for extraction of features are illustrated in Figure 1.5.[72]

XIC scan of *m/z* bins A simpler strategy of slicing *m/z* dimension as shown in Figure 1.5a yields list of ion chromatogram towards chromatographic direction for each extracted *m/z* bins. Such XICs can be further processed independently using different peak annotation methods. This approach has successfully been employed for various studies successfully and is featured by a few open-source software tools.[40,65] Selection of optimum bin width parameter for *m/z* slicing becomes a challenging task and also a limitation in applying this strategy. A broader bin width results in merging of co-eluting peaks that have *m/z* value within a half of the bin width parameter, whereas narrower bin width for a lower resolution MS data results in multiple feature being selected for same metabolite ion in consecutive bins. With current generation HRMS instruments, the optimal bin width for efficient feature extraction needs to be compressed to a very narrow value leading to an explosion in number of bins studied as opposed to the total metabolites practically present per spectral profile.

Peak detection from both dimensions Another straightforward approach for feature detection in LC-MS or GC-MS based data follows peak finding in both directions - m/z followed by chromatographic domain (Figure 1.5b). Although this methodology is computationally intensive, depending upon the algorithm chosen for peak finding in both directions, various available open-source tools exercise this approach because of its simplicity in implementation and efficiency in terms of coverage in feature extraction.[38–40,42,45]

Fitting isotope model In an alternative approach, isotopic model fit of individual features detected was used as an added level of scrutinization for feature extraction (Figure 1.5c). Use of generic mass-dependent isotope pattern (with consecutive peak list separated by ∼1 Da) may assist in reducing the false positive feature detection and hence improve quality of data.[73] Application of such a concept has shown to provide improved quality of feature extraction assisting in peptide sequence annotation[74] along with complex lipidome characterization and its quantitative analysis.[75] Although open source tools such as, MapQuant[74] and Isoconv[73] use such isotopic clusters for improved feature extraction, usage is limited to heavier molecules with m/z above 1000 Da.[73]

Approaches for peak detection

Numerous algorithms for extraction of peak profiles from MS data have surfaced in last few decades. A recent review[76] details various available peak detection algorithms, their pitfalls and mathematical underpinnings that can assist in identifying suitable robust method for feature extraction from high-throughput studies such as metabolomics.

Owing to Gaussian nature of peaks in both the dimensions (*m/z* and time domain), peak extraction algorithm following a common or combination of various strategies can be employed in each direction. Commonly used approaches for peak detection are based upon (i) ion intensity threshold; (ii) template (usually Gaussian) function correlation and (iii) wavelet transform techniques.

Use of ion intensity threshold The ion intensity threshold level is a simple approach, also termed as S/N based method, which uses ion response cut-off value for peak identification. This cut-off or noise value is either user defined or identified statistically using methods mentioned before.[40,62,68,69] Based on defined cut-off value, peaks are identified by scanning for local maximum within specific window along any dimension that qualifies the cut-off threshold. This approach has advantages of simplicity in implementation, faster performance and minimal user inputs. The major drawback of this method is its disregard for peak quality/shape. Also, sensitivity to noise estimation method may lead to vulnerability with erroneous peak prediction.

Correlation with template function Considering the Gaussian response of peaks in either dimension, scanning for a Gaussian function, *i.e.* matching filter function, with characteristic peak width equivalent to a predefined resolution would provide accurate peak selection.[40,42] Based upon a correlation threshold value with the template function, peaks can be characterized. Unlike intensity threshold based method, use of template function limits the false positive results for peak identification. However, the presence of high frequency noise can influence the correlation value and affect peak identification. This can be controlled using smoothing filter with larger window size before feature detection. Similarly, peak shape abnormalities caused by improper optimization of chromatographic condition or aberrations from ionization process, further contributes to false negative results in peak detection. Strategies that employ convolution of multiple filter functions, such as Gaussian-Exponential function[67,77] - termed as exponentially modified gaussian (EMG) distribution and Gaussian-Gaussian function,[78] were showcased as alternate template functions. Although feasible, the need for selecting such data treatment or use of alternative methodology, which is dependent as per case basis, makes this task further tedious.

Wavelet transformation Sophisticated techniques such as wavelet transformation were used as a solution that indirectly adapts the fitting function as per the data. For MS data analysis both continuous wavelet transform (CWT)[40,79] and discrete wavelet transform (DWT)[67] have been explored successfully. In brief, a mother wavelet function is translated at different m/z or RT locations and simultaneously scans for a range of wavelet scaling (frequency) values, generating daughter wavelets with varying peak width. An array of wavelet coefficient values were identified from this, providing index of matching score for daughter wavelets with data peaks. As an example, following wavelet transformation on chromatographic data results into time-frequency representation of the spectrum, in case of CWT, or dyadically discretized wavelet frequency representation, in case of DWT. Such representations provide characteristic information for spectral peaks, such as peak width and area based on scale and coefficient values respectively, along with peak positions. Flexibility to choose different mother wavelet function in addition to its scaling for varying peak width asserts this method with higher efficiency and robustness for feature extraction. Other advantages offered by wavelet method include insensitivity towards noise in data, with application of scale threshold for high-frequency noise, and also towards baseline drift from data.[79] A comparative performance evaluation of this method for mass spectral data demonstrated improved peak detection using CWT in comparison to publicly available peak detection algorithms.[79,80] With the complexity of algorithm, this method demands higher computational power and tuning large number of method parameters, which is difficult for a person having little expertise or understanding

Table 1.2 Molecular adducts observed in positive and negative ion mode MS sources operating at atmospheric pressure (such as ESI, MALDI, DART, DESI *etc.*)

Molecular adduct **Positive mode**	**Exact Mass shift**^{a}	**Molecular adduct** **Negative mode**	**Exact Mass shift**^{a}
$[M+H]^+$	1.00728	$[M-H]^-$	-1.00728
$[M+Li]^+$	7.01546	$[M-H+H2O]^-$	17.00329
$[M+NH4]^+$	18.03383	$[M+F]^-$	18.99895
$[M+H+H2O]^+$	19.01784	$[M-H+CH3OH]^-$	31.01894
$[M+Na]^+$	22.98922	$[M+Cl]^-$	34.96940
$[M+H+CH3OH]^+$	33.03349	$[M+HCOO]^-$	44.99820
$[M+K]^+$	38.96316	$[M+NO2]^-$	45.99345
$[M+H+CH3CN]^+$	42.03383	$[M+CH3COO]^-$	59.01385
$[M+H+H2O+CH3OH]^+$	51.04406	$[M+NO3]^-$	61.98837
$[M+Na+CH3CN]^+$	64.01577	$[M+Br]^-$	78.91888
$[M+Ag]^+$	106.90455	$[M+HSO4]^-$	96.96011
$[2M+H]^+$	-	$[M+H2PO4]^-$	96.96962
$[2M+Na]^+$	-	$[M+CF3COO]^-$	112.98559
$[2M+K]^+$	-	$[M+I]^-$	126.90503
$[3M+Na]^+$	-	$[2M-H]^-$	-
$[3M+K]^+$	-		

[a] Exact mass shift value defines shift in m/z peak from exact mass of metabolite.

Adaptation from citation ref 81.

about the technique. Very few open source tools provide wavelet based feature extraction method following both DWT[67] as well as CWT[40,42,79] strategies.

1.2.4 Deconvolution of data

Owing to the high sensitivity and ability to capture high resolution data with modern HRMS instruments, the resultant data not only contains metabolic information alone but is also accompanied with additional spectral peaks that can be a characteristic of the analytical workflow. Typically in a soft ionization source, such as electrospray ionization (ESI) that is commonly used with LC-MS studies, an array of ions may be generated on account of (i) different adducts formation, (ii) fragmentation and (iii) isotopic clusters for individual metabolite compound. A list of possible adduct ions observed in positive and negative ion mode MS acquisition using atmospheric ionization (API) sources such as ESI, DART, MALDI and DESI *etc.* have been listed in Table 1.2. An extended list of in-source fragmentation behavior observed in relation with various functional groups of metabolites and the expected mass shift from metabolite adduct ion peak is illustrated in ref 81. Typically for isotopic clusters, a rigorous workflow is followed, which involves iterative identification of elemental composition for each selected monoisotopic peak corresponding to strict mass accuracy constrains and simulation of isotopic patterns for identified elemental composition.[43] But these approaches become unfavorable for higher masses, given the exponential increase in number of chemical formulas constituted using different combinations of elements from CHONSP (roughly 7×10^9 molecular formulas possible for mass $\leq$ 1500 Da[43]). Similarly, simulation of isotopic pattern for higher mass becomes computationally expensive for routine algorithms, such as binomial expansion,[43] or even modern adaptations of polynomial expansion and Fourier transformation based method.[82] Such issues were handled in past for peptide related compounds, with either the use of statistical occurrence of amino acids from various database entries for peptide sequences - definition of averagine molecule[83] or least square fitting of relative intensity for isotopic peaks as function of monoisotopic masses of compounds.[73] Such approaches work better within a limited range of masses (∼1000 Da - 2000 Da). With increasing mass, the uncertainty for estimation increases to larger extent.[73]

While performing metabolic annotation based upon complete features for all ions, such redundant number of peaks for ions on account of different adducts, fragment ions or isotopic clusters originating from a single metabolite molecule may contribute to false positive predictions. Dispersion of quantitative features per metabolite into various ions leads to attenuation of sensitivity offered. By deconvoluting the data, such issues can be circumvented.

Common strategy for deconvolution is based on a simple concept that all ions originating from same metabolite should have similar RT value i.e. they elute out simultaneously in a chromatographic separation run.[72] All such ions with RT values within specified tolerance limit are grouped together. Though on account of various analytical reasons, such as inefficient separation method for metabolites, complexity of biological sample, network of metabolites participating sharing a common pathway, structural isomers *etc.*, different metabolites can share similar RT values or observed in proximal region within given tolerance limit. Using statistical similarity measures like Pearson's correlation coefficient for scan-by-scan signal intensity variation across XICs for grouped ions, such issues can be circumvented. Open source tool MAVEN[38] provides this approach that assigns a metric for identified isotopic peaks using cut-off criterion over correlation statistics to handle unrelated peaks grouping. Decon2LS[84] is another software tool, which works with spectral data and deconvolutes redundant features by following an algorithm, THRASH[85]. The algorithm follows a linear interpolation model based upon averagine molecule to estimate isotopic distribution. Subsequently, isotopic clusters are deleted to retain only monoisotopic peaks for each metabolite. Since, elemental composition for averagine molecule has been defined specifically with peptides in consideration, the applicability of this tool for metabolomics data becomes limiting. MapQuant[74] provides an alternative approach for deconvolving isotopic peaks following a tree data structure. Peak clusters were fitted with a binomially distributed sum of 2-D Gaussian functions as shown in eqn 1.2.

$$f(m, r; A, r_0, m_0, \sigma_m, \sigma_r, c, z) = A \sum_i \frac{B(i; c, p)}{2\pi\sigma_m\sigma_r} e^{-\frac{(r-r_0)^2}{2\sigma_r^2}} e^{-\frac{\left(m-(m_0+\frac{i}{z})\right)^2}{2\sigma_m^2}} \tag{1.2}$$

Here, the parameters that define this bivariate function of m/z (m) and retention time (r) are: total abundance of isotopic cluster (A), RT centroid (r_0) and m/z (m) for monoisotopic peak, standard deviation for Gaussian peak along RT dimension (σ_r) and m/z dimension (σ_m), number of carbon in molecule (c), charge state (z) and finally, binomial distribution function

$$B(i; c, p) = \binom{c}{i} p^{(c-i)}(1-p)^i$$

with natural isotopic abundance (p) for carbon-13 (^{13}C). Such a method efficiently helps to deconvolve isotopic clusters even if there are peak groups with intertwined isotopic clusters.

1.2.5 Peak alignment and retention time correction

In a routine untargeted analysis, also used for targeted approach, set of detected peaks for metabolites are grouped together based upon their characteristic response in different class of samples. Parameters like RT for chromatographic elution of metabolites and matching of exact mass for metabolites of interest are used for such grouping. Shifts in the RT dimension along with m/z direction must be accounted for before grouping such features towards metabolite annotations. In a chromatographic experiment, RT shifts can arise due to (i) column performance or column aging, (ii) temperature or pressure variations and (iii) sample matrix effect with varying range of concentration and salt contents. Though, drifts along m/z dimension are generally smaller and can be controlled with use of internal calibration during spectral acquisition. Given the non-linear nature of RT shifts,[40] alignment becomes a necessary task before further downstream analysis for chromatographic analysis. Typically, two approaches are followed for correction in RT shifts,

1. Use of raw data for generating RT mapping using a reference standard (Chromatogram alignment),

2. Metabolite features clustering after peak detection, with bi-dimensional proximity criteria along RT and *m/z* dimension (Peak alignment).

A commonly followed technique involves addition of internal standard mixture with all the samples. Detected features for these can be used to linearly shift metabolic features in the neighbourhood. But such a technique poses two further difficulties such as, use of linear response correction for non-linear behavior and possibility of competitive ion suppression effects with the presence of additional internal standards mix. Therefore, many of the open source data analysis tools follow practice of identifying a reference template, which may be evolved iteratively for improving alignment.[38,40,42]

A simpler approach is to carryout pairwise mapping of total ion chromatogram (TIC), which is known as correlation optimized warping (COW).[86] Since each feature from TIC may represent many metabolites, using only uni-dimensional TIC based correction may lead to alignment of non-specific features. Hence, similar methods like continuous profile model (CPM), which divides *m/z* dimension into four segments, may still not work efficiently.[87]

A second approach of alignment follows clustering of detected peak features. Since, feature detection helps reduce the data size, this method is computationally less demanding. This technique has been applied in several fashions such as (i) pairwise matching,[88] (ii) matching across set of replicates,[89] (iii) successively matched across all runs against first run,[42,47] and (iv) matched against an adaptive reference template with iterative cycle.[38,40,42] The reference set of features used were adapted following different criteria, which are specific to software packages, like XCMS identify a group of well behaving peaks that are observed in almost all samples and preferably without any other conflicting features in neighborhood,[40] whereas in case of MZmine's Random Sample Consensus (RANSAC) method, an average profile peaklist that is iteratively evolving over cycles of alignment is used. Since this approach follows bi-dimensional matching across both m/z and RT dimensions, selection of appropriate threshold over both these dimensions for clustering becomes very critical. In case of MetAlign,[47] the RT width used for matching features becomes progressively smaller till it reaches to a lower limit equivalent to peak-width for feature. Another important aspect is the model structure used for

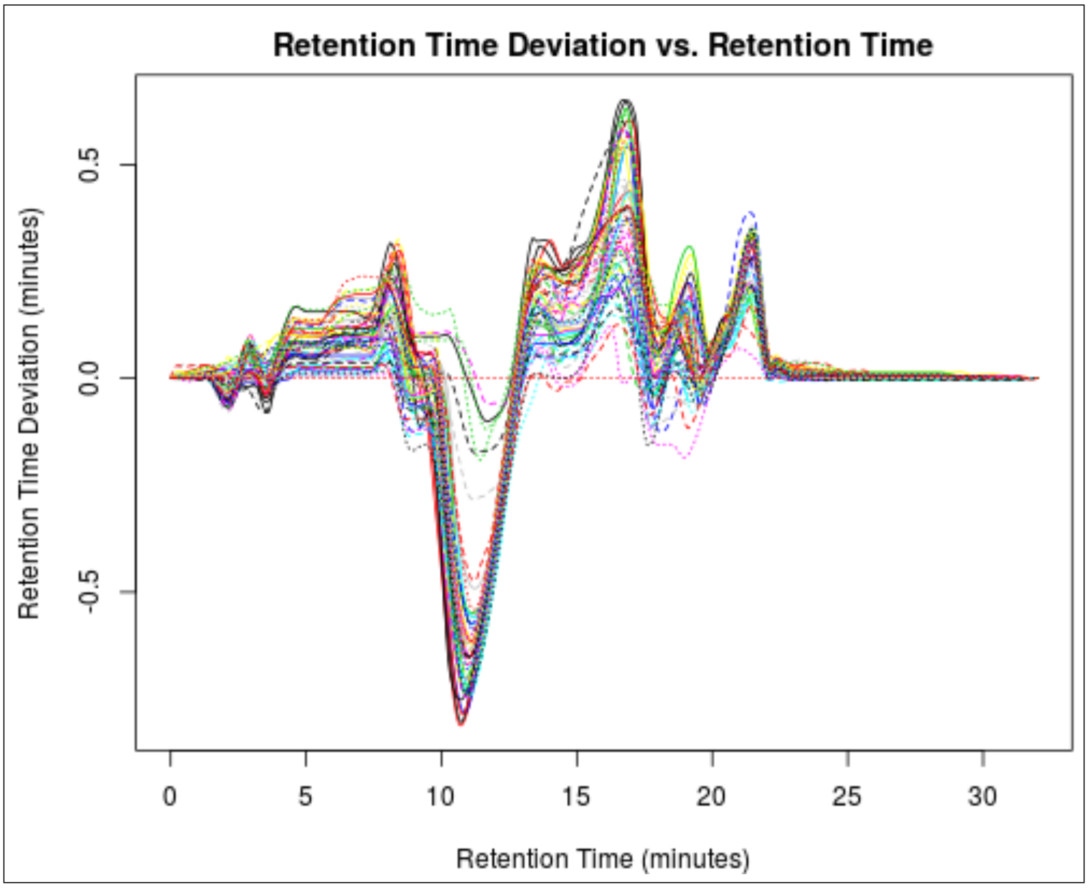


Figure 1.6 Retention time correction profile for a set of 53 LC-MS analyses runs. Retention time deviation towards positive region indicates that the RT for sample was higher than median RT and *vice versa*.

alignment. Considering the non-linear nature of RT shifts, most of the tools/methods follow a single non-linear model fit for RT differences from median point of clustered features.[38] Whereas XCMS offers a local regression fit (LOESS) method that follows local segmented low-order polynomial fits, which offers robustness against local perturbations as well. A comparative study for performance evaluation of alignment algorithms from Xalign (MET-Xalign),[48] msInspect,[53] MZmine,[42] OpenMS,[37] SpecArray and XCMS[40] shows that the algorithms for XCMS and MZmine perform better for metabolomics data.[90] Figure 1.6 shows non linear profile for alignment model using LOESS method of XCMS for a set of LC-HRMS profiles of sample extracts for Gram-negative bacteria, *Chromobacterium violaceum*.

1.2.6 Metabolite annotation

Analyte annotation is an important feature for any algorithm for MS-based metabolomics analysis platform and GUI. Considering the biological complexity in metabolomics study, accurate annotation of unknown metabolites (subsequent to data filtering and feature detection) is a challenging task. In comparison with GC-MS based workflow, metabolite identification from LC-MS becomes further challenging given the list of variables involved in a study such as instrumental parameters, different types columns, separation conditions and fragmentation mechanism employed *etc.* With current developments in modern HRMS instruments with improved mass accuracy, perennial increase in wealth of information about fragmentation profiles studied over last few decades and databases resources along with novel algorithmic development, the task of metabolic annotation has become reasonably feasible. Common strategy for metabolic identification involves:

- the use of tandem MS data for screening of specific fragments ions, neutral loss and precursor ion relation through database search or with standard sample run,

- mass accuracy-based confirmation while using HRMS instruments such as Orbitrap or FT-ICR.

A comprehensive overview of annotation process along with basic information about MS data is reported for detailed understanding.[81] This report provides brief introduction about available database resources and tools for analysis.

A list of commonly used mass spectral libraries is listed in Table 1.3. In general, while scanning for metabolite features through databases, the experimental spectra is matched against reference spectra from database and a matching score is estimated for each feature. Stein *et. al.*[91] have compared different algorithms for spectral matching such as, dot-product function, which measures the cosine angle between spectral features represented as vectors, probability-based matching system that uses peak occurrence statistics, Euclidean distance *etc.* The dot-product function with square root intensity scaling was found to be better method for ranking metabolite features in spectral matching.

Various databases listed in Table 1.3 provide information for matching accurate mass (m/z), fragmentation product (MS/MS) along with the details about ionization mode, ion type and various ionization interfaces such as electrospray ionization (ESI), electron ionization (EI), chemical ionization (CI) *etc.* Because of the data complexity with the presence of different adduct ions, neutral loss fragments and isotopic clusters, spectral matching becomes a difficult task even with the higher mass accuracy offered by HRMS instruments. Use of isotopic clusters to discern elemental formula have been considered as an approach for metabolite annotation.[43] A systematic approach with the help of set of rules for filtering false positive hits from database matching, which includes matching of isotopic abundances as well, has been reported by Kind *et. al.*[92,93] Further, online database tool MZedDB[44] has provided facility to process information related to neural loss, adduct information for each chemical structure. MZmine package provide

Table 1.3 Available database resources for mass spectral library

Database name	Availability	Web address
NIST 14	Commercial access	http://www.nist.gov/srd/nist1a.cfm
NIST MSMS Library	Commercial access	http://www.nist.gov/srd/nist1a.cfm
Wiley Registry of Mass Spectral Data	Commercial access	http://onlinelibrary.wiley.com/book/10.1002/9780470175217
FiehnLib	Download possible	http://fiehnlab.ucdavis.edu/projects/FiehnLib/index.htm
GolmMetabolome Database	Download possible	http://gmd.mpimp-golm.mpg.de/
Human metabolome database (HMDB)	Download possible	http://www.hmdb.ca/
KEGG ligand database	Download possible	http://www.kegg.jp/kegg/ligand.html
Madison metabolomics consortium database (MMCD)	Download possible	http://mmcd.nmrfam.wisc.edu/
Manchester metabolomics database (MMD)	Download possible	http://dbkgroup.org/MMD/
MassBank	Download possible	http://www.massbank.jp/
ReSpect	Download possible	http://spectra.psc.riken.jp/
LipidBank	Web access only	http://www.lipidbank.jp/
METLIN	Web access only	http://metlin.scripps.edu

an interface to search m/z list, along with adduct and fragment information across various online chemical compound databases such as, HMDB, METLIN, and KEGG *etc.* Similarly, output results from XCMS analysis are linked with METLIN database queries for list of peaks identified. Apart from mass information, biological interactions and relationship network can also used as a measure of metabolite detection.[94,95] Few open-source packages offer this approach for untargeted profiling of metabolites in a high-throughput manner.[96,97] An extensive list of tools, which assist in identification of metabolites is reported in review by Yi *et. al.*[98]

1.3 Downstream processing of metabolomics MS data

On the basis of different objectives for a given metabolomics analysis, the longitudinal data involving vast amount of metabolite data obtained through HRMS workflow can be handled with different linear or non-linear data analysis methods. Typically, the processed data obtained through data analysis workflow represented in Figure 1.3 is given in a matrix format with linked information about m/z, ion response and RT index for chromatographic analysis. Various statistical tools with visualization support (optional) can be used to extract further in-depth information about data structure. MetaboAnalyst[39] is a web based tool, specifically targeted for metabolomics analysis, analyses XCMS processed data to draw further conclusions. Apart from routine multivariate statistical approaches for sample classification or functional characterization using quantitative metabolomics profile, other hyphenated analysis tools with applications in metabolic engineering have also surfaced in recent past.[99] In this section a brief introduction about methods used is illustrated. For more comprehensive list of data analysis/transformation methods readers are advised to consult other reviews.[98–100]

1.3.1 Extraction of metabolite features

Various biological experiments are aimed at identifying key list of metabolites involved or affected with specific experimental treatment. This can be achieved by studying statistical features of metabolite or with the help of optimization problem to discern list of optimal metabolite lists relevant to the experimental design. Data analysis approaches for this are either (i) metabolite ranking or (ii) metabolite selection.[101]

Metabolite ranking

Common interest in carrying out metabolite ranking is to distinguish highly important list of metabolites with strong relevance towards the biological variations. Routine approach for metabolite ranking involves variable ranking based on partial least square (PLS) or PCA,[102] loading or factor weights (LW),[103] variable importance on projection (VIP),[104] regression coefficient (RC)[105] and selectivity ratio (SR).[106] A comparative performance study conducted with clinical samples suggest better efficiency using VIP in comparison with LW and RC.[107] Since these methods uses different approaches for ranking of variables, the results generated may also be different. A recent report by Yun *et. al.*[108] employed aggregation of ranked orders generated using different methods.

Metabolite selection

In a similar fashion like metabolite ranking, selecting set of features is also of interest for certain studies. The methods used for metabolite ranking can be extended for selection of metabolites as well with the help of threshold criteria. Using an objective function to evaluate the predictive efficiency of selected metabolites list towards class of sample, a classification model can be generated. The performance measure for this can be obtained with the help of classical bootstrapping method (with re-sampling of data points) or with the help of cross validation (CV) error estimations. In order to avoid over-fitting of the data, care must be taken to employ appropriate performance evaluation strategies such as re-sampling CV dataset and size distribution across training data and test data. An alternate approach with increased robustness of analysis was employed in the form of random forest (RF)[109] and model population analysis for variable selection (MPA).[110] These strategies involves sampling of the dataset into n sub-datasets following Monte Carlo sampling (MCS), and evaluation of classification model developed for each subsets.

Further, these approaches categorize variables on the basis of their contribution for prediction error as informative, uninformative or interfering.

1.3.2 System level understanding of metabolic physiology

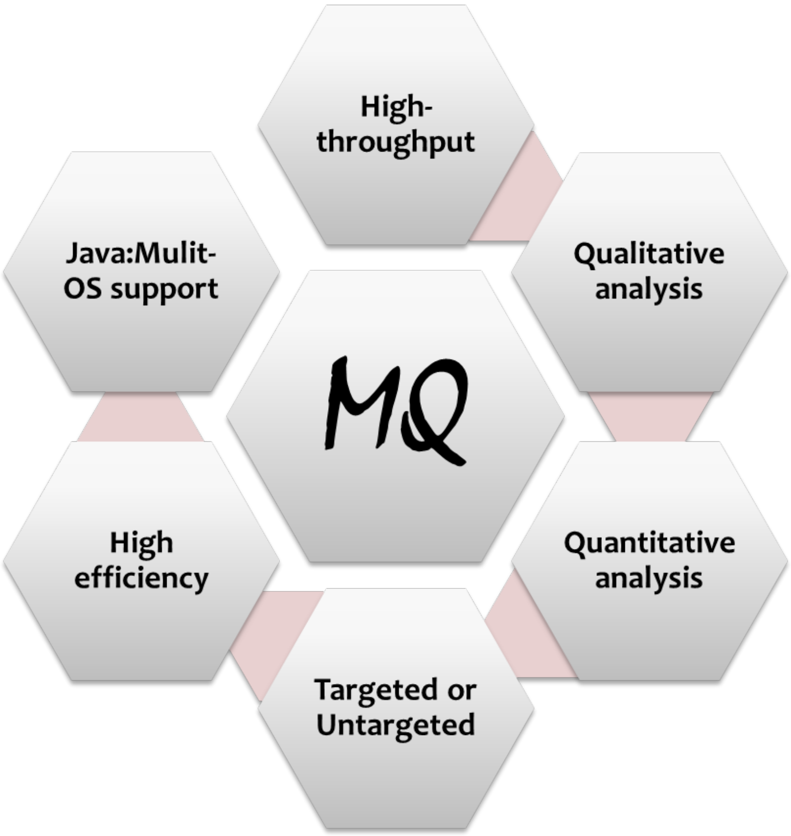
Rapid development in the field of metabolomics was made feasible with availability of global metabolic profiling following HRMS workflows. With increased accuracy of cellular physiology, depicted by cellular metabolome in comparison to its predecessor 'omics' technologies (genomics or proteomics), metabolomics has already paved its way in clinical disease diagnostic applications.[111,112] But unlike other 'omics' platforms with established databases and analysis workflows, metabolomics is still an evolving field with potential showcased by only few analytical strategies. Metabolic network analysis is one such strategy that utilizes metabolic profiling data to dictate in depth mechanistic perspective of cellular physiology.[113,114] Most routinely used metabolic network analysis approaches are, (a) metabolic flux analysis (MFA) and (b) flux balance analysis (FBA) following flux constraints identified using metabolic profiling.

In MFA, systematic tracing of stable isotope labeled (usually ^{13}C) metabolites' fate in metabolic pathway is studied. Distribution of isotopomers across metabolic pathways following differential carbon-carbon transitions provides mechanistic differences across case studies.[115] Such MFA based strategies were shown to assist in applications involving yeast species selection based on efficient aerobic fermentation on glucose or in designing metabolically engineered strains for overexpression of transaldolase and transketolase.[116,117] A list of freeware tools for MFA such as, SUMOFLUX[118], PFA toolbox[119] and many others can be found at http://fiehnlab.ucdavis.edu/staff/kind/metabolomics/flux-analysis.

In case of FBA, an optimal flux solution for a selected biological function is identified for a metabolic network stoichiometry having additional experimentally calculated constraints. This is achieved using linear programming (LP) approach for estimation of flux solution space for all participant reactions under pseudo-steady state condition. The additional experimentally defined constraints are extracellular efflux/uptake rate of targeted list of metabolites. Hence, FBA solution offers a global system-level perspective of possible intracellular metabolic physiology for a system while using very few targeted list of extracellular metabolic profiles. Such versatility offered by FBA led to the development of its various supplement methods (constraints-based methods) with different application.[99] Minimization of Metabolic Adjustment (MOMA) is one such supplementary FBA method, which aims to find feasible flux distribution nearest to original FBA solution in response to a gene knockout.[120] Alper *et. al.*[121] have showed increased lycopene production from an already high producing strain of *E.coli* with application of MOMA. Another such constraints based method is synthetic lethal analysis (to identify cell/tissue specific) based strategies to control cellular survival in the presence of cellular mutational profile.[122] Many published reports have shown application of synthetic lethal analysis in devising therapeutic strategies for cancer treatment.[123,124] Additionally, for interrogation of network's structural interactions over long-range pathway nodes, constraints based approaches of flux variability analysis (FVA) or uniform random sampling of steady-state flux space can be availed. FVA estimates steady state flux range for each reaction for a defined optimal flux through objective reaction. Whereas, random flux sampling populates flux for each reaction randomly within the solution space defined by FVA such that the constraints imposed by steady state of system and optimal flux through objective reaction are respected. This enables unbiased appraisal of metabolic network structure, allowing functional readout of tightly interlinked list of reactions. Taken together, these methods illustrate network-wide effects of changes in an experimentally identified phenotypic state with possible impact on cascade of downstream network reactions. Most of these analyses can be achieved by open-source freeware tool, Constraints Based Reconstruction and Analysis (COBRA) toolbox for MATLAB and Python language platform.[125]

Chapter 2

MQ: High throughput data analysis tool for HRMS data



Salient features offered by MQ

2.1 Introduction

Contemporary high-resolution mass spectrometry (HRMS) provide millidalton (mDa) level resolution and mass-to-charge ratio (m/z) measurement within a few parts per million (ppm) accuracy of the exact mass as opposed to nominal mass based analysis on conventional low resolution mass spectrometry (MS) platforms. It is now theoretically possible to accurately profile and qualify a few thousands of distinct metabolites from diverse biological sources within a single HRMS scan. HRMS coupled to a chromatographic front end (gas chromatography (GC) or liquid chromatography (LC)) is usually the preferred choice for metabolomic analysis.[13,32] Various chromatography-free direct and ambient ionization methods used in conjunction with HRMS analyzers further aid in measurements that offer analysis directly from sample surfaces along with higher throughput.[18] Mass spectral imaging (MSI) using direct ionization methods has shown potential to obtain mechanistic and molecular insights at a cellular, organ and systemic level.[126]

Analysis of HRMS data for diverse applications requires efficient algorithmic tools to extract and process relevant information from raw data. A common data analysis pipeline followed for processing HRMS data is represented in Figure 2.1. Various proprietary and a few open source tools that can support HRMS data analysis are currently available.[40,42,127–130] (Refer Table 1.1 in Chapter 1 for brief list of available tools and supported platform) Usually proprietary softwares for data analysis come bundled with mass spectrometry instruments and support only specific file formats. Inherent raw data inflexibilities can at times be limiting especially if one were to work across platforms from different manufacturers. A few open-source tools such as mMass[127,128], MZmine[42], XCMS[40], Mascot[129] and TOPP[130] are also available. Some of these tools do not include options for absolute quantitation and are geared towards particular workflows (for example proteomics, metabolomics, and lipidomics profiling). Also, data handling in the majority of data processing tools for quantitative analysis is pegged with chromatography-based workflows leaving little room for those exploring direct HRMS measurements devoid of chromatography. Most of the available open source tools were developed using Python, C++ or R platform depending upon different features offered such as the availability of numerical or statistical libraries and modular structure making it flexible to integrate in third party applications. Availability of Java based open source tools for MS data analysis is limited although the Java platform is powered with many more features and has an array of libraries for numerical or multivariate analysis along with recent active development of Mass Spectral Development Kit (MSDK) (https://msdk.github.io/). Herein, we report 'MQ', an algorithm developed using Java platform that attempts to support high throughput HRMS-based targeted metabolite quantitation workflows subsequent to global metabolomic profiling and qualification. MQ has been developed in our group with a broad perspective to aid qualitative and quantitative HRMS data analysis for direct MS analysis, especially to support MALDI-MS workflows. The algorithm has been continuously improvised, rigorously tested and benchmarked vis-à-vis proprietary Xcalibur™(Thermo Scientific) software that is code of federal regulations compliant (CFR, FDA). Herein, we describe various features of MQ and showcase a comparative study for the assessment of its quantitative performance in direct ionization and LC-HRMS metabolomic analysis for a targeted set of analytes. Preliminary versions of MQ were previously used in quantitative non-chromatographic laser desorption ionization mass spectrometry-based analysis workflows[112,131].

Availability and implementation: Freely available on web (for academic use only) at, http://www.ldi-ms.com/services/software. Software is accompanied by example files and user manual. MQ is implemented in Java and supported on Linux and Microsoft Windows system having Java runtime environment (JRE) pre-installed. Additional details and processed data used for this dissertation work can be found at http://bit.ly/dissertationDataAG.

Credits towards sample preparation: Samples for S-adenosylmethionine (SAM), S-adenosine-L-homocysteine (SAH) and atorvastatin were obtained from collaborative experiments with research colleagues Dr. Nivedita and Ijaz. Mammalian cell culture medium and extracellular sample extracts for two neuronal cancer cell lines (U87MG and NSP) were obtained from cell culture experiments performed by Rupa as a part of collaborative study with Dr. Anu Raghunathan.

2.2 Materials and Methods

2.2.1 Chemicals

LC-MS grade methanol and acetonitrile was procured from J.T.Baker (India). SAM, SAH, trifluoroacetic acid (TFA), verapamil ((RS)-2- (3,4-dimethoxyphenyl)-5-2-prop-2-ylpentanenitrile) and 2, 5-dihydroxy benzoic acid (2, 5-DHB) were purchased from Sigma-Aldrich. Melamine (2, 4, 6-triamino-1, 3, 5-triazine) was purchased from Loba Chemie (India). Atorvastatin ((3R,5R)-7-[2-(4-Fluorophenyl)-3-phenyl-4-(phenylcarbamoyl)-5-propan-2-ylpyrrol-1-yl]-3,5- dihydroxyheptanoic acid) was obtained as gratis sample from Mylan Laboratories limited, Hyderabad, India. Deionized water with specific resistance 18.2 MΩ cm^{-1} was obtained from Milli-Q unit (Merck Millipore). Mammalian cell culture medium composed of DMEM (Sigma-Aldrich, D6046) supplemented with MEM Non-essential amino acids solution (Sigma-Aldrich, M7145), was used as standard mixture for all amino acids. All the chemicals used in this study were of analytical grade.

2.2.2 LC-HRMS based quantitative analysis of amino acids

The LC-HRMS instrumentation consisted of autosampler (Accela Open Autosampler, Thermo Scientific) and liquid chromatograph (Accela 1250, Thermo Scientific) in tandem with the Q-Exactive (Thermo Scientific) high resolution mass spectrometer equipped with a heated electrospray ionization (HESI) interface. Instrument operation and data acquisition was performed using the 'Xcalibur™' platform software (Thermo Scientific). A C18 Hypersil gold column (10 cm x 2.1 mm x 3.0 μM) by Thermo Scientific was used for eluting the samples prior to the ESI. The mass analyzer was operated in positive ion mode and data was acquired in triplicates within a mass range of 60-900 m/z at 70,000 FWHM resolution.

For quantitative analysis, mammalian cell culture medium and extracellular sample extracts for two neuronal cancer cell lines (U87MG and NSP) that were characterized in a parallel published study[132], were utilized. Cell culture medium standard mixture was serially diluted to generate calibration curves for the ranges reported in Table 2.2. A total of 10 calibration levels and 2 quality control (QC) samples were used. To investigate differential metabolic exchange profile, sample extracts from every 24 hr were pooled across cellular culture growth over 7 days. A 100 μL of such sample extract was mixed with 400 μL of chilled methanol (previously stored in -80°C). The solution was thoroughly mixed for 2 mins followed by centrifugation for 15 mins at 5000 rpm (4°C). The tubes were carefully removed, 300 μL of supernatant was withdrawn and transferred into a fresh tube. A two-step serial dilution of supernatant was performed using 50 % acetonitrile in water. In the first step, 50 μL of supernatant was thoroughly mixed with 450 μL of diluent. This solution was further diluted by mixing 100 μL of sample solution with 400 μL of diluent. These samples along with standard mix were uniformly spiked with the 2 μM solution of verapamil as internal standard to evaluate the performance and for data normalization. The solutions were thoroughly mixed and were then analyzed on LC-HRMS system. Following the acquisition, the raw data was analyzed using proprietary 'Quan-browser' module from Xcalibur™. Separately, data analysis using MQ was carried out as outlined below.

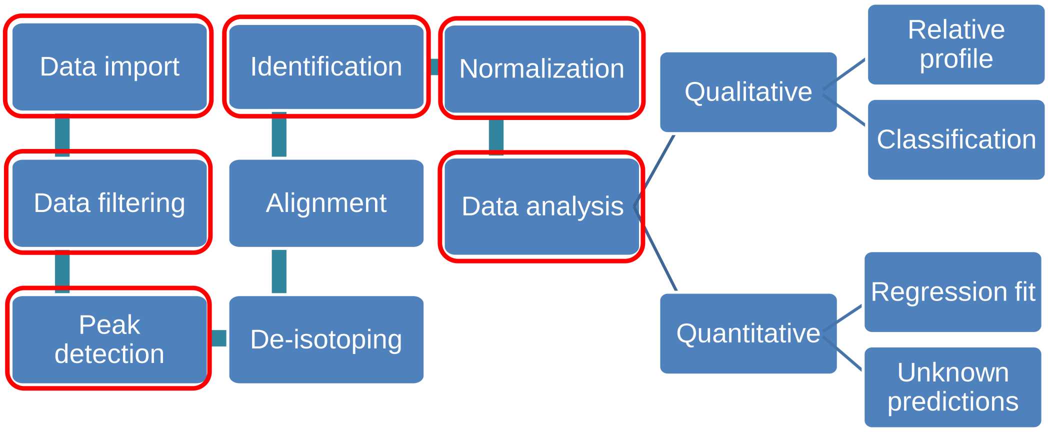


Figure 2.1 Detailed workflow for HRMS Data analysis. Highlighted aspects are featured in MQ

2.2.3 MALDI Q-TOF MS based quantitative analysis of biomarker metabolites

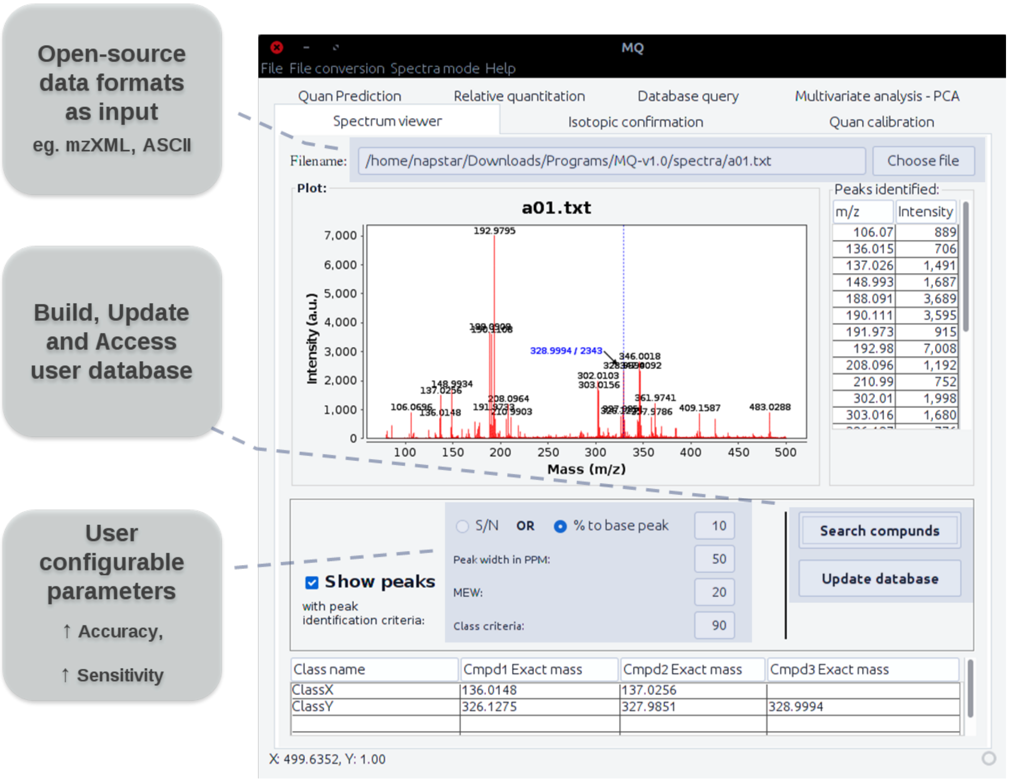
For generating MALDI Q-TOF MS data, a standard solution of SAM and SAH was prepared in methanol : water (1:1, v/v). For generating calibration curves, standard solutions were serially diluted to obtain the predetermined calibration levels and QCs as shown in Table 2.3. An internal standard (IS) solution consisting of 5.34 μM melamine was uniformly spiked in all the samples before analysis to evaluate the performance and for data normalization. To carry out MALDI Q-TOF MS analysis, previously prepared standards were mixed in 1:1 ratio with 10 mg/ml of 2,5-DHB matrix solution, which was separately prepared in acetonitrile : 0.1% TFA in water (1:1, v/v). All samples were spotted on the MALDI target plate by dispensing 1 μL of matrix-analyte mixture in 4 replicates for each of the calibrants. The data was acquired on Waters Synapt HDMS™MALDI Q-TOF instrument in positive ion mode.

2.2.4 Quantitative analysis of atorvastatin using AP-MALDI HRMS

Quantitative analysis of atorvastatin was performed on Q-Exactive (Thermo Scientific) mass spectrometer equipped with an atmospheric pressure (AP) matrix-assisted laser desorption ionization (MALDI) (AP-MALDI) source from Mass Tech Inc. USA. The instrument was operated in positive ion mode and the data was acquired within a mass range of 200-600 *m/z* at 70,000 FWHM resolution. 100 μM stock solutions of atorvastatin were prepared in methanol : water (1:1, v/v). The stock solution was serially diluted to prepare predetermined calibration levels and QCs. 2, 5-DHB (10 mg ml^{-1}) was prepared in acetonitrile : 0.1% TFA (1:1, v/v) and used as the MALDI matrix. 1.5 μL of matrix was spotted on MALDI target plate. Samples premixed with 0.5 μM verapamil, used as an internal standard, were subsequently spotted on dried matrix layer. Following the acquisition, the raw data was analyzed using MQ.

2.3 Overview of MQ

MQ features key aspects of data analysis as depicted in Figure 2.1 and includes a graphical user interface (GUI). Several modules within MQ have been designed to enable peak qualification, feature extraction, relative as well as absolute quantification, and untargeted analysis. Modules namely ‘Spectrum viewer’, ‘Isotopic confirmation’, ‘Quan calibration’, ‘Quan prediction’, ‘Relative quantitation’, ‘Database query’ and ‘Multivariate analysis - PCA’ are bundled into a single platform, as shown in Figure 2.2, to aid seamless user experience. An overview of HRMS data analysis workflow using MQ is illustrated in Figure 2.3. Various aspects of data processing using MQ are described below.



Filename: /home/napstar/Downloads/Programs/MQ-v1.0/spectra/a01.txt

Peaks identified:

m/z	Intensity
106.07	889
136.015	706
137.026	1,491
148.993	1,687
188.091	3,689
190.111	3,595
191.973	915
192.98	7,008
208.096	1,192
210.99	752
302.01	1,998
303.016	1,680

☑ Show peaks with peak identification criteria:

○ S/N OR ● % to base peak 10

Peak width in PPM: 50

MEW: 20

Class criteria: 90

Class name	Cmpd1 Exact mass	Cmpd2 Exact mass	Cmpd3 Exact mass
ClassX	136.0148	137.0256	
ClassY	326.1275	327.9851	328.9994

X: 499.6352, Y: 1.00

Figure 2.2 Screenshot for MQ user interface with available data analysis modules

2.3.1 Data preprocessing

For data analysis with MQ, time averaged HRMS data is used as input data source in generic ASCII format (spectral m/z, intensity list) or mzXML format (common open source format developed by Seattle Proteome Center). Data analysis using time averaged MS spectrum serves as common platform for chromatography-based as well as direct MS-based approaches. LC-HRMS data can be considered as a stack of mass spectra acquired over a period of chromatographic run time. It contains details of ions detected during the process, generating significantly large datasets of information. In a typical LC-HRMS based quantitative analysis peak area, estimated as an extracted ion chromatogram (XIC), is used as a quantitative parameter. This is usually a response of ion current observed from an elution profile specific to an analyte of interest. For quantitative estimations using MQ, this multidimensional data (ion current over the m/z

range as a function of LC runtime) is transformed into an averaged two dimensional mass spectral profile over a specific LC runtime. An averaged HRMS data profile was found to preserve the quantitative features specific to analyte peaks from sample. This approach was benchmarked against Xcalibur™(Thermo Scientific), which follows XIC based quantitative workflow.

Supporting module for generating average MS profile spectra from encrypted manufacturer specific file format is provided as an accompanying file conversion tool in MQ. File conversion using this tool is a two-step process. First the instrument specific file formats are converted to their respective MS1 (MS level 1) profiles using 'MSConvert' module from ProteoWizard package[133]. In the second stage, the MS1 profile data is averaged using in-house built tool 'MSAvg', written in Perl scripting language and GNU Octave environment, to generate two dimensional mass spectral profiles. Owing to the non-homogeneity in distribution of m/z positions across each scan, MSAvg identifies a list of unique m/z data entries from all mass spectral scans acquired in a chromatographic run. While averaging each scan, an interpolated spectral profile is generated using these unique m/z lists as input key for linear interpolation. For such interpolation, the spectral

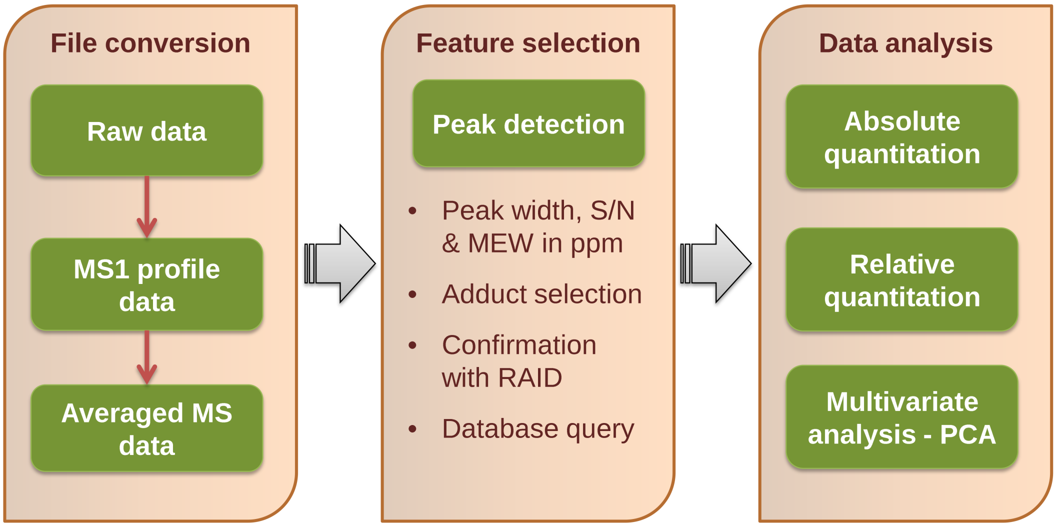


Figure 2.3 Schematic of HRMS data processing steps using MQ

profile needs to be acquired in profile mode, which also imposes as a limitation for its inflexibility towards centroid data. The final averaged mass spectral profile is used for all qualitative and quantitative analysis under different modules of MQ.

2.3.2 Spectrum viewer

Direct visual analysis of MS spectrum is often the quickest way for evaluating various qualitative checks such as presence/absence of a peak, mass accuracy (measured in ppm), signal intensity in ion counts, and the peak width. Spectrum viewer module of MQ offers direct visual analysis with additional options of optimizing peak finding criteria and subsequent database search. The spectrum viewer incorporates many user friendly and handy features such as, zoom-in/out, *m/z* value and intensity annotation. Additional options for spectra exporting (plot graphics to clipboard), saving spectra in image (png) format, and 'properties' option for improving visual attributes can also be availed. Users can also generate a list of peaks from the spectrum through optimization of 'Peak finding criteria' filters. Peaks can be qualified based upon signal to noise ratio (S/N), relative percentage intensity with respect to the highest or base peak, mass extraction window (MEW) within a set ppm and peak width. A convenient database search option is also available in the same window where annotated metabolite list or user created databases can be quickly referred. Additionally, in order to perform a high-throughput database query, a separate database search module is also available featuring similar 'Peak finding criteria' filters. Facility to add list of peaks generated from the spectrum viewer module as a user entry into database is possible.

2.3.3 Peak detection

Various methods for peak detection from MS data have been reported in last few decades. A recent review[76] provides an account on available feature extraction method along with their limitations. Few popular software tools such as, MZmine[42] and XCMS[40] follows fitting of a template (Gaussian or Exponentially Modified Gaussian) function with given mass resolution setting for peak qualification. Various mass analyzers show varying response for ion distribution (resolution) at different *m/z* region[31]. This leads to difficulty in following template function based approach befitting for data from diverse list of MS instruments. MQ uses two step method for feature extraction. In the initial step, first derivative downward zero-crossing over method as peak picking algorithm with constraints over the slope threshold value is used[98]. First derivative spectra are subjected to moving average window (with width equivalent to half of MEW specified in ppm) smoothing. This helps in removing minor kinks that are a result of noise from the data, and increases computational efficiency in peak searching in such areas. Post smoothing first derivative spectra are subjected to peak finding for downward zero-crossing points, which essentially represents the highest point in a peak. Detected peaks are qualified based upon amplitude threshold and slope threshold, which keeps peak kurtosis in check. Slope threshold is defined based upon user specified peak width value as 0.5*(peak width points)$^{-2}$, whereas amplitude threshold is based upon user defined filters such as S/N or relative percentage intensity to the base peak. Subsequently, in the second step a Gaussian function is fitted to the ion distribution observed in the proximity of peak. For this, a second order polynomial function is fit to log transformed ion count response for a set of points within a user defined MEW (in ppm) of peak that helps to capture Gaussian behavior of mass spectral profile. Peak area and peak intensity are estimated from this Gaussian function as area under the curve or based upon peak amplitude value estimation for polynomial fit, respectively.

2.3.4 Qualitative confirmation of analytes using RAID

Natural abundance for heavier isotopes of various elements leads to characteristic relative abundance for isotope intensity distribution (RAID) of analyte peaks, which is specific to elemental compositions. Various reports in past decade highlighted significance of RAID over and above mass accuracy offered by ultra-high resolution MS in characterization of analytes assertively[92,134]. In case of complex and large molecules, prediction of RAID becomes more complicated and computationally intensive[82,135]. In a recent publication, a web based tool for Molecular Isotopic Distribution Analysis (MIDAs) was developed with two improved algorithms for RAID calculation based on polynomial and Fourier-transform methods, having better performance in comparison to published tools for RAID estimation[82]. In MQ, an implementation of polynomial based algorithm from MIDAs web tool for fine grained RAID (used with high-resolution MS) estimation has been incorporated. Provision for adjustable mass accuracy and customizable isotopic abundances are salient features making it amenable to adapt for different experimental designs, such as isotopic labeling.

In brief, RAID estimation following polynomial based method involves multiplication of polynomial expressions for each element. These polynomial expressions are constituted from observed natural isotopic abundance values for each element. For fine grained RAID estimation, polynomial expansion was achieved following multinomial theorem with constraints over allowable exponent for individual isotope abundances in each element. These constraints are function of natural abundance of isotopes and elemental composition of analyte. Further details of algorithm can be found in article by Alves G. *et al.*[82].

With the provision of aforementioned algorithm in MQ for RAID estimation, screening and qualification of analyte with user specified elemental composition can be achieved from given mass spectrum. Confidence measure for qualification is provided as estimates of mass accuracy and percentage error in peak intensity for calculated heavier isotopic peak against observed peak.

2.3.5 Quantitative analysis

Quantitative analysis in MQ can be carried out based upon peak area or peak intensity. A weighted or non-weighted quantitative model with a linear/quadratic regression model fit can be generated for response curves extracted from peak areas or intensity values. Additionally, support for regression model fit for log transformed data is also offered. This can be used for adjusting varying instrumental responses and for fitting regression models of analytes whose concentration ranges vary over several orders of magnitude that usually affect their linear responses.

Calibration models for absolute quantitation can be generated under 'Quan-calibration' module. Here, users can create and utilize a data library consisting of analytes and corresponding monoisotopic masses. An array of parameters can be specified for feature extraction and regression fit for list of analytes. These parameters are, analyte adduct ion(s), MEW (in ppm), weighted/non-weighted calibration model, *m/z* of internal standard used (optional), and spectral file names for the calibrants along with replicates. These parameters for a specific analysis project can be saved in a data library file. A least-square regression fit for all the analytes of interest is then generated that allows high-throughput simultaneous quantification in a single batch processing step. Adduct specific regression data is provided as an output. The best fitting adducts and models can then be selected based on regression statistics such as, slope, %RSD of technical replicates and intercept. Calibration models generated through 'Quan-calibration' module should be loaded in 'Quan-prediction' module to process the samples.

Relative quantification module allows direct comparison of internal standard normalized analyte responses across samples. Input parameters such as, *m/z* list for analyte ions, an internal standard adduct ion *m/z*, and MEW are required for processing. The output of relative quantitation provides the absolute and internal standard normalized peak area or intensity of analytes along with mass accuracy in ppm.

2.3.6 Untargeted profiling and multivariate analysis

MQ supports untargeted profiling and feature extraction using multivariate analysis in a metabolomics study. Data generated through full-scan HRMS has a multidimensional profile that poses a challenge for untargeted analysis. Unsupervised multivariate analysis is an unbiased means to identify a signature set of metabolites, which can be accounted as discriminative features. We have incorporated principal component analysis (PCA) as linear and unsupervised method for multivariate analysis of metabolomics data. PCA orthogonally transforms input spectral data by rotating the variables in coordinate space such that newly formed variables (Principle component factors- PC) should have maximum relevance with the variance within data. These transformations are a result of projecting original data points onto PC space identified by linear combination of the original variables, and thus it does not lead to loss of information. Additionally, a detailed analysis of these PC would help in identifying relevant list of analyte peaks, which holds higher coefficient values for linear combinations. These set of peaks are responsible for features represented by respective PC. For this multivariate analysis input variable data is provided either in terms of identified peak list with the use of 'Peak finding criteria' as mentioned in 'Spectrum viewer' module or intensity response for list of metabolites obtained by database query. Additional available set of parameters are: MEW for *m/z* grouping and normalization of peak intensity response using internal standard.

2.4 Results and Discussion

The performance of MQ was evaluated by simultaneous quantification of multiclass analytes that includes amino acids, metabolite biomarkers and pharmaceutical drug. The set of these three case studies was chosen to establish performance scaling for handling and analysis of data generated from varied analytical complexity, in addition to benchmarking using proprietary data analysis software.

A conventional chromatographic quantitative workflow for LC-HRMS analysis of a set of amino acids from biological matrix was used for comparison of data analysis using MQ and Xcalibur™(Thermo Scientific) software. Acquired data in native format with LC profiles was used as input for Xcalibur™based quantitative analysis, where peak area represented an area under curve of ion response observed from a chromatographic elution profile specific to an analyte of interest. In case of MQ, as discussed before, average mass spectral profile was used as input dataset for analysis. The accuracy and efficiency of quantitative workflows in MQ was evaluated at different levels of data processing such as, data conversion, identification of peak, peak integration, calibration curve fitting and unknown prediction. The calibration curves were examined in terms of %RSD of technical replicates, intercept, slope and regression coefficient (r^2).

Evaluation of quantitative characteristic preserved by MSAvg, post data transformation of LC-HRMS data into two dimensional mass spectral profiles, is showcased in Table 2.1. A test sample representing a dilution level from calibration solutions of mammalian cell culture medium standard mixture was used for this comparison. Although the peak area estimated by Xcalibur™for LC-HRMS data was different in comparison to peak area, for respective list of analytes, estimated from averaged mass spectral profile using MQ. A significant Pearson's correlation coefficient for a pairwise comparison across these different modes of peak area estimation was observed. This represents merits of preserved linear dependency of ion abundance with signal count for data transformed using MSAvg (Table 2.1). Further these peak areas were evaluated following sample test for equal variance (F-test). Significantly low estimated F-test statistic (in comparison to F-critical value of 0.0256) and p-value for rejection of null hypothesis, ratifies the unhindered quantitative nature of peak response following data transformation (Table 2.1).

Further, the output results of absolute quantitation of metabolites using MQ following LC-HRMS analysis is shown in Table 2.2. Results reported contain the comparison of slope and linear regression coefficients obtained for the calibrations processed using MQ and Xcalibur™. Regression coefficients of above 0.9 r^2 indicate excellent linearity for the calibration curves within the used concentration range of 0.025 - 19.98 μM, that are estimated in close proximity, by both MQ and Xcalibur™. The slopes for both the cases were also a close match largely and indicate similar responses and sensitivity for the methods. Two sets of QC samples to cover the broad calibration range were used to test the calibrations generated. The results obtained were reported as percentages relative to the expected recovery of 100%. The recoveries for all the amino acid QC samples for both the higher and lower concentration ranges were quite close to the expected recoveries and well within a generally acceptable precision of 15% relative standard deviation (%RSD). The data for proline shows greater deviation between MQ and Xcalibur™data. Most significantly, the recoveries obtained for the rest of the amino acid QC samples using MQ and Xcalibur™were strikingly similar. This is in spite of the fact that input for Xcalibur™was native raw data format, while for MQ it was transformed into averaged spectrum format. These results clearly establish that the spectral averaging preserves the quantitative characteristics of the data and benchmarks both qualitative and quantitative analysis using MQ.

Table 2.1 Statistical evaluation of quantitative features preserved post data transformation into two dimensional mass spectral profile

	Mean peak area (MQ)	**Mean peak area (Xcalibur™)**	**Pearson correlation(r^2)**	**F-test statistics**	**F-test P value**
Ala	7.90E+05	2.30E+07	0.9900	9.17E-11	1.83E-10
Arg	2.73E+06	1.87E+08	0.9974	2.47E-09	4.94E-09
Asn	2.39E+05	1.67E+07	0.9678	1.45E-09	2.90E-09
Asp	4.49E+05	3.22E+07	0.7993	4.42E-11	8.83E-11
Cyst	4.23E+05	2.90E+07	0.9999	3.94E-09	7.87E-09
Glu	4.08E+05	2.77E+07	0.9894	1.27E-09	2.55E-09
Gln	8.40E+06	5.70E+08	0.997	1.12E-09	2.24E-09
Gly	1.15E+06	7.31E+07	0.9995	1.97E-10	3.93E-10
His	1.33E+06	9.16E+07	0.9997	1.50E-09	2.99E-09
Leu	4.06E+07	2.78E+09	0.9815	8.11E-10	1.62E-09
LYS	4.31E+06	2.95E+08	0.9999	1.05E-09	2.10E-09
MET	1.97E+06	1.35E+08	0.9995	1.15E-09	2.31E-09
PHE	6.22E+06	4.25E+08	0.9904	2.29E-09	4.59E-09
Pro	3.43E+06	7.34E+07	0.9982	5.34E-10	1.07E-09
Ser	1.22E+06	8.23E+07	0.9997	5.88E-10	1.18E-09
Thr	4.01E+06	2.77E+08	0.9989	7.31E-10	1.46E-09
Try	4.59E+05	3.12E+07	0.9996	4.47E-09	8.93E-09
Tyr	2.14E+06	1.47E+08	0.9993	2.43E-09	4.86E-09
Val	2.12E+07	9.58E+08	0.9714	1.04E-09	2.07E-09

Table 2.2 Quantitative information obtained subsequent to the LC-HRMS analysis of amino acids from a chemically defined mammalian cell culture media. Results obtained from MQ data processing were benchmarked using Xcalibur™(Thermo Scientific).

Analyte	**Ion (m/z)**	**Calibration range (in μM)**	**Slope**		**Regression**		**QC sample 1**			**QC sample 2**		
								MQ	**Xcalibur™**		**MQ**	**Xcalibur™**
			MQ	**Xcalibur™**	**MQ**	**Xcalibur™**	**μM**	**% Recovery (% RSD)**	**% Recovery (% RSD)**	**μM**	**% Recovery (% RSD)**	**% Recovery (% RSD)**
Ala	90.0557	0.025-0.495	0.131	0.118	0.982	0.989	0.099	108.9(9.10)	107.1(8.73)	0.174	107.9(5.17)	102.5(2.42)
Arg	175.1192	0.100-1.990	1.026	1.004	0.999	0.998	0.398	113.7(2.07)	113.7(1.47)	0.701	106.7(4.38)	107.1(4.02)
Asn	133.0607	0.025-0.495	0.088	0.088	0.991	0.988	0.099	124.8(3.48)	123.5(2.31)	0.174	106.5(7.14)	106.0(6.22)
Asp	134.0450	0.025-0.495	0.159	0.163	0.994	0.993	0.099	114.4(3.27)	115.6(3.06)	0.174	104.5(4.02)	103.3(3.30)
Cystine	241.0311	0.050-0.995	0.158	0.154	0.998	0.996	0.199	113.9(5.41)	113.5(5.62)	0.350	108.3(5.57)	108.4(4.91)
Gln	148.0607	0.999-19.980	3.010	2.915	0.996	0.992	3.996	116.2(2.23)	123.3(3.26)	7.035	106.2(8.34)	106.0(5.11)
Glu	147.0767	0.025-0.495	0.146	0.142	0.995	0.994	0.099	119.8(2.67)	117.1(3.15)	0.174	104.7(5.36)	106.9(8.09)
Gly	76.0393	0.125-2.495	0.401	0.372	0.995	0.992	0.499	119.2(2.81)	121.5(2.28)	0.879	106.0(4.24)	107.0(4.47)
His	156.077	0.050-1.000	0.495	0.486	0.999	0.996	0.200	120.3(2.55)	120.0(2.71)	0.352	109.0(5.24)	109.2(5.05)
Leu	132.1021	0.200-4.000	14.711	14.467	0.996	0.995	0.800	111.3(3.15)	111.1(2.79)	1.408	103.2(3.22)	102.3(2.68)
Lys	147.1130	0.200-3.995	1.616	1.575	0.999	0.997	0.799	114.6(2.01)	117.1(1.88)	1.407	106.7(4.56)	107.8(4.42)
Met	150.0577	0.050-1.005	0.731	0.690	0.993	0.992	0.201	114.1(3.09)	114.2(3.58)	0.354	104.1(3.70)	104.4(4.18)
Phe	166.0864	0.100-1.995	2.225	2.166	0.992	0.990	0.399	112.4(1.71)	114.0(1.84)	0.702	103.5(3.74)	103.8(2.97)
Pro	116.0706	0.025-0.495	0.440	0.340	0.939	0.903	0.099	123.1(4.56)	105.7(18.85)	0.174	117.3(6.97)	107.9(4.96)
Ser	106.0504	0.125-2.495	0.419	0.408	0.992	0.990	0.499	127.8(3.29)	128.3(3.10)	0.879	106.2(6.11)	106.5(5.82)
Thr	120.0659	0.198-3.950	1.420	1.395	0.992	0.990	0.790	120.2(3.31)	120.9(3.06)	1.391	107.8(5.29)	106.4(4.95)
Try	205.0966	0.020-0.390	0.166	0.160	0.991	0.988	0.078	114.2(4.61)	115.0(4.50)	0.137	103.1(3.33)	103.4(3.29)
Tyr	182.0816	0.115-2.300	0.763	0.747	0.993	0.992	0.460	118.6(1.56)	119.1(1.73)	0.810	107.1(4.79)	107.5(4.23)
Val	118.0868	0.201-4.010	5.825	5.067	0.992	0.997	0.802	110.3(2.95)	109.3(3.18)	1.412	105.1(2.97)	103.9(2.41)

Note: Analyte names are specified as three letter codes for amino acids. Cystine is the oxidised dimer form of amino acid cysteine. Quality control samples were used at two different concentrations levels, with QC sample 1 - lower concentration level, QC sample 2 - higher concentration level. Quantitative features used in this study correspond to area under the extracted ion response - for Xcalibur™and cumulative ion intensity response from averaged mass spectral profile - for MQ.

MQ was further tested successfully for the analysis of data acquired using chromatography free direct ionization source, MALDI coupled with TOF MS for metabolite disease biomarkers (S-adenosylmethionine: SAM, S-adenosylhomocysteine: SAH) and AP-MALDI coupled with Q-Exactive HRMS for pharmaceutical drug (Atorvastatin). SAM and SAH concentration levels are considered as a measure for cellular DNA methylation capacity and have been implicated in various pathological disorders[136,137]. Whereas, atorvastatin is one of the most prescribed drugs belonging to the 'statin' class for treating high cholesterol levels. The calibration models were successfully generated and recoveries of QC samples were also estimated. Table 2.3 summarizes the quantitative information obtained from MQ data processing subsequent to data acquisition. The calibration ranges were within the concentration range of 0.5 to 10 μM for SAM, SAH analysis and 1 to 10 μM for atorvastatin. Calibration curves with excellent linearity were obtained with regression coefficient of above 0.9 r^2 for all three analytes. Two QC samples, which cover higher and lower concentration range of calibration model, were used to estimate quantitative performance using measures of percentages recovery and percentage relative standard deviation for the estimations. For most of the QC samples % recovery obtained were closer to expected recovery concentration and with estimation precision below 15% in terms of %RSD. For SAM and SAH, % recovery estimated for lower concentration range QC was lower, especially with a higher deviation across replicates for SAM. This can be attributed to the spot-to-spot variations commonly observed in MALDI analysis, which can be improved using startegies like stable isotope labelling. Nonetheless, these results showcase the applicability of MQ for quantitative estimation of analytes using chromatography free direct ionization sources. Previously, MQ based quantitation of analytes from complex matrices such as plasma, urine and food matrix using MALDI based direct MS methods have also been reported.[112,131]

In addition, to showcase the specificity and sensitivity offered by MQ for feature selection

Table 2.3 Quantitative information obtained from MQ data processing subsequent to MALDI-HRMS analysis of various analytes

Analyte	**Ion (m/z)**	**Calibration range (in μM)**	**Slope**	**Regression**	**QC sample 1**		**QC sample 2**	
					μM	**% Recovery (% RSD)**	μM	**% Recovery (% RSD)**
SAM[a]	399.1395	0.75-10	0.054	0.97	2.5	78 (22.5)	8	110 (10.4)
SAH[a]	385.1286	0.75-10	0.237	0.93	2.5	73 (10.9)	8	116 (12.2)
Atorvastatin[b]	559.2603	1-10	0.089	0.96	3.5	100 (12.62)	7.5	102 (5.3)

Note: [a]Biomarker metabolites analysed using Waters Synapt HDMS™MALDI Q-TOF.
[b]Pharmaceutical drug analysed using Q-Exactive (Thermo Scientific) equipped with AP-MALDI from Mass Tech Inc. USA.

and quantitative estimation of metabolites from complex biological matrix, extracellular milieu from a glioblastoma cancer cell line U87MG and its phenotypically different subpopulation of neurospheroidal (NSP) cell line was analyzed. LC-HRMS data of cell culture samples was used for comparative evaluation of relative quantitative estimations across the two cell lines (U87MG/NSP), for a set of amino acids, using relative quantitation module of MQ and 'Quan browser' module from proprietary software Xcalibur™. The results for this analysis are illustrated in Figure 2.4. As expected all the amino acids showed a comparable estimation of relative quantitative profiles, across MQ and Xcalibur™. These results further illustrates application of MQ for reliable quantitative analysis of metabolites from a complex biological matrix, with the help of time averaged mass spectral profile of LC-HRMS data.

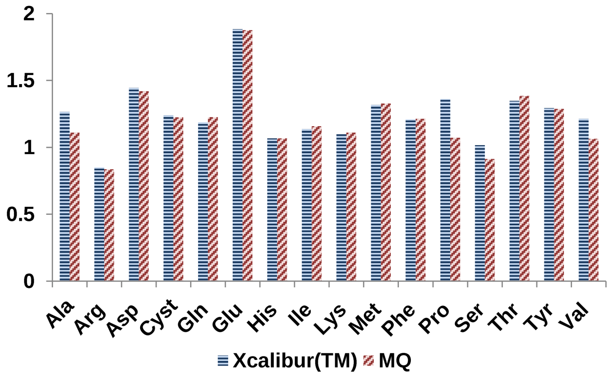


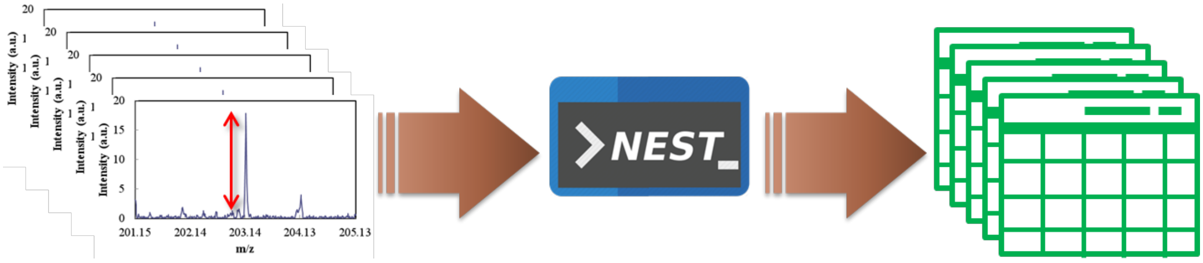
Figure 2.4 Quantitative performance of MQ in comparison to proprietary software Xcalibur^{TM}for a list of metabolites from cell culture samples. Analyte names are specified as three letter codes for amino acids. Cystine is the oxidised dimer form of amino acid cysteine.

2.5 Conclusion

Significant mass resolution offered by modern mass analyzers has encouraged the application of full scan mode MS analysis for reliable metabolite feature annotation along with qualitative and quantitative analysis. In order to accomplish this, a rigorous set of constraints that take into account high mass accuracies for peak qualification along with naturally present isotopic peak distributions are widely accepted criteria[31,92,134]. MQ incorporates the efficient MIDAs algorithm[82] for relative isotopic abundance confirmation towards this end. A high-throughput database query workflow and PCA based multivariate clustering analysis can further benefit qualitative metabolic profiling. MQ offers flexibility with features such as, (a) availing mzXML and ASCII input data formats that are independent from proprietary raw data, (b) user configurable parameters for peak feature detection and (c) compatibility with both chromatography based and direct mass spectrometry methods. Seamless Qual-Quan integration is feasible using MQ through the benchmarked quantitative module that caters to both relative and absolute quantitation. Applications beyond the experimental scenarios showcased in this chapter are possible and include broad areas of food, pharmaceutical and clinical analysis.

Chapter 3

NEST: Tool for high-throughput estimation of S/N from HRMS data



Schematic for high-throughput S/N estimation using NEST

3.1 Introduction

Chromatography-free mass spectrometry methods using ambient and direct ionization sources have found use in a several applications such as pharmaceutical characterization, bio-molecular identification, forensic studies, food contaminant analysis and targeted metabolite analysis.[30,73,81,112] Together coupled with high resolution analyzers such as time of flight mass analyzer (TOF), reliable and in some cases, semi-quantitative analysis is also feasible. In an analysis coupled with chromatography, background signal response in proximity of the resolved chromatographic peak, constituting mostly chemical noise based ion current response, is generally used for the determination of signal to noise ratio (S/N) and determining the limits of detection.[69,72] Direct MS analysis necessitates appropriate approaches for S/N determination in the absence of chromatographic separation.

Broadly mass spectral noise specific to a mass spectrometer can be attributed to: (a) white noise, which is also termed as electrical noise, and is independent of signal response, (b) shot noise, which is also known as Poisson noise, due to the discrete nature of signal response, and (c) chemical noise, which originates from matrix ions forming weakly bound complexes of analyte ions with solvent molecules.[138] The amount of contribution from these three sources could vary as per different mass analyzers. TOF analyzer data shows confluence of all the three with strong influence of chemical noise, whereas FT-MS analyzer like, Orbitrap, has minimal chemical noise in comparison to shot noise along with thermal noise from preamplifier.[139,140] Mass spectral noise poses as one of major limiting factor in reliable detection as well as quantitation of trace level analytes. Various approaches for S/N ratio estimation ranges from a simple visual discrimination, variance from background signal[69] to complex fitting of signal or noise model to mass spectrometry data.[138] These methods provide variable performance for different type of analyzer and hence it becomes difficult for their orthogonal usage when one is performing a comparative analysis. Apart from fitting specific analytic equation to mass spectral profile, numerical estimation of noise can be an alternative.[85]

Various reports discussing different strategies for estimation of mass spectral noise, adapted for diverse set of mass analyzers, can be found in existing literature.[138,141,142] A common approach is to use either maximum peak-to-peak signal deviation or median absolute deviation (MAD) in the baseline response for a selected region with low peak density[143] or across mass range of analysis.[144] But such estimations can be an over estimate for the background noise when using the maximum amplitude of background variation in specified region. They can be skewed sometimes because of the stochastic nature of electronic noise, and inapplicable to newer generation Fourier transform mass spectrometry (FT-MS), where the acquisition software removes background signal and replace it with zero-padding.[142] A different approach for estimation of S/N using spectral data, has been demonstrated earlier.[85] The underlying assumption was that selected spectral region is abundant with background peaks for accurate estimation. Though, this strategy could fairly estimate noise and baseline values of the spectral profile from data originating from most mass analyzers, it suffers for data with less intense background response or sparsely distributed background peaks.

Herein, we present NEST, an algorithmic implementation of this strategy to offer workflow automation, using open source GNU Octave language. Adaptation and benchmarking of the algorithm for non-chromatographic direct high resolution mass spectrometry analysis workflows is demonstrated using dimethyl arginine as an example.[85]

Availability and implementation: Freely available on web at (for academic use only): http://bit.ly/NEST-MS. NEST is implemented in GNU Octave and supported on Linux and Microsoft Windows system having GNU Octave platform installed. Additional details and processed data used for this dissertation work can be found at http://bit.ly/dissertationDataAG.

Credits towards sample preparation: Samples for simulated urine matrices representative of normoglycemic and proteinuric conditions were obtained from collaborative experiments with research colleague - Dr. Nivedita.

3.2 Materials and Methods

3.2.1 Chemicals

Commercial standards of ultrapure 2,5-dihydroxybenzoic acid (2,5-DHB), NG, NG'-dimethyl L-arginine di (p–hydroxyazobenzene – p'-sulphonate) salt (SDMA), NG, NG-dimethyl arginine hydrochloride (ADMA), potassium chloride, sodium chloride, urea, citric acid, potassium phosphate, creatinine, sodium hydroxide, sodium bicarbonate, bovine serum albumin (BSA), LC-MS grade acetonitrile (ACN), methanol and trifluoroacetic acid (TFA) were purchased from Sigma Aldrich. Ascorbic acid was purchased from Loba Chemie (India). Sulfuric acid was purchased from Merck. Deionised water with specific resistivity 18.2 MΩ cm^{-1} was collected from SG ultrapure water unit (Germany).

3.2.2 Sample preparation

In order to represent biological complexity, simulated urine matrices representative of normoglycemic and proteinuric conditions were prepared according to a previously published protocol.[145] For preparing proteinuric simulated urine samples, BSA in concentrations of 350 μg protein/mg creatinine was added to simulated urine matrix. The prepared samples were aliquoted and stored at -20°C.

Stock solutions of ADMA, SDMA and 2,5-DHB was prepared in 50% acetonitrile (0.1% TFA). Mixture of 4 μM ADMA and SDMA was prepared in three different tubes and evaporated to dryness. 100 μL of different simulated matrices were added to the different tubes and 300 μL of ice cold methanol was added to each of these tubes. The simulated matrices were diluted 10 and 50 times with acetonitrile: water (0.1% TFA).

3.2.3 MALDI mass spectrometry (MS) based analysis of simulated matrix

Cross-platform comparative analysis were performed on Waters Synapt HDMS with the matrix-assisted laser desorption ionization (MALDI) ionization source operated in reflectron V-positive ion mode and AB Sciex 5800 MALDI TOF/TOF mass spectrometer. For Synapt, detector voltage of 1750 V was used following manufacturer recommended detector sensitivity test. Optimized laser source (Nd:YAG, 355 nm) energy was used for all acquisition of spectra. Root mean square mass accuracy was maintained within 5ppm by instrument calibration with PEG (mixture of PEG 200, 600 and 1000) before acquisition for samples.

1 μL of matrix was spotted on individual wells of MALDI target plate and dried in air. All the simulated matrices and their dilutions were spotted on an AB Sciex 96-well MALDI-TOF/TOF target plate pre-spotted with matrix. Only normoglycemic simulated urine matrix was also spotted on Waters 96-well MALDI Q-TOF target plate pre-spotted with matrix.

3.3 Results and Discussion

3.3.1 Algorithm for estimation of noise using NEST

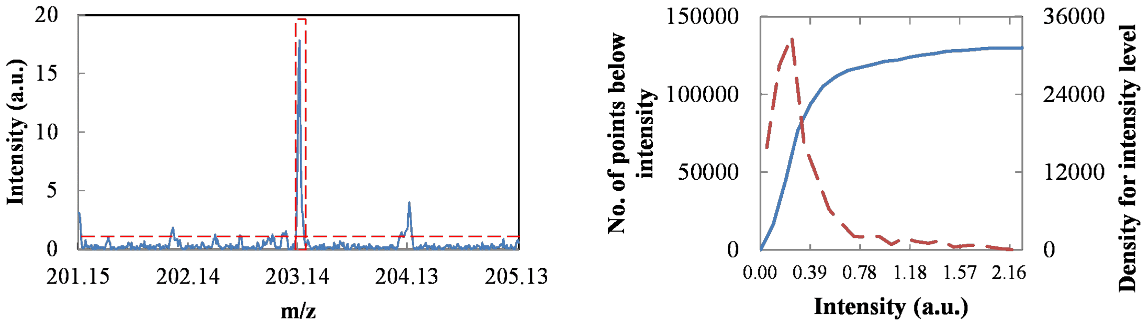
Detailed description of the S/N estimation has been previously reported.[85] The spectral profile is transformed into a paired list of *m/z* and ion abundances in either profile mode or centroid mode (needed for FT-MS data) to process data using the current algorithm. The following user defined parameters are required by the algorithm, (a) *m/z* of peak of interest, (b) m/z extraction window (MEW$_{peak}$) size represented in ppm for signal identification, (c) m/z range (MEW$_{background}$) for noise estimation, and (d) bin size for density profiling of the peak intensities in the defined MEW$_{background}$. A subset of the paired list of m/z and ion abundances in the proximity of peak of interest are extracted within the MEW$_{background}$. This subset list is processed to identify the S/N for peak of interest by the following equation:

$$\frac{S}{N} = \frac{1}{n}(s - b) \tag{3.1}$$

Here, s denotes signal response for peak of interest, while b and n denote the baseline intensity and noise intensity within MEW$_{background}$. Figure 3.1a, illustrates the estimation of s within the specified 50 ppm MEW$_{peak}$ and 4 Da MEW$_{background}$. A cumulative distribution function profiling the number of ion abundances having a signal response equal or below a given intensity value was estimated for the above ranges (Figure 3.1b - solid blue line). First derivative transformation of this distribution function, which provides the probability density function of ion abundance, is also depicted (Figure 3.1b - dashed red line). This is also a representation of the frequency of occurrence of the ion abundances under consideration. The highest point in this probability density function is a measure for the baseline intensity b, while the peak distribution of probability density function, FWHM of peak, denotes noise intensity n. The signal-to-noise ratio can thus be calculated by following Equation 3.1.

3.3.2 Benchmarking with other methods

Performance of NEST was evaluated vis-à-vis methods and tools currenty available either as freeware or instrument specific software. These are US pharmacopeia method[146], MALDI-Quant tool's S/N estimation method, mMass feature annotation tool, Data Explorer™(AB Sciex tool for instrument data analysis) feature annotation tool.



(a) Spectral region of interest

(b) Density profile for the baseline intensity

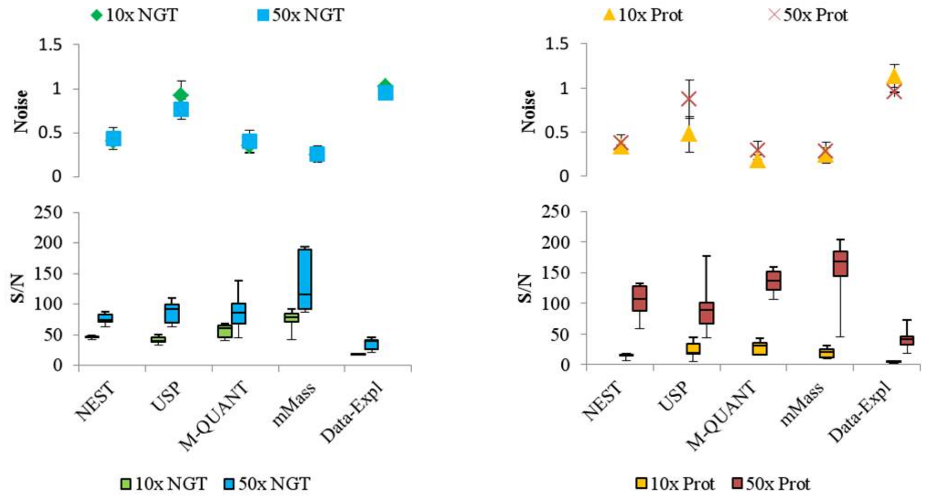
Figure 3.1 Data processing for estimation of S/N for peak of interest. (a) Mass spectral window showing analyte peak of interest. Highlighted dashed lines represent user specified mass extraction window (in ppm) for peak selection along with background signal used for analysis (bottom horizontal) line. (b) Density profile for the baseline intensity. Solid blue line represents cumulative distribution profile for ion abundances for the baseline noise. The dashed red line denotes the frequency of occurrences for ion abundance over the range of values studied.

All these methods utilize different approaches to estimate noise and S/N ratio and hence such study will provide a rigorous outlook for benchmarking our method. In brief, the US pharmacopeia method[146] calculates noise as difference in observed maxima and minima of background ion intensity signal in proximity of analyte peak spanning at least 5 times of its FWHM. MALDI-Quant utilizes R platform base median absolute deviation (MAD) method that estimates median of the absolute deviation from median of ion intensity values. Open source mMass calculates S/N ratio with Equation 3.1, by estimating ion intensity for analyte peak signal and baseline signal from spectral noise as median of all data points that can be tuned by user set parameters. Data Explorer™annotates peaks in similar fashion by using background signal from user specified trace of spectral region to estimate root mean square (RMS) noise. Amongst these, for the US pharmacopeia published method, one needs visual determination of the background maxima and minima, making it susceptible to user-to-user variation. In the case of MALDI-Quant, estimation of a single noise signal value per spectra is performed. But, shot noise component of signal noise is proportional to square root of the analyte signal intensity. Hence, spectra that consist of peaks with intensity spanning over multiple orders of magnitude, such peaks could influence noise intensity in proximal *m/z* range. It would thus be apt to estimate noise based on background signal in the vicinity of analyte peak of interest.

3.3.3 Estimation of S/N of dimethyl arginine

Matrix effects in samples of biological origin present a challenge in the accurate measurement of metabolites endogenously present. Dimethyl arginine is a marker for renal insufficiency and is present in urine. In the case of urine from renal subjects, the presence of protein and potential binding of the metabolite might influence the signal and pose difficulty in comparisons with normal subject samples. In such cases, appropriate dilution with or without solid phase extractions are used to optimize the S/N. The algorithm developed herein can serve as a useful tool in such method optimization especially for direct mass spectrometry analysis.

Towards this end, MALDI MS of dimethyl arginine spiked at physiologically relevant concentrations in simulated urine was performed. Surrogate urine was prepared as per existing protocols and protein content was added separately to simulate the diseased condition. A series of 2 dilutions levels on the samples (10x dilution and 50x dilution) were studied. Noise and S/N estimations using this algorithm described have been showcased in Figure 3.1 (A and B) with comparisons using existing methods / algorithms (Figure 3.2). All these methods show enhanced S/N across dilution levels from both types of samples, which can be attributed to lowered matrix effects. As expected, dilution of urine samples results in an increase in S/N value for the dimethyl arginine. This can be attributed to the ion suppression effect of ion source. Owing to the biological complexity with biomolecules ranging in different concentration levels, analytes present in trace concentrations have to compete for ionization from other pool of biomolecules leading to ion suppression or enhancement effects. Dilution of samples would address such undue



(a) Simulated normoglycemic urine samples

(b) Simulated proteinuric urine samples

Figure 3.2 Effect of different levels of dilution on noise and S/N values for dimethyl arginine from two classes of simulated urine samples. (a) Normoglycemic (NGU), (b) Proteinuric (PU). Dilution levels: 10x dilution – green (NGU), yellow (PU), 50x dilution – blue (NGU), red (PU). Scatter plot at the top represents average noise estimated from replicate samples with standard deviations as error bars. Box plot at the bottom represents average S/N estimation showcasing minima, 1^{st} quartile, median, 3^{rd} quartile and maxima from replicate samples.

ion suppression effects. These trends can be seen in Figure 3.2 from S/N estimation using all methods. DataExplorer™estimates higher noise values (∼1) leading to lower S/N values compared to other methods. As a measure of performance evaluation using these methods, we have compared the standard deviation of predictions from replicate samples across different methods. It can be observed that the deviation in predicted signal noise value is higher for US pharmacopeia method as a result of stochastic nature of baseline noise in addition to manual errors from user interventions. For S/N estimation, a higher variation in calculated values for 50x proteinuric sample sets was observed for US pharmacopeia and mMass tool. Similar higher variations can be observed for 50x normoglycemic urine mix sample sets for S/N estimation following methods from US pharmacopeia, MassQuant and mMass tool. Overall, performance of NEST for S/N estimation was found to be at par with publicly available data anlysis tools, such as MALDI-Quant and

Table 3.1 Comparison of S/N and noise value estimation across different MS platforms for normoglycemic simulated urine samples

		Average	RSD
AB SCIEX TOF/TOF	S/N	1.03	47.81
	Noise	0.39	21.8
Waters Synapt G1	S/N	170.08	46.44
	Noise	12.60	13.92

mMass. Consistency of noise estimations across replicates (Figure 3.2), having minimal variance, and amenability to use in high-throughput manner using batch mode, makes it a versatile tool for analyst pursuing analytical method development.

For a cross platform comparison, we have analyzed data generated using AB Sciex 5800 series MALDI TOF/TOF MS instrument and Waters Synapt G1 HDMS instrument in tandem with MALDI ion source. As a case study, dimethyl arginine mix in normoglycemic simulated urine matrix was analyzed using both of these instruments and results are illustrated in Table 3.1. The estimated values for S/N along with noise were an order higher in case of data generated using Waters Synapt instrument compared to AB-Sciex TOF/TOF instrument. The difference in the instrument architecture for analyzer and especially detectors relating to differential sensitivity and ion signal response can be expected. This might lead to the altered estimates for noise and subsequent S/N values.

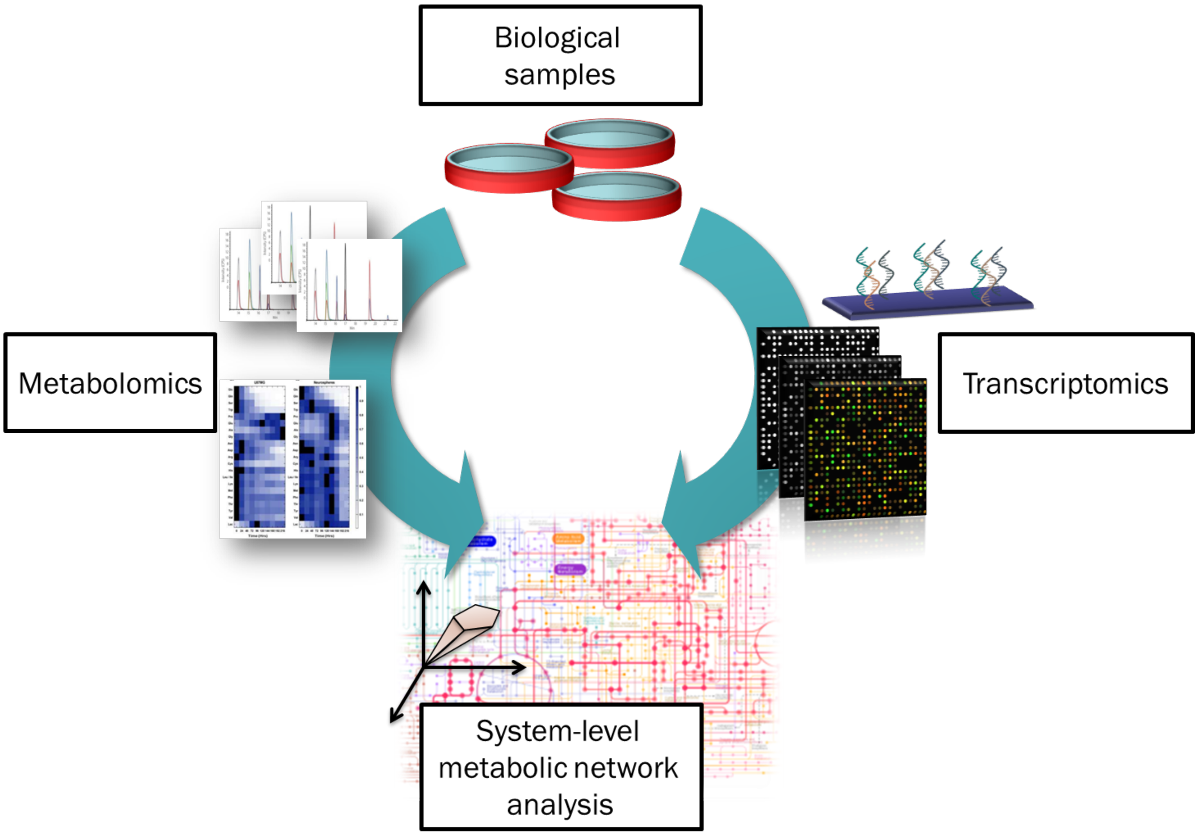
3.4 Conclusion

Advancements in high-resolution mass spectrometry have enabled high-throughput analysis by benefitting from analyte coverage over a broad mass range from individual scan.

Thus increased opportunities to explore direct mass spectrometry approaches devoid of chromatography can be observed in recent past. In order to offer at par accuracy and robustness with such chromatography-free analytical workflows consideration towards adaptation of data analysis strategies is equally important. NEST offers automation for high-throughput signal-to-noise ratio estimation for chromatography-free high-resolution mass spectrometry (HRMS) data in MS platform independent fashion. The current algorithm performs well in comparison to the estimation following various freeware software tools along with instrument specific proprietary data analysis tools. Signal-to-noise ratio forms the primary criteria in reliable analyte feature annotation and hence becomes vital for both qualitative as well as quantitative workflows. An automated tool such as NEST not only offers ease of evaluation of novel direct MS workflows but also a method consistent across various MS platforms that makes it a better alternative method for reliable feature extraction post NEST's integration into any open source data analysis tools.

Chapter 4

Interfacing HRMS data with genome-scale metabolic modeling



Integrative analysis using multi-omic data for system-level interrogation

4.1 Introduction

The need to develop tools for personalized medicine and individualized therapy is heightened especially for diseases like cancer where heterogeneity plays a big role. The metabolic coverage offered by high-resolution mass spectrometry (HRMS) based quantitative workflow offers a means to differentiate diseased phenotype from normal cells. But for development of efficient and selective treatment strategies, reprogrammed metabolic networks can be exploited. Although, HRMS based metabolic profiling can be availed to decipher metabolic reprogramming, quantitation of complete metabolome becomes increasingly complex. In order to accelerate the pace of therapeutic research, a number of computational tools have been developed. Constraints based modeling (CBM) using genome scale metabolic (GSM) network reconstructions of human metabolism have recently gained interest for formalizing potential novel targets for cancer treatment.[123,124] CBM takes into account mass balance, thermodynamic constraints and context specific 'omics' data, which is crucial for building system specific contextualized GSM model that can offer interrogation of specific targets for therapy. Amongst other omic platforms, metabolic profiling using HRMS have been exploited to build context specific models that were validated with growth or metabolic phenotypes.[114,147,148] The simplicity and flexibility of interrogating constraints based models with an ever-increasing list of data acquisition methods extends their application domain from measurement tools to tools used to discern functional or mechanistic insights of cellular metabolism for cellular engineering or individualized therapy.[149] In this chapter, constraints based metabolic models were built for glioblastoma cancer cell lines U87MG and a phenotypically different subpopulation, neurospheroidal (NSP) cell line, using HRMS based metabolic temporal profiles as context specific constraints. These models were evaluated for their phenotypic predictions. Further, flux sampling analysis predicted characteristic metabolic network reprogramming for the differential growth and maintenance needs of U87MG and NSP cancer cell lines. Since the number of reactions participating in GSM network exceeds the number of metabolites, making it an underdetermined system, the optimal solution for any system constrained using metabolic uptake/secretion profiles alone, is not unique. Additional 'omics' data platforms such as 'genomics', can also be integrated into GSM models to further reduce flux space offered by optimization process.

One purpose of developing these methods is to accurately predict context-specific intracellular metabolic flux distribution and the other is to develop tissue specific models for multi-cellular organisms and probe computational phenotypes. These tend to provide holistic as well as mechanistic understanding of metabolism and can be extended to different levels of cellular architecture using interpretations from molecular portraits.

Several algorithms have been developed for the incorporation of gene expression data into flux balance models. These include (a) Gene Inactivity Moderated by Metabolism and Expression (GIMME) algorithm[150], (b) Integrative Metabolic Analysis Tool (iMAT)[151,152], (c) Metabolic Adjustment by Differential Expression (MADE)[153], (d) E-Flux[154], and (e) Probabilistic Regulation Of Metabolism (PROM)[155]. Based on the gene expression data, thresholds are set to tightly constrain reactions in the metabolic network reconstruction. The flux cone is thus capped by changing the upper bound on a reaction based on a Boolean representation that functionally connects the reactions to the genes/transcripts. For the GIMME, iMAT, and MADE algorithm, gene expression levels in the data are reduced to binary states (by setting the upper bounds of a reaction to some large constant or zero), the E-Flux method however attempts to directly incorporate mRNA levels or transcript abundance data as maximum feasible rates of reactions in the FBA optimization problem. Although, E-Flux represents a more physiologically accurate reaction activity gradient than other parallel algorithms, the use of a direct linear relationship between transcript abundances and corresponding reaction rates, lacks a biological mechanistic basis. PROM is a method that integrates regulatory and metabolic networks. It calculates the probability of a metabolic target gene being expressed relative to the activity of its regulating transcription factor from a large dataset of gene expression data, and the flux maxima of the metabolic reaction associated with the metabolic target gene is constrained by a factor of this probability. All the algorithms are based on the assumption that mRNA transcript levels are a strong indicator for the level of protein activity.

We have developed an algorithm ScalEX to integrate gene expression data in an upgraded human metabolic reconstruction RECON1[156] and represent cancer cell lineages for 9 different tissues from NCI-60 panel (Table 4.4). ScalEX contextualize upper bounds on the reaction flux to shrink solution space with the aid of global gene expression profile parsed through a non-linear function. The non-linear function involves a scaling exponent that is contextualized and can thus define the cell type or lineage. Such application of constraints to the flux balance problem allows the optimality criterion to predict growth rates for cancer cells, determine flux distribution patterns, identify rigidity of the networks and ultimately explain the heterogeneity of all cell lineages. Comparisons of *in silico* model predictions with experimental data for growth rate of cell line models from NCI-60, were used as primary validations. These results are just representative of the wide spectrum of applications plausible with *in silico* system level models that can eventually fill a critical need for predictive models of tumor growth, proliferation and metabolic outcomes in personalized medicine.

Availability and implementation: Freely available on web at (for academic use only): http://bit.ly/ScalEXcode. ScalEX is implemented in Perl, GNU Octave and Python. It is supported on Linux and Microsoft Windows system having respective command-line interface for Perl, GNU Octave and Python installed. Additional details and processed data used for this dissertation work can be found at http://bit.ly/dissertationDataAG.

Credits towards project supervision and sample preparation: Bulk of research work described in this chapter was performed under supervision of Dr. Anu Raghunathan. Extracellular sample extracts for two neuronal cancer cell lines (U87MG and NSP) were obtained from cell culture experiments performed by Rupa as a part of collaborative study with Dr. Anu Raghunathan.

4.2 Materials and Methods

4.2.1 CBM models for cancer cell line U87MG and NSP using LC-HRMS based metabolic profiles

Sample extraction for metabolic profiling using LC-HRMS

Extracellular cell culture extracts, growth statistics and exome sequencing based genomic variants were obtained for two neuronal cancer cell lines (U87MG and NSP) that were characterized in a parallel published study.[132] Cell line for U87MG (HTB-14; Human Glioblastoma Multiforme from ATCC; IC50$_{TMZ}$: 745.6 μM) and its phenotypically different subpopulation of neurospheroidal (NSP) cell line (IC50$_{TMZ}$: 1039 μM) were treated with 10 μM dosage of alkylating agent temozolamide (TMZ). Growths of these cell lines were profiled via cell count over a period of 216 hours (9 days). In case of genomic variants based on exome sequencing profile for both cell lines, functional characterization was achieved using web tool Oncotator[157] (http://portals.broadinstitute.org/oncotator/), for annotation of mutations into synonymous or non-synonymous category. For LC-HRMS based temporal quantitative profiling of extracellular metabolites, samples were harvested every 24 hours over a period of seven days. A 100 μL of sample extract was mixed with 400 μL of chilled methanol (previously stored in -80°C). The solution was thoroughly mixed for 2 min followed by centrifugation for 15 min at 5000 rpm (4°C).

The tubes were carefully removed, 300 μL of supernatant was withdrawn and transferred into a fresh tube. A two-step serial dilution of supernatant was performed using 50% acetonitrile in water. In the first step, 50 μL of supernatant was thoroughly mixed with 450 μL of diluent. This solution was further diluted by mixing 100 μL of sample solution with 400 μL of diluent. 10 μL of sample solution for each time point was pooled for meta-analysis post LC-HRMS metabolic profiling, with the help of multivariate statistical tools. All the solutions were thoroughly mixed before analysis using LC-HRMS system.

LC-HRMS based metabolic profiling

The LC-HRMS instrumentation consisted of autosampler (Accela Open Autosampler, Thermo Scientific) and liquid chromatograph (Accela 1250, Thermo Scientific) in tandem with the Q-Exactive (Thermo Scientific) high resolution mass spectrometer equipped with a heated electrospray ionization (HESI) interface. Instrument operation and data acquisition was performed using the Xcalibur™platform software (Thermo Scientific). A C18 Hypersil gold column (10 cm x 2.1 mm x 3.0 μM) by Thermo Scientific was used for eluting the samples prior to the ESI. The mass analyzer was operated in positive ion mode and data was acquired in triplicates within a mass range of 60-900 *m/z* at 70,000 FWHM resolution. For quantitative analysis, mammalian cell culture medium standard mixture (composed of DMEM - Sigma-Aldrich, D6046, supplemented with MEM Non-essential amino acids solution - Sigma-Aldrich, M7145) was serially diluted to generate calibration curves for the ranges reported in Table 2.2. A total of 10 calibration levels and 2 quality control (QC) samples were used. These samples and standard mix along with extracellular sample extracts were uniformly spiked with the 2 μM solution of verapamil as internal standard to evaluate the performance and for data normalization. The solutions were thoroughly mixed and were then analyzed on LC-HRMS system.

Following the acquisition, the raw data was analyzed using proprietary 'Quan-browser' module from Xcalibur™for quantitative estimations.

Generation of core cancer metabolic model for U87MG and NSP

A published model for central core metabolism[148] consisting of 382 reactions, that are highly conserved in cancer, was used to build U87MG and NSP specific models. The model consisted of reactions involved in metabolic functions such as, biomass precursor synthesis, core energy metabolism, co-factor transfer and regeneration reactions, and relevant pathways for high secretion/uptake metabolites etc. Neuronal cell specific biomass composition was determined for U87MG and NSP to define biomass macromolecular composition as conversion of individual metabolite precursors into biomass, as illustrated for neuronal cancer cell line in previously published article from Prof. Palsson's research group.[148] This biomass reaction was also constrained using experimental growth rates of 0.021 hr^{-1} and 0.0096 hr^{-1} for U87MG and NSP, respectively. Systemic effects of

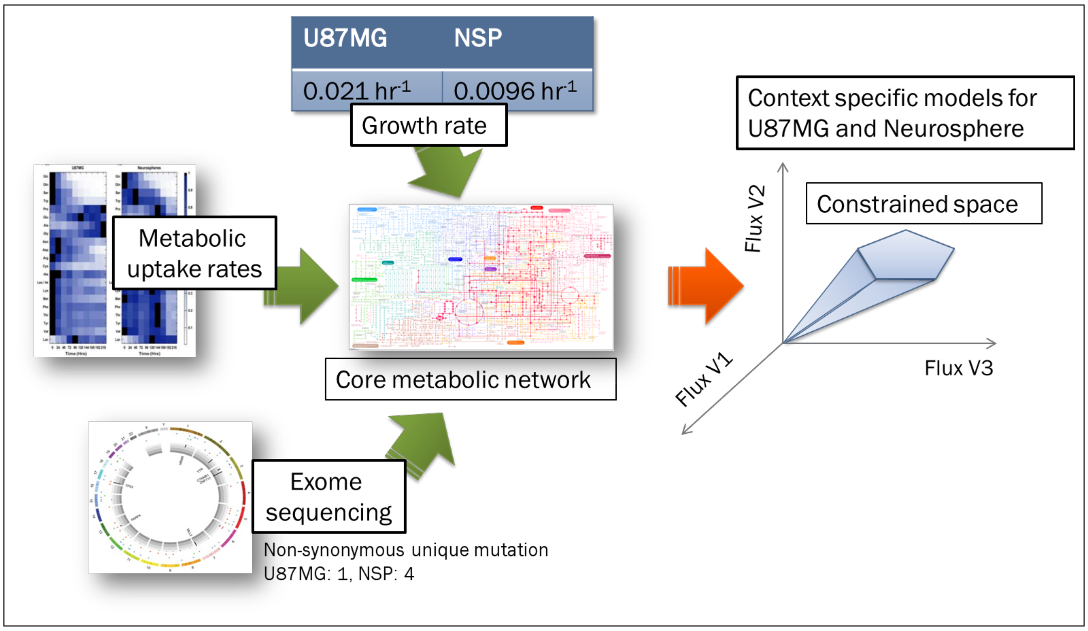


Figure 4.1 Illustrative workflow of building context specific metabolic models for U87MG and NSP

genomic variants on cellular metabolism are well known.[158,159] In order to incorporate enzymopathic effects of system specific unique mutations for these two cell lines, a list of reactions were identified (as shown in Table 4.1), following gene-protein-relation having genes with mutations. Such intracellular reaction's flux bounds was constrained following Equation 4.1.[160]

$$newV_{i,max} = v_{i,min} + \frac{v_{i,min} - v_{i,max}}{4} \tag{4.1}$$

In Equation 4.1, $v_{i,min}$ and $v_{i,max}$ represent feasible flux range in each reaction from an unaltered model system, identified using flux variability analysis (FVA). Models for both these cell lines were further contextualized using metabolic footprinting data from LC-HRMS based analysis. By making use of the quantitative metabolic estimations for a list of 21 metabolites (see Figure 4.2), rates of nutrient uptake/release were calculated by plotting the concentrations (μM/gDCW) over time. The slope of the curve was used to calculate the maximum flux through respective exchange reactions as showcased in a published research study.[148] Figure 4.1 illustrates schematic representation of contextualized model developed using various phenotypic information, discussed above for U87MG and NSP cancer cell lines. These contextualized models were evaluated following their *in-silico* phenotypic predictions using constraints based methods such as, flux balance analysis and uniform random flux sampling of flux solution space.

4.2.2 Genome scale metabolic models contextualized using gene expression constraints

The development of gene expression integrated ScalEX model of human metabolism primarily requires (i) the human metabolic network reconstruction (ii) a flux balance model with environmental conditions and (iii) gene expression data sets for a particular

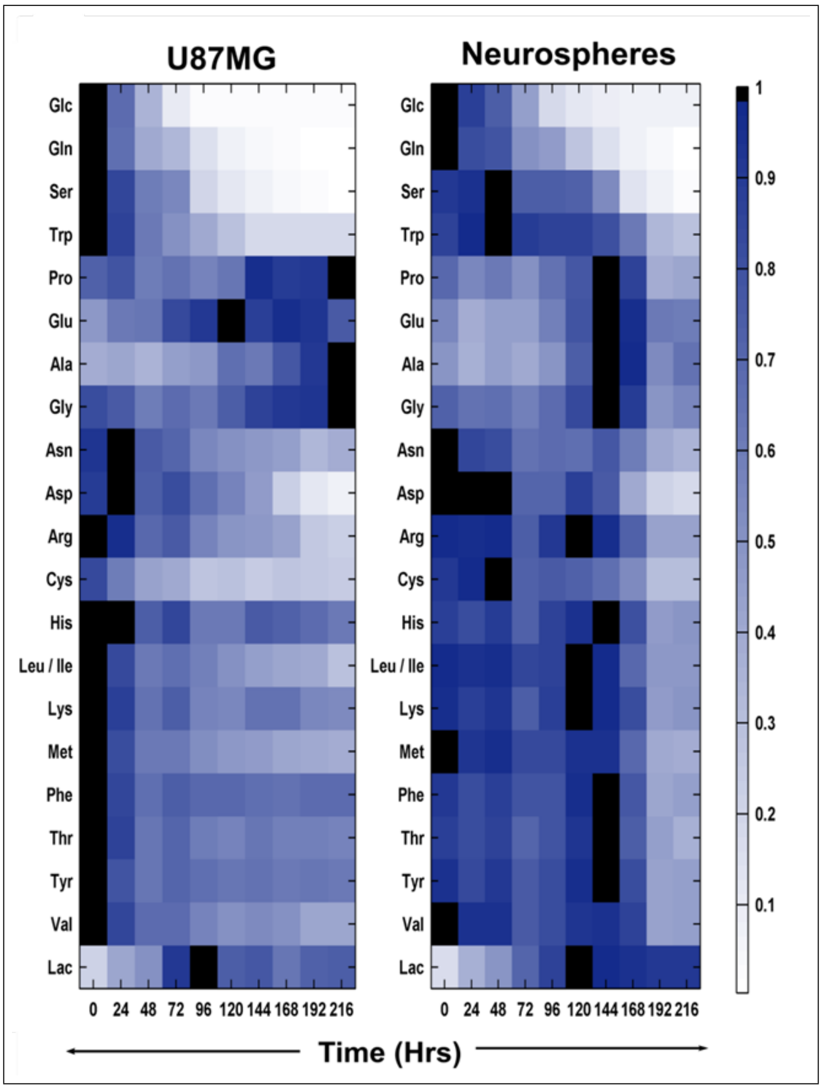


Figure 4.2 Exchange profiles of U87MG and NSP cell lines for list of metabolites. To illustrate differential uptake profiles, concentration profiles are max-normalized across each row. Analyte names were specified as three letter code of amino acids. Glc and Lac represents profiles for glucose and lactate, respectively.

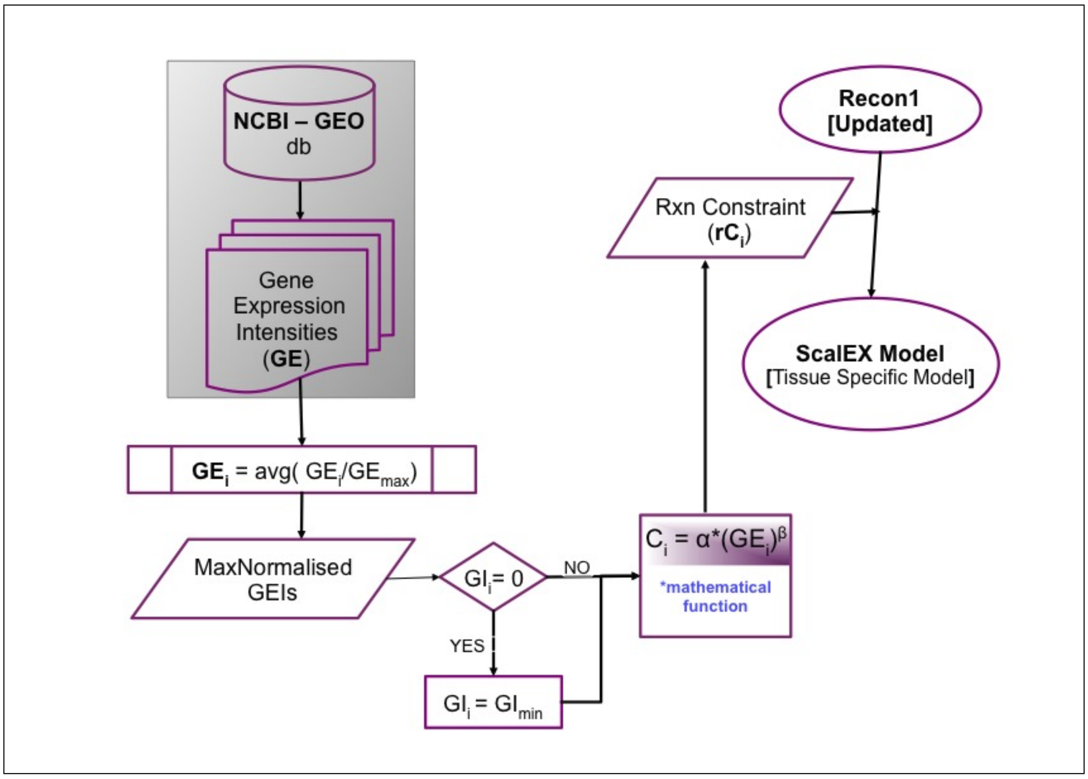


Figure 4.3 Schematic pipeline for ScalEX algorithm. GE_{max} defines maximum gene expression intensity while GI_{min} is minimum gene expression intensity from gene expression data. α and β in mathematical transformation function defines scaling constant and exponent, respectively. Following sections describes α and β in detail.

Table 4.1 List of genes with unique mutations leading to enzymopathies for associated reactions

GeneSymbol	RxnID	RxnName	RxnFormula	Cell line	FVA min. flux	FVA max. flux	Updated max. flux
AMT	GCCam	glycine-cleavage complex (lipoylprotein), mitochondrial	gly[m] + h[m] + lpro[m] <=> alpro[m] + co2[m]	U87MG	0.0413	0.2350	0.1866
AMT	GCCbim	glycine-cleavage complex (lipoylprotein) irreversible, mitochondrial	alpro[m] + thf[m] –> dhlpro[m] + mlthf[m] + nh4[m]	U87MG	0.0413	0.2350	0.1866
AMT	GCCcm	glycine-cleavage complex (lipoylprotein), mitochondrial	dhlpro[m] + nad[m] <=> h[m] + lpro[m] + nadh[m]	U87MG	0.0413	0.2350	0.1866
CYC1	CYOR_u10m	CYOR u10m	2.0 ficytC[m] + 2.0 h[m] + q10h2[m] –> 2.0 focytC[m] + 4.0 h[t] + q10[m]	NSP	0.5180	6.3929	4.9242
CYC1	CYOOm2	CYOOm2	4.0 focytC[m] + 8.0 h[m] + o2[m] –> 4.0 ficytC[m] + 2.0 h2o[m] + 4.0 h[t]	NSP	0.2590	3.1965	2.4621
ME1	ME2	malic enzyme (NADP)	mal_L[c] + nadp[c] –> co2[c] + nadph[c] + pyr[c]	NSP	0.0000	0.1624	0.1218
NSDHL	C3STDH1Pr	C-3 sterol dehydrogenase (4-methylzymosterol)	4mzym_int1[r] + nadp[r] –> 4mzym_int2[r] + co2[r] + h[r] + nadph[r]	NSP	0.0005	0.0005	0.0005
NSDHL	C4STMO2Pr	C-4 methyl sterol oxidase	4mzym_int2[r] + nadp[r] + o2[r] –> co2[r] + h[r] + nadph[r] + zym_int2[r]	NSP	0.0005	0.0005	0.0005
SLC38A4	PHEt4	L-phenylalanine transport in via sodium symport	na1[e] + phe_L[e] –> na1[c] + phe_L[c]	NSP	0.0032	0.0035	0.0034
SLC38A4	SERt4	L-serine via sodium symport	na1[e] + ser_L[e] <=> na1[c] + ser_L[c]	NSP	0.0259	0.0286	0.0279
SLC38A4	GLNt4	L-glutamine reversible transport via sodium symport	gln_L[e] + na1[e] –> gln_L[c] + na1[c]	NSP	0.1492	0.1649	0.1609
SLC38A4	LEUt4	L-leucine transport in via sodium symport	leu_L[e] + na1[e] –> leu_L[c] + na1[c]	NSP	0.0157	0.0173	0.0169
SLC38A4	ASNt4	L-asparagine transport in via sodium symport	asn_L[e] + na1[e] <=> asn_L[c] + na1[c]	NSP	0.0011	0.0012	0.0012

cell type/lineage. A pipeline for ScalEX algorithm implemented using Perl scripting language, is represented in Figure 4.3. Any additional information regarding exhibited phenotypes and objectives that are biologically or clinically relevant, are useful to validate the model.

Gene expression data processing

The gene expression data sets accessed from NCBI-GEO databases are processed to rescale the gene expression intensities. The gene expression intensities are normalized within a particular data set to the maximum absolute expression value in the data set. Such a normalization, essentially a linear transformation on the original gene expression data set, results in rescaled values in the range [0,1]. If the X_{min} and X_{max} are the minimum and maximum values, respectively, for gene expression intensity in the microarray dataset, the new values will be scaled in the range [X_{min} / X_{max} , 1]. This makes comparisons across data sets and microarray platforms feasible and allows for accurate interpretations. Considering the specificity and sensitivity of gene expression profiling methods, the interpretation of 'zero' intensity may be related to measurement sensitivity of method in contrast to the gene actually being turned off. Hence, to avoid such misinterpretations we have replaced zero intensity values by the minimum observed absolute gene expression intensity. These normalized values from the array are then scaled to the upper bound of the possible flux using a transformation function defined by ScalEX as discussed in the following subsection.

Scaling mRNA levels to reaction flux

ScalEX as the name suggests implements a scaling function that correlates the observed gene expression intensity/mRNA abundance to the maximum velocity of the enzyme catalyzed reaction *i.e.* maximum flux (upper bound v_j) that the catalyzed reaction can carry. The maximum flux, $V_{i,max}$ is scaled to the mRNA transcript/gene expression data using an exponential scaling function (as shown in Equation 4.2), with a scaling constant 'α' and a scaling exponent 'β'.

$$V_{i,max} = \alpha(X_i)^{\beta} \tag{4.2}$$

In Equation 4.2, X_i is the gene expression intensity of gene i, which codes for enzyme that catalyzes reaction j.

Calculation of scaling parameters

The scaling exponent 'β', is a conditional metabolic-transcript fraction (quite akin to a mole fraction of chemical species). It defines the theoretical metabolic expression potential of any cell based on its gene expression under given micro-environmental conditions. Thus, the exponent β is defined as the ratio of the so-called total metabolic gene expression to the total gene expression of cells.

$$\beta = \frac{\sum_{i=1}^{n} Xmet_i}{\sum_{i=1}^{m} X_i} \tag{4.3}$$

Xmet = gene expression of metabolic function

The exponent β thus reflects a non-linear regulation transposed on gene expression, on account of post transcriptional and post translations effects and dictate the tacit relation between mRNA and Vmax.

Empirical value is set for 'α' as 10, defining the two orders of magnitude difference from the scaling exponent. The scaling constant α, can be considered as reflective of k_{cat} values of the enzyme, which is known to have median value of 10 across prokaryotes and eukaryotes organisms.[161].

Apart from the non-linear relationship represented by Equation 4.2, between mRNA abundance level and Vmax, empirical studies for alternative linear relationship function (instead of exponential, multiplication of β with gene expression) and different values for α ranging over orders of magnitude, were also carried out. It was observed that the transformation function represented by Equation 4.2 with $\alpha = 10$ shown better performance.

Using Boolean expressions of gene protein reaction relation to define flux

Boolean rules represent the gene protein reaction relation (GPR), that essentially specify gene combinations necessary and/or sufficient for a protein catalyzed reaction to be functional or to carry flux in a cell. The simplest case of a one to one relation between gene and reaction as for phosphoglucose isomerase PGI, which can be reflected as,

"2821.1 $\Rightarrow$ PGI"
PGI $\equiv$ PGI

GPRs are Boolean logic expressions that can be written using standard operators AND and OR. Thus for multi-gene proteins, protein complexes and isozymes, GPR relationships are complex and are utilized for scaling mRNA abundances accordingly for catalyzed reaction rates. Intuitively, the AND operator will be limited by the lowest expressing gene and hence reduces the capping vector to the minimal expression value amongst the genes required for function, while the OR operator increases the capping vector to include activity of all isozymes. The GPR for succinate dehydrogenase in human indicates 4 subunits coded by transcripts *6389.1* (SdhA), *6392.1* (SdhB), *6391.1* (SdhC) and *6390.1* (SdhD) to form the functional protein SDH. So, the rule would be

"SdhA and SdhB and SdhC and SdhD ⇒ SDH"

"*6389.1* AND *6392.1* AND *6391.1* AND *6390.1* ⇒ SDH"

The gene expression of SDH would be proportional to that of the subunit lowest in abundance.

$X_{SDH} \equiv MIN\{X_{SdhA}, X_{SdhB}, X_{SdhC}, X_{SdhD}\}$

The AND operation between transcripts would thus be represented by a minimization filter.

When either gene can independently decide activity, like isozymes, the OR operation is used as in Glyceraldehyde 3 phosphate dehydrogenase.

"*2597.1* or *26330.1* ⇒ GAPDH"

$X_{GAPDH} \equiv X_{GAPDH1} \bigcup X_{GAPDH2}$

The OR operation is thus implemented as a union filter.

All multi operation expressions defining complex GPRs follow a combination of rules, e.g., "(*3030.1* AND *3032.1*) OR *38.1* ⇒ ACACT1"

4.2.3 Constraints based approaches for interrogation of metabolic reconstruction models

Flux balance analysis (FBA)

FBA is a modeling formalism based on stoichiometry and linear optimization that computes capabilities of metabolic networks. Material balance can be written around a system comprising such networks. The consequence of the quasi steady-state assumption (due to metabolic transients being rapid as compared to cell growth or environmental changes) is that all metabolic fluxes that lead to the formation or degradation of a metabolite must be balanced, leading to the flux balance equation

$$\frac{dX}{dt} = S \cdot v = 0 \tag{4.4}$$

wherein S is a $m \times n$ stoichiometric matrix of the reactions, m is the number of the metabolites, n is the number of fluxes, and v is the flux vector of the network. The elements in S matrix correspond to the stoichiometric coefficients of the reactions. Equation 4.4 is typically an under determined system of linear equations (more unknown fluxes than metabolites) and has innumerable possible solutions. When a biologically relevant objective function is used, only the solution that gives the maximum or minimum value is relevant and is obtained using the following linear program,

$$max\{v_{obj}\}$$

subject to,

$$S \cdot v = 0 \tag{4.5}$$

$$v_i^{min} \leq v_i \leq v_i^{max} \tag{4.6}$$

In addition to the mass balance constraints defined by Equation 4.5, upper bound (v_i^{max}) and lower bounds, (v_i^{min}) are imposed through Equation 4.6, to enforce thermodynamic reversibility and certain cell-environment characteristics (uptake/secretion rates).
Based on ScalEX, additional constraints are imposed via these bounds on intracellular maximum reaction rates calculated through the mRNA abundance data. The maximum

reaction rates calculated using Equation 4.2 are applied to the model based on the reaction thermodynamics for reversible (Equation 4.7) and irreversible (Equation 4.8) reactions.

$$-\alpha(X_i)^\beta \leq v_i \leq \alpha(X_i)^\beta \tag{4.7}$$

$$0 \leq v_i \leq \alpha(X_i)^\beta \tag{4.8}$$

The constraints file thus generated using ScalEX can be fed to an updated Human Recon1 FBA model. Such FBA model for different cancer cell lineages was implemented in MATLAB R2012b (The MathWorks Inc., Natick, MA, USA). The linear program was solved with the Tomlab (Tomlab Optimization Inc., Seattle, WA) CPLEX linear programming solver.

Flux Variability Analysis (FVA)

Plurality of solutions exists for the FBA problem, since the cell can choose multiple flux distributions to result in a unique objective function. FVA identifies the set of feasible fluxes at the optimal objective. The method calculates the minimum and maximum allowable fluxes through each reaction using a double optimization linear programming approach for each reaction of interest. (Equations 4.5, 4.6).

The FVA problem, an extension of the FBA, is set up as

$$max_v\{v_i\}/min_v\{v_i\}$$

subject to,

$$S \cdot v = 0$$

$$v_{obj} \geq \gamma Z_0$$
$$v_i^{min} \leq v_i \leq v_i^{max}$$

where v_{obj} is an optimal solution for (Equation 4.4). γ is a control parameter to define the problem with respect to the default optimal state ($\gamma = 1$) or alternate sub-optimal network states ($0 \leq \gamma < 1$).

The non-uniqueness of the FBA solution allows calculation of a range of flux that is feasible for each reaction, thus defining the rigidity and plasticity of the network.

Uniform random sampling of reaction flux

Similar to FVA, properties of metabolic flux states can be deciphered by random sampling of feasible flux space within the enclosing parallelepiped solution space.[160] This can be achieved by choosing a random point uniformly along each edge of parallelepiped following Monte Carlo sampling. Equation 4.9 illustrates how random points are chosen within the solution space.

$$\alpha_i = \alpha_{i,min} + Rn(\alpha_{i,max} - \alpha_{i,min}) \tag{4.9}$$

In Equation 4.9, Rn is a random number chosen between 0 and 1 while $\alpha_{i,max}$ and $\alpha_{i,min}$ defines the flux range of feasible flux state along each reaction vector identified using FVA. These points can then be further compared to the set of constraints imposed on a constrained based metabolic model, in order to verify whether the random point falls in solution space.

Solution sampling in this manner not only offers insights about plasticity of metabolic network but offers latent information about metabolic flux states such as, co-regulated list of trans-acting metabolic reactions. Additionally, information about rewiring of metabolic network imposed by system specific constraints can also be elucidated by

inspecting the population distribution of random sampling for each reaction.[160]
Flux sampling for U87MG and NSP cell line models was carried out using a Markov Chain Monte Carlo method of Artificial centering Hit-and-Run (ACHR) Sampler from COBRA toolbox. Faster mixing and better coverage for irregularly shaped solution space are the attributes that makes ACHR smapler a better choice over other available sampling methods. The initial point for the sampler was chosen amongst 1000 warmup point identified by combining random and orthogonal point. A total of 50000 randomly distributed sampling points were computed with 1000 iterations between each stored point. Distribution of individual reaction flux values across the sampling population was represented as a histogram of feasible flux value and associated frequency in the convex polytope of solution space. Comparison of such flux distributions across both models for U87MG and NSP enabled shortlisting of possible metabolic network rewiring pertaining to these cancer cell models.

4.2.4 Meta-analysis of differential metabolic LC-HRMS profile

For extraction of global metabolic features from LC-HRMS data, publicly available tools, MZmine2 (http://mzmine.github.io/)[42] and XCMSonline (https://xcmsonline.scripps.edu)[41] were utilized as described below. For this analysis, metabolic profiles for time pooled samples of U87MG and NSP cell lines were utilized.

Untargeted systems-level metabolic profiling using MZmine2

For analyte peak feature extraction using MZmine2 (ver 2.30) from LC-HRMS data of cancer cell lines, HRMS profiles were reduced to centroid mode. Optimal peak picking parameters with intensity cut-off of 1.0e4 and m/z shift tolerance of 2.5 ppm were used to generate chromatograms with minimum peak width of 20 secs. Following Savitzky-Golay smoothing with filter width of 5, these chromatograms were subjected to deconvolution using the local minimum search algorithm. Criteria of 20% minimum relative height, minimum RT range of 6 secs and maximum peak width of 90 secs were used. Isotopic peaks and duplicate peaks were removed using tolerance over m/z of 2.5 ppm and RT of 6 secs. To correct any linear or non-linear deviation in RT, RANSAC aligner was used with over 5 iterations of alignment.

These aligned features were subjected to unsupervised clustering using principal component analysis (PCA) to identify differential metabolites across the two cancer cell lines. Annotations of these extracted features were achieved using an online database search module within MZmine. Features were queried against the human metabolome database (HMDB) with mass accuracy of 10 ppm. Annotated features were further mapped to human metabolic pathway with the help of MetaboAnalyst (http://www.metaboanalyst.ca/)[39] using KEGG reference pathway database.

Untargeted metabolic profiling using XCMSonline

Centroided LC-HRMS data was utilized for feature extraction using XCMSonline. Wavelet transformation based centWave method was used with a 5 ppm tolerance for m/z and 10 to 90 secs chromatographic peak width range. For chromatogram integration, peak limits were found using a mexican hat based filter with a S/N threshold of 6. Followed by RT correction using obiwarp method, peak density chromatograms were aligned using tolerance for RT of 5 secs and 0.015 Da width of m/z slices, as grouping criteria.

Similar to MZmine, XCMS online also offers PCA method to evaluate the extracted analyte features in a graphical manner to easily identify similar and dissimilar samples, thus highlighting the variability in the multivariate data set. For annotation of features and their pathway mapping, XCMS online makes use of Metlin database from within their web interface. A search criterion of 10 ppm mass accuracy was used.

Multivariate analysis of targeted list of metabolites using MetaboAnalyst

For exploratory statistical analysis of metabolic profiles, multivariate analysis tools were employed with the help of web interface tool, MetaboAnalyst (http://www.metaboanalyst.ca/)[39]. Metabolite concentration estimates of time pooled samples for a list of metabolites shown in Figure 4.2, were used for this analysis. To address the skew in concentration data and the large range of different metabolites, standard normal variate correction was applied. Processed data was analyzed using chemometric tools such as, PCA and partial least square discriminant analysis (PLS-DA) from statistical analysis section of MetaboAnalyst interface.

4.3 Results and Discussion

4.3.1 Differential metabolic phenotype for cancer cell line U87MG and NSP

By virtue of quantitative accuracy and analyte coverage offered by HRMS, its application in metabolic profiling of clinical and biological samples has been widespread.[14,26,88,99,132] Extensive list of tools offering statistical interpretations, functional mapping and methods for variable ranking are available for comprehensive data analysis (refer Table 1.1 for few open-source tools). Such tools, either make use of targeted list of metabolic quantitative profiles or extract metabolic features in an untargeted manner. We have made use of two such popular open-source data analysis tools, MZmine[42] and XCMSonline[41] for untargeted metabolic feature extraction and statistical interpretation. Additionally, quantitation of selected metabolites (shown in Figure 4.2) allowed delineation of differential profiles across the two cancer cell lines, U87MG and NSP.

Statistical clustering of cancer cell lines using targeted metabolic profiles

The temporal quantitative estimation of a selected set of metabolites (Figure 4.2) were used to estimate cell specific exchange rate constraints for metabolic network models. In order to evaluate the metabolites in this set that had differential significance, we made use of supervised clustering algorithm of PLS-DA. Figure 4.4a illustrates the projection score plot using PLS-DA model with distinct clustering of samples for U87MG and NSP. It is noteworthy, that PLS-DA models are developed with emphasis on increasing covariance across input variables (metabolite concentrations) and sample group (U87MG and NSP). Hence, the weighted sum of squared correlations across partial least square (PLS) components and input variables, rank metabolites according to their importance in discriminating the sample groups. The variable importance on projection (VIP) scores of the input variables for the PLS-DA model are shown in Figure 4.4b. Apart from glucose and lactate, VIP scores for arginine and malic acid were found to be significant indicating the potential impact of these metabolites' differential consumption profiles to heterogeneity in the two cancer cell lines.

Untargeted metabolic profiling of U87MG and NSP using LC-HRMS data

Functional characterization of both cancer cell lines was obtained following untargeted metabolic profiling of LC-HRMS data. To ascertain extraction of maximum feasible list of features, MZmine2 and XCMSonline were utilized. Although, number of features extracted by both methods were significantly different, with XCMSonline identifying 416 features while MZmine2 extracting 3309 features, Figure 4.5 shows distinct clusters of both cancer cell lines, even with application of unsupervised clustering method of PCA. Following putative annotation of extracted features using public online databases, pathway mapping was carried out to identify likely affected list of pathways. Table 4.2 illustrates mapped pathways for putatively annotated metabolic features having differen-

Table 4.2 Functional analysis of untargeted features extracted for U87MG and NSP

Pathway	Overlapping metabolites found	p-value
Using MZmine2		
Tyrosine metabolism	2	0.014
Selenoamino acid metabolism	1	0.054
Cysteine and methionine metabolism	1	0.13
Purine metabolism	1	0.21
Using XCMSonline		
Ketolysis	2	0.1
Nicotine degradation III	2	0.1
tRNA charging	3	0.17
4-aminobutyrate degradation	3	0.17
Nicotine degradation IV	4	0.24
Urea cycle	5	0.32

tial metabolite levels across the two cell lines. It is noteworthy that irrespective of the analysis method, the predominant pathways involved reactions for cofactor and nucleotide metabolism. These results corroborate with differential growth and co-factor demands across these two cell lines (as shown in Table 4.3). To investigate further mechanistic network reprogramming across the two cell lines, we have developed constraint based context specific metabolic network models for both cancer cell lines. The results of the validation against experimental data sets are discussed in the following sections.

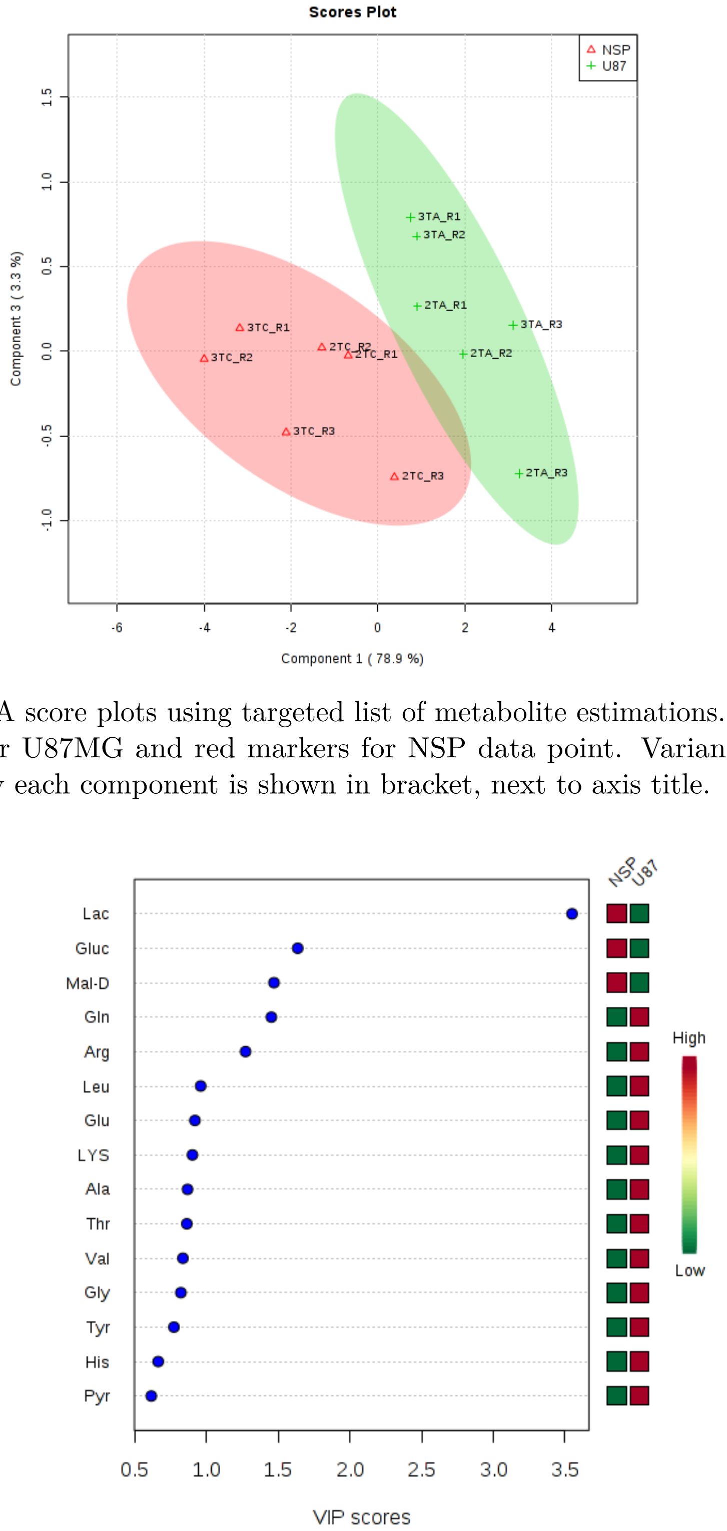
Table 4.3 Phenotypic predictions of co-factor demands for U87MG and NSP cell line model

	U87MG model	**NSP model**
Experimental observations (hr^{-1})		
Growth	0.021	0.0096
Model predictions (mM $gDCW^{-1}$ hr^{-1})		
NADH	33.1639	26.4382
NADPH	31.8479	25.3376
ATP	100.7259	83.1454

4.3.2 Core metabolic models for cancer cell line U87MG and NSP

Beyond the genetic heterogeneity, perceived as common characteristic in most cancer cells transformation, metabolic reprogramming has also been proposed as a cancer hallmark.[162] Typically metabolic reprogramming manifests as altered uptake of nutrients and their eventual metabolic fate. A systematic investigation of metabolic pathway that regulates these differential traits can be accomplished using constraints based metabolic model analysis. With application of various phenotypic constraints, as illustrated in Figure 4.1, context specific models developed for U87MG and NSP cell lines were used to compute phenotype related to cellular metabolism.

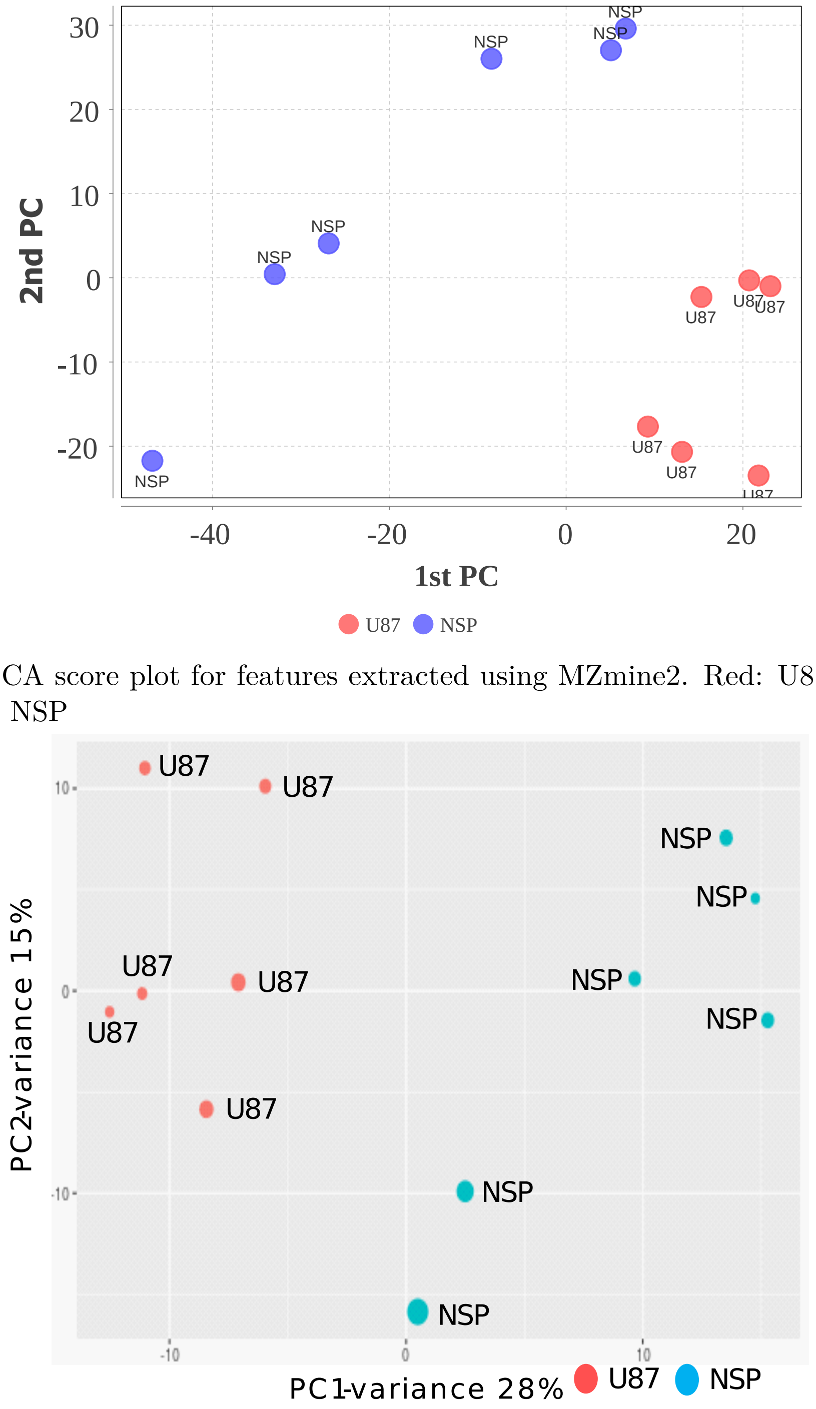
Typically an excess of cofactors production, such as NADH, NADPH and ATP over growth and energy demands are known to be related to metabolic flexibility.[148] The *in silico* representation of U87MG and NSP showed variable demand of co-factors (NADH, NADPH and ATP). As shown in Table 4.3, models predicted an excess of co-factor production for U87MG vis-à-vis NSP, consistent with their differential growth rates.



(a) PLS-DA score plots using targeted list of metabolite estimations. Green markers for U87MG and red markers for NSP data point. Variance contributed by each component is shown in bracket, next to axis title.

(b) VIP scores of PLS-DA model for input variables. Heatmap shows concentration levels of these metabolites.

Figure 4.4 Supervised clustering of metabolic profiles using PLS-DA method



(a) PCA score plot for features extracted using MZmine2. Red: U87MG, Blue: NSP

(b) PCA score plot for features extracted using XCMSonline. Red: U87MG, Cyan: NSP

Figure 4.5 Unsupervised clustering using PCA for untargeted features extracted using MZmine2 and XCMSonline

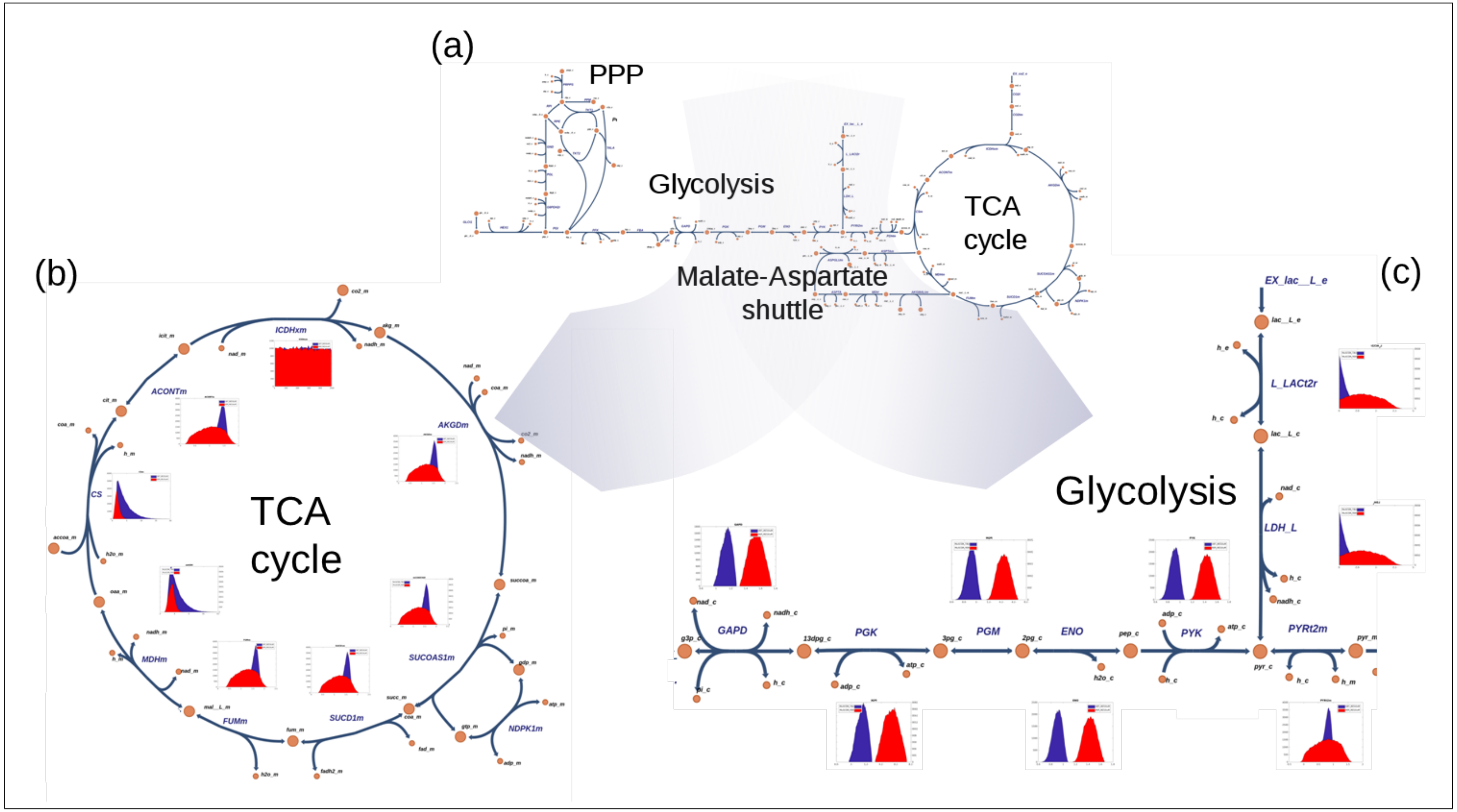


Figure 4.6 Differential metabolic rewiring across U87MG (represented by blue colored histogram) and NSP (represented by red colored histogram) cell line models for glucose catabolic pathway observed using uniform random flux sampling. (a) Overview of glucose metabolism consisting of Glycolysis, Pentose phosphate pathway and TCA cycle. Illustrations of differential metabolic flux density distribution observed across sampling population, in reactions from (b) TCA cycle & (c) Glycolysis.

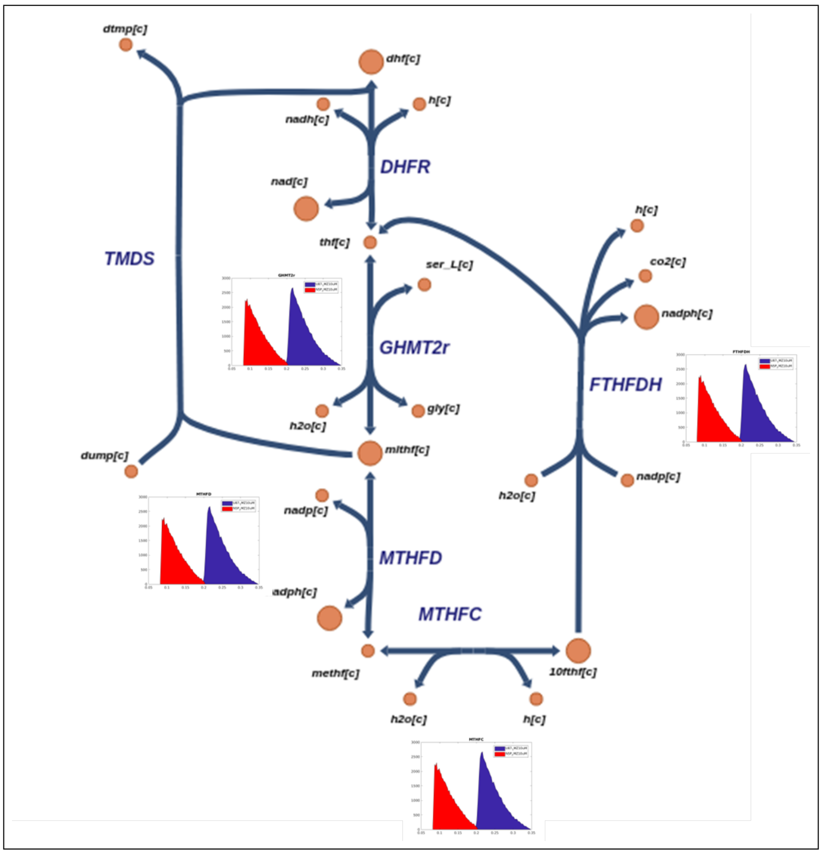


Figure 4.7 Differential metabolic rewiring in folate metabolism across U87MG and NSP cell line models observed using uniform random flux sampling

differentiate phenotypes across the two cell types. This highlights the need for addressing the mechanism at the whole genome scale level to drive any further therapy regimes.

4.3.3 Context specific genome scale metabolic model using ScalEX

Although the core model was helpful in formalizing context specific models, a genome scale model would help address redundancy and mechanistic basis of a rewired metabolism. An effort to reduce this redundancy by constraining intracellular reactions was carried out by developing an in house algorithm ScalEX. ScalEX uses gene expression data as input and generates list of constraints for all reactions catalyzed by enzymes linked through GPR for the model.

Table 4.4 Spearman rank correlations achieved by ScalEX in predicting growth rates for individual tissue specific cell-line models from NCI-60

Tissue type	No. of cell lines	Calculated β value	Correlation coefficient	Significance of correlation (p-value)
Breast	6	0.09735	0.429	4.19E-01
CNS	6	0.09706	0.886	3.33E-02
Colon	7	0.09823	0.964	2.78E-03
Lung	9	0.09753	0.733	3.11E-02
Leukemia	6	0.09703	0.543	2.97E-01
Melanoma	9	0.09804	0.786	2.79E-02
Ovarian	7	0.09714	0.771	1.03E-01
Prostrate	2	0.09756	1.000	1.00E+00
Renal	8	0.09729	0.714	5.76E-02

Note: Higher p-value for correlation estimate in case of models built for Prostrate tissue is on account of only 2 cell line data availability under NCI-60 panel.

Evaluation of model predictions for cell-specific models with different lineages

For establishing validations of models built using ScalEX, a list of tunor cell line's data from NCI-60 panel was utilized. An array of models for 60 tumor cell lines were built using published gene expression data (Accession ID: GSE5846) as an input for ScalEX along with additional constraints for secretion and uptake profiles for a list of 23 metabolites from published literature[148,165]. The estimated metabolic-transcript fraction was found to vary within a small range for the cancer cell lines (Table 4.4). However, with application of this metabolic-transcript fraction as scaling exponent (β) along with scaling constant (α) from Equation 4.2, transformation of individual metabolic gene expression values into reaction flux constraints was achieved in context specific manner. The fact that this panel is constituted by cell lines from a list of 9 tissue types (as illustrated in Table 4.4),

Table 4.5 Spearman rank correlations in predicting growth rates for all tumor cell-line specific models from NCI-60.

Method	**Correlation coefficient**	**Significance of correlation (P value)**
ScalEX	0.444	4.24E-04
ScalEX + Exch	0.783	2.48E-13
iMAT	-0.07	0.59
Eflux	0.43-0.44	3.6E-04
PRIME	0.69	1.2E-09

Note: List of different algorithms employed for generating context specific models are listed in first column. ‘ScalEX’ represents constraints identified using ScalEX algorithm, whereas ‘ScalEX + Exch’ defines models having constraints from both ScalEX and metabolite exchange rates identified experimentally

having significant differences in phenotypic and metabolic architecture, offers robust evaluation of model performance. Besides, NCI-60 tumor cell lines have been used as model cell lines for cancer studies, which provides plethora of phenotypic information for validations and also in improving model’s behavior.

Here, correlations comparing growth rate predictions from models against experimental values were used as a measure of evaluation. Table 4.5 shows Spearman’s rank correlations (with p-value of significance $<$ 1E-04) comparing growth rates for experimental and model predictions. The results showed improved performance of models constrained using ScalEX and experimental exchange rates, in comparison with other published tools such as, iMAT, Eflux and PRIME, which also utilizes gene expression data to constrain GSM models for flux balance analysis (Table 4.5).

Further investigation of model’s performance for individual tissue types revealed that models of cell line for tissue types, Breast and Leukemia shows poor correlations for *in silico* growth rate predictions in comparison to experimental observations (Table 4.4). As mentioned before, primary assumption for FBA of quasi steady state reckons that the metabolic interplay events in comparison to cellular growth or environmental conditions, are short lived and hence can be considered to be at steady state. But for cancer cells the transient genetic regulations, dictated by various factors including environmental inputs and/or stress, may not be uniform with diversity of cell population within a tumor. And this may affect the model's performance, built using gene expression data as input. Nevertheless, cell line models for colon tumour samples with 7 candidate gene expression dataset showed significant correlations of growth rate predictions (Spearman R=0.964, p-value=2.78E-03), as shown in Table 4.4.

4.4 Conclusion

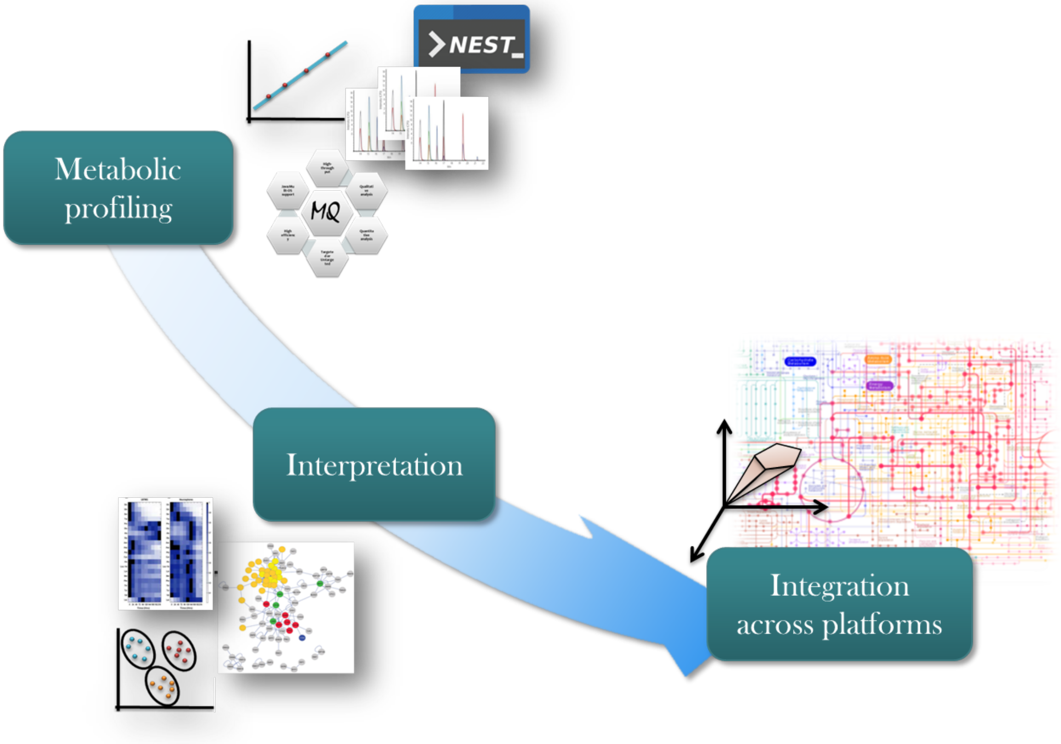
Advances in HRMS technologies have enabled overall throughput of metabolomic analysis. With simultaneous Qual/Quan capabilities, scope of metabolic profiling based applications has widened from basic biochemical investigations to clinical diagnostic analysis. The amount of data generated demands advanced computational strategies to capitalize on these merits. Although targeted and untargeted metabolic features can be analyzed using functional analysis tools and help identify differentiating features, but a mechanistic biological basis cannot be delineated using such tools. Here we have showcased application of constraints-based modeling (CBM) with use of metabolic reconstruction network constrained using such metabolic profiles. Context specific models were developed using metabolic exchange profiles for in house cancer cell lines U87MG and its phenotypically variant sub-population, NSP. Metabolic reprogramming, associated with differential phenotypic behavior in presence of drug, was delineated using model predictions.

Expansion of poly-omic data available for various organisms has led to continuous expansion of reconstruction models to genome-scale levels that are biochemically, genetically and genomically (BiGG) structured.[156] Such multi-omic nature for metabolic reconstructs can be availed for development of more context specific models using corresponding 'omic' data platforms. A similar attempt was showcased using constraints derived from metabolic exchange rates along with ScalEX based constraints for intracellular reaction using gene expression data.

All of these exercises, with their successful phenotypic validations, illustrate the growing scope of applications for hyphenated methodologies such as, CBM in conjunction with high-throughput multi-omic platforms. Similar efforts might help us get closer to the ambitious goal of personalized medicine with more precise and effective method treatments specific to each individual.

Chapter 5

Conclusion



Comprehensive workflow for metabolomic studies using HRMS

Growing list of applications with use of HRMS in various scientific fields can be attributed to its suitability for both targeted and untargeted analysis. Features like high sensitivity, mass accuracy and dynamic range in full-scan acquisition mode enables a variety of tasks, such as pre- and post-target analysis along with retrospective analysis. These set of features aptly gear HRMS workflows for researchers pursuing metabolomic analysis with emphasis on discovery of metabolite transformation products or untargeted analysis. Moreover advancement in HRMS technologies and improved features has also instigated development of direct mass spectrometry (MS) based analysis platform beyond routine chromatography studies.

In these various research fronts, the development of robust data analysis tools and methods supporting cross platform analysis becomes relevant. Work described in the dissertation has attempted to address this challenge and capitalize on the advantages that HRMS offers by developing/showcasing data analysis tools at different levels of hierarchy. Developed tools, such as MQ/NEST (validated by proprietary/freeware software tools) or hyphenated applications using CBM in conjunction with HRMS data (with experimental validations) not only establish the robustness merits for HRMS data but also illustrate broader perspective of applications made feasible using HRMS workflows.

It is noteworthy that there are few following aspects having greater importance in data processing and analysis using HRMS data, which still needs to be addressed.

Effective parameter optimization strategy

Although various data analysis tools exists, such as XCMS[40] and MZMine[42], that offers an array of algorithms for individual data pre-processing step, the effective coverage of features extracted largely depends on parameters for these algorithms defined by user. With diversity of analytical methods, developed for analytes with divergent chemistry or by availing different analytical platforms with chromatography or non-chromatography front-end considerations, demand for specific data handling methods with processing parameters tuned for analytical method in question is pertinent. Further, the understanding of mathematical complexity of these data processing methods such as, wavelet transformation filter or mexican hat filter, necessitates a data analysis expert to effectively choose such user input parameters. For a routine analytical practitioner, to make the most of these advanced data analysis tools, strategies for optimization of data processing method parameters is equally essential as much as analytical method optimization. Tools like, IPO[166] that offers automated optimization of such method parameters can be availed to ease this process. However applicability of IPO[166] is restricted to data analysis following offline version of XCMS. A systematic sensitivity analysis of different data processing methods in response to user input parameters might help formalize list of guidelines that effectively benefit in better use of such data analysis tools.

Analyte feature annotation

Accurate feature annotation of HRMS data for untargeted analysis has always been challenging. Redundancy of analyte features on virtue of different adduct ions, neutral loss, in source fragments and isotopic peaks increases the complexity of this task. Although, with the help of literature sourced information and consideration towards analyte peaks deconvolution using criteria of common retention time or isotopic peaks cluster, offers an alternative to resolve such data complexity. But more often such deconvolution strategy follows generic understanding of isotopic peaks clustering or generic list of neutral loss fragments led mass differences.[73] Essentially such data deconvolution strategies lead to excessive filtering of data and affects features annotation contrariwise. A recently published review article offers critical evaluation of various feature annotation methods available for analytical community.[167] Alternative strategies with application of Markov

Chain principle to estimate probability of accurate metabolite annotation have been discussed.[167] Similar strategies can also be extended with application of machine learning approaches such as neural networks or random forest classifiers, that have established their classification performance since long.[109] Having an accurate metabolite annotation methodology, exclusively using the full scan MS spectrum, will not only utilize throughput of HRMS but can benefit in terms of ease for integration of HRMS data with system level modeling approaches discussed in this dissertation. Any such cellular functional interpretations can further be evaluated against HRMS data in retrospective manner, employing true potential of HRMS to use.

Advancement in MS instrumentation and expanding analysis method spectrum each year, have led to increased application domain to more and more feasible fields of study. New methods capable of handling such novel platforms, extracting maximum potential using contemporary analysis methods are expected to rise. Such conglomeration of benefits offered by analysis platforms and data analysis methods holds the potential in bridging the gap between bench side analytical-academic efforts to the clinical applications for guiding effective translational efforts.

Bibliography

[1] J. Van Der Greef and A. K. Smilde, “Symbiosis of chemometrics and metabolomics: Past, present, and future,” *Journal of Chemometrics*, vol. 19, no. 5-7, pp. 376–386, 2005.

[2] S. C. Gates and C. C. Sweeley, “Quantitative metabolic profiling based on gas chromatography.,” *Clinical chemistry*, vol. 24, pp. 1663–73, oct 1978.

[3] O. Fiehn, “Metabolomics - The link between genotypes and phenotypes,” *Plant Molecular Biology*, vol. 48, no. 1-2, pp. 155–171, 2002.

[4] W. Weckwerth, “Metabolomics in systems biology.,” *Annual review of plant biology*, vol. 54, pp. 669–89, 2003.

[5] V. de Lorenzo, “From the selfish gene to selfish metabolism: revisiting the central dogma.,” *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 36, pp. 226–35, mar 2014.

[6] G. J. Patti, O. Yanes, and G. Siuzdak, “Innovation: Metabolomics: the apogee of the omics trilogy.,” *Nature reviews. Molecular cell biology*, vol. 13, pp. 263–9, apr 2012.

[7] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, “Systematic functional analysis of the yeast genome.,” *Trends in biotechnology*, vol. 16, no. 9, pp. 373–378, 1998.

[8] J. K. Nicholson, J. C. Lindon, and E. Holmes, “’Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.,” *Xenobiotica; the fate of foreign compounds in biological systems*, vol. 29, pp. 1181–9, nov 1999.

[9] G. G. Harrigan, R. H. LaPlante, G. N. Cosma, G. Cockerell, R. Goodacre, J. F. Maddox, J. P. Luyendyk, P. E. Ganey, and R. A. Roth, “Application of high-throughput Fourier-transform infrared spectroscopy in toxicology studies: contribution to a study on the development of an animal model for idiosyncratic toxicity.,” *Toxicology letters*, vol. 146, pp. 197–205, feb 2004.

[10] H. E. Johnson, D. Broadhurst, D. B. Kell, M. K. Theodorou, R. J. Merry, and G. W. Griffith, “High-Throughput Metabolic Fingerprinting of Legume Silage Fermentations via Fourier Transform Infrared Spectroscopy and Chemometrics,” *Applied and Environmental Microbiology*, vol. 70, no. 3, pp. 1583–1592, 2004.

[11] R. F. Adams, “Determination of amino acid profiles in biological samples by gas chromatography.,” *Journal of chromatography*, vol. 95, pp. 189–212, aug 1974.

[12] K. Tanaka, A. West-Dull, D. G. Hine, T. B. Lynn, and T. Lowe, “Gas-chromatographic method of analysis for urinary organic acids. II. Description of the procedure, and its application to diagnosis of patients with organic acidurias,” *Clinical Chemistry*, vol. 26, no. 13, pp. 1847–1853, 1980.

[13] E. Jellum, E. A. Kvittingen, and O. Stokke, “Mass spectrometry in diagnosis of metabolic disorders,” *Biological Mass Spectrometry*, vol. 16, no. 1-12, pp. 57–62, 1988.

[14] K. Dettmer, P. a. Aronov, and B. D. Hammock, “Mass spectrometry-based metabolomics.,” *Mass spectrometry reviews*, vol. 26, no. 1, pp. 51–78, 2007.

[15] G. A. Theodoridis, H. G. Gika, E. J. Want, and I. D. Wilson, “Liquid chromatography-mass spectrometry based global metabolite profiling: A review,” *Analytica Chimica Acta*, vol. 711, pp. 7–16, jan 2012.

[16] T. Soga, Y. Ohashi, Y. Ueno, H. Naraoka, M. Tomita, and T. Nishioka, “Quantitative Metabolome Analysis Using Capillary Electrophoresis Mass Spectrometry,” *Journal of Proteome Research*, vol. 2, pp. 488–494, oct 2003.

[17] C. Wittmann and E. Heinzle, “Application of MALDI-TOF MS to lysine-producing Corynebacterium glutamicum: a novel approach for metabolic flux analysis.,” *European Journal of Biochemistry / FEBS*, vol. 268, pp. 2441–2455, apr 2001.

[18] M. Zhou, J. F. McDonald, and F. M. Fernández, “Optimization of a direct analysis in real time/time-of-flight mass spectrometry method for rapid serum metabolomic fingerprinting.,” *Journal of the American Society for Mass Spectrometry*, vol. 21, pp. 68–75, jan 2010.

[19] A. Jackson, S. Werner, N. Talaty, Y. Song, K. Campbell, R. Cooks, and J. Morgan, “Targeted metabolomic analysis of Escherichia coli by desorption electrospray ionization and extractive electrospray ionization mass spectrometry,” *Analytical Biochemistry*, vol. 375, no. 2, pp. 272–281, 2008.

[20] S. C. Beu, M. W. Senko, J. P. Quinn, F. M. Wampler, and F. W. McLafferty, “Fourier-transform electrospray instrumentation for tandem high-resolution mass spectrometry of large molecules,” *Journal of the American Society for Mass Spectrometry*, vol. 4, no. 7, pp. 557–565, 1993.

[21] I. Gertsman, J. A. Gangoiti, and B. A. Barshop, “Validation of a dual LC-HRMS platform for clinical metabolic diagnosis in serum, bridging quantitative analysis and untargeted metabolomics.,” *Metabolomics : Official journal of the Metabolomic Society*, vol. 10, pp. 312–323, apr 2014.

[22] J. D. Pleil and K. K. Isaacs, “High-resolution mass spectrometry: basic principles for using exact mass and mass defect for discovery analysis of organic molecules in blood, breath, urine and environmental media,” *Journal of Breath Research*, vol. 10, p. 012001, mar 2016.

[23] J. Kouassi Nzoughet, C. Bocca, G. Simard, D. Prunier-Mirebeau, J. M. Chao de la Barca, D. Bonneau, V. Procaccio, F. Prunier, G. Lenaers, and P. Reynier, “A Nontargeted UHPLC-HRMS Metabolomics Pipeline for Metabolite Identification: Application to Cardiac Remote Ischemic Preconditioning,” *Analytical Chemistry*, vol. 89, pp. 2138–2146, feb 2017.

[24] H. Z. Senyuva, V. Gökmen, and E. A. Sarikaya, “Future perspectives in Orbitrap™-high-resolution mass spectrometry in food analysis: a review,” *Food Additives & Contaminants: Part A*, vol. 32, pp. 1568–1606, oct 2015.

[25] F. Hernández, J. V. Sancho, M. Ibáñez, E. Abad, T. Portolés, and L. Mattioli, “Current use of high-resolution mass spectrometry in the environmental sciences,” *Analytical and Bioanalytical Chemistry*, vol. 403, pp. 1251–1264, may 2012.

[26] S. Ríos Peces, C. Díaz Navarro, C. Márquez López, O. Caba, C. Jiménez-Luna, C. Melguizo, J. C. Prados, O. Genilloud, F. Vicente Pérez, and J. Pérez Del Palacio, “Untargeted LC-HRMS-Based Metabolomics for Searching New Biomarkers of Pancreatic Ductal Adenocarcinoma: A Pilot Study.,” *SLAS discovery : advancing life sciences R & D*, vol. 22, pp. 348–359, apr 2017.

[27] A. Singh, N. Bhattacharya, A. Ghanate, and V. Panchagnula, “Rapid and Direct Quantitation of Pharmaceutical Drugs from Urine Using MALDI-MS,” *Current Trends in Mass Spectrometry*, vol. 11, no. 1, March, pp. 24–29, 2013.

[28] K. Dettmer and B. D. Hammock, “Metabolomics - A new exciting field within the "omics" sciences,” *Environmental Health Perspectives*, vol. 112, no. 7, pp. 396–397, 2004.

[29] B. D. Bennett, J. Yuan, E. H. Kimball, and J. D. Rabinowitz, “Absolute quantitation of intracellular metabolite concentrations by an isotope ratio-based approach.,” *Nature protocols*, vol. 3, no. 8, pp. 1299–1311, 2008.

[30] A. S. Benk and C. Roesli, “Label-free quantification using MALDI mass spectrometry: Considerations and perspectives,” *Analytical and Bioanalytical Chemistry*, vol. 404, pp. 1039–1056, sep 2012.

[31] B. Rochat, E. Kottelat, and J. McMullen, “The future key role of LC–high-resolution-MS analyses in clinical laboratories: a focus on quantification,” *Bioanalysis*, vol. 4, pp. 2939–2958, dec 2012.

[32] E. J. Want, P. Masson, F. Michopoulos, I. D. Wilson, G. Theodoridis, R. S. Plumb, J. Shockcor, N. Loftus, E. Holmes, and J. K. Nicholson, “Global metabolic profiling of animal and human tissues via UPLC-MS.,” *Nature protocols*, vol. 8, pp. 17–32, jan 2013.

[33] J. Boccard, J.-L. Veuthey, and S. Rudaz, “Knowledge discovery in metabolomics: an overview of MS data handling.,” *Journal of separation science*, vol. 33, pp. 290–304, feb 2010.

[34] S. Castillo, P. Gopalacharyulu, L. Yetukuri, and M. Orešič, “Algorithms and tools for the preprocessing of LC-MS metabolomics data,” *Chemometrics and Intelligent Laboratory Systems*, vol. 108, pp. 23–32, aug 2011.

[35] K. A. Veselkov, L. K. Vingara, P. Masson, S. L. Robinette, E. Want, J. V. Li, R. H. Barton, C. Boursier-Neyret, B. Walther, T. M. Ebbels, I. I. Pelczer, E. Holmes, J. C. Lindon, and J. K. Nicholson, “Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery.,” *Analytical chemistry*, vol. 83, pp. 5864–5872, aug 2011.

[36] O. Kohlbacher, K. Reinert, C. Gröpl, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, and M. Sturm, “TOPP–the OpenMS proteomics pipeline.,” *Bioinformatics (Oxford, England)*, vol. 23, pp. e191–7, jan 2007.

[37] K. Reinert and O. Kohlbacher, “OpenMS and TOPP: open source software for LC-MS data analysis,” *Methods in molecular biology (Clifton, NJ)*, vol. 604, pp. 201–211, 2010.

[38] E. Melamud, L. Vastag, and J. D. Rabinowitz, “Metabolomic analysis and visualization engine for LC-MS data.,” *Analytical chemistry*, vol. 82, no. 23, pp. 9818–9826, 2010.

[39] J. Xia, N. Psychogios, N. Young, and D. S. Wishart, “MetaboAnalyst: a web server for metabolomic data analysis and interpretation.,” *Nucleic acids research*, vol. 37, pp. W652–60, jul 2009.

[40] C. A. Smith, E. J. Want, G. O’Maille, R. Abagyan, and G. Siuzdak, “XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.,” *Analytical Chemistry*, vol. 78, pp. 779–787, feb 2006.

[41] R. Tautenhahn, G. J. Patti, D. Rinehart, and G. Siuzdak, “XCMS Online: a web-based platform to process untargeted metabolomic data.,” *Analytical chemistry*, vol. 84, pp. 5035–9, jun 2012.

[42] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic, “MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.,” *BMC bioinformatics*, vol. 11, no. 1, pp. 395–405, 2010.

[43] S. Böcker, M. C. Letzel, Z. Lipták, and A. Pervukhin, “SIRIUS: decomposing isotope patterns for metabolite identification.,” *Bioinformatics (Oxford, England)*, vol. 25, pp. 218–24, jan 2009.

[44] J. Draper, D. P. Enot, D. Parker, M. Beckmann, S. Snowdon, W. Lin, and H. Zubair, “Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour ’rules’.,” *BMC bioinformatics*, vol. 10, p. 227, jan 2009.

[45] T. Yu, Y. Park, J. M. Johnson, and D. P. Jones, “apLCMS–adaptive processing of high-resolution LC/MS data.,” *Bioinformatics (Oxford, England)*, vol. 25, pp. 1930–6, aug 2009.

[46] J. E. Katz, D. S. Dumlao, S. Clarke, and J. Hau, “A new technique (COMSPARI) to facilitate the identification of minor compounds in complex mixtures by GC/MS and LC/MS: Tools for the visualization of matched datasets,” *Journal of the American Society for Mass Spectrometry*, vol. 15, no. 4, pp. 580–584, 2004.

[47] A. Lommen, “MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing.,” *Analytical chemistry*, vol. 81, pp. 3079–86, apr 2009.

[48] W. Zhang, Z. Lei, D. Huhman, L. W. Sumner, and P. X. Zhao, “MET-XAlign: a metabolite cross-alignment tool for LC/MS-based comparative metabolomics.,” *Analytical chemistry*, vol. 87, pp. 9114–9, sep 2015.

[49] C. D. Broeckling, I. R. Reddy, A. L. Duran, X. Zhao, and L. W. Sumner, “MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics,” *Analytical Chemistry*, vol. 78, no. 13, pp. 4334–4341, 2006.

[50] X. Wei, W. Sun, X. Shi, I. Koo, B. Wang, J. Zhang, X. Yin, Y. Tang, B. Bogdanov, S. Kim, Z. Zhou, C. McClain, and X. Zhang, “MetSign: A computational platform for high-resolution mass spectrometry-based metabolomics,” *Analytical Chemistry*, vol. 83, no. 20, pp. 7668–7675, 2011.

[51] A. L. Duran, J. Yang, L. Wang, and L. W. Sumner, “Metabolomics spectral formatting, alignment and conversion tools (MSFACTs),” *Bioinformatics*, vol. 19, no. 17, pp. 2283–2293, 2003.

[52] P. M. Palagi, D. Walther, M. Quadroni, S. Catherinet, J. Burgess, C. G. Zimmermann-Ivol, J. C. Sanchez, P. A. Binz, D. F. Hochstrasser, and R. D. Appel, “MSight: an image analysis software for liquid chromatography-mass spectrometry,” *Proteomics*, vol. 5, no. 9, pp. 2381–2384, 2005.

[53] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh, “A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS.,” *Bioinformatics (Oxford, England)*, vol. 22, pp. 1902–9, aug 2006.

[54] B. L. LaMarche, K. L. Crowell, N. Jaitly, V. A. Petyuk, A. R. Shah, A. D. Polpitiya, J. D. Sandoval, G. R. Kiebel, M. E. Monroe, S. J. Callister, T. O. Metz, G. A. Anderson, and R. D. Smith, “MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis.,” *BMC bioinformatics*, vol. 14, p. 49, jan 2013.

[55] L. N. Mueller, O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M.-Y. Brusniak, O. Vitek, R. Aebersold, and M. Müller, “SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling.,” *Proteomics*, vol. 7, pp. 3470–80, oct 2007.

[56] P. G. Pedrioli, “Trans-proteomic pipeline: a pipeline for proteomic analysis,” *Methods Mol Biol*, vol. 604, pp. 213–238, 2010.

[57] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. a. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, and P. Mallick, “A cross-platform toolkit for mass spectrometry and proteomics,” *Nature Biotechnology*, vol. 30, no. 10, pp. 918–920, 2012.

[58] M. Daszykowski and B. Walczak, “Use and abuse of chemometrics in chromatography,” *TrAC Trends in Analytical Chemistry*, vol. 25, pp. 1081–1096, dec 2006.

[59] A. N. Krutchinsky and B. T. Chait, “On the nature of the chemical noise in MALDI mass spectra,” *J. Am. Chem. Soc. Mass. Spectrom.*, vol. 13, pp. 129–134, 2002.

[60] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, “Electrospray ionization for mass spectrometry of large biomolecules,” *Science*, vol. 246, no. 4926, pp. 64–71, 1989.

[61] P. Haimi, P. Haimi, A. Uphoff, A. Uphoff, M. Hermansson, M. Hermansson, P. Somerharju, and P. Somerharju, “Software tools for analysis of mass spectrometric lipidome data.,” *Analytical chemistry*, vol. 78, no. 24, pp. 8324–31, 2006.

[62] A. A. C. Sauve and T. P. T. Speed, “Normalization, baseline correction and alignment of high-throughput mass spectrometry data,” in *Proceedings Gensips*, 2004.

[63] W. Wang, C. H. Becker, H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, and M. Anderle, “Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards,” *Analytical Chemistry*, vol. 75, no. 18, pp. 4818–4826, 2003.

[64] P. H. C. Eilers, “A perfect smoother.,” *Analytical chemistry*, vol. 75, pp. 3631–6, jul 2003.

[65] D. Radulovic, S. Jelveh, S. Ryu, T. G. Hamilton, E. Foss, Y. Mao, and A. Emili, “Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry.,” *Molecular & cellular proteomics : MCP*, vol. 3, pp. 984–97, oct 2004.

[66] C. A. Hastings, S. M. Norton, and S. Roy, “New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data,” *Rapid Commun Mass Spectrom*, vol. 16, no. 5, pp. 462–467, 2002.

[67] X. Wei, X. Shi, S. Kim, L. Zhang, J. S. Patrick, J. Binkley, C. McClain, and X. Zhang, “Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics,” *Analytical chemistry*, vol. 84, pp. 7963–71, sep 2012.

[68] S. Purvine, N. Kolker, and E. Kolker, “Spectral quality assessment for high-throughput tandem mass spectrometry proteomics.,” *Omics : a journal of integrative biology*, vol. 8, pp. 255–65, jan 2004.

[69] H. Xu and M. A. Freitas, “A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra.,” *BMC bioinformatics*, vol. 11, p. 436, jan 2010.

[70] X.-j. Li, E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold, “A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry.,” *Molecular & cellular proteomics : MCP*, vol. 4, pp. 1328–40, sep 2005.

[71] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biol*, vol. 5, no. 10, p. R80, 2004.

[72] M. Katajamaa and M. Oresic, "Data processing for mass spectrometry-based metabolomics.," *Journal of chromatography. A*, vol. 1158, no. 1-2, pp. 318–28, 2007.

[73] M. Wehofsky, R. Hoffmann, M. Hubert, and B. Spengler, "Isotopic deconvolution of matrix-assisted laser desorption/ionization mass spectra for substance-class specific analysis of complex samples," *Eur J Mass Spectrom (Chichester, Eng)*, vol. 7, pp. 39–46, 2001.

[74] K. C. Leptos, D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church, "MapQuant: Open-source software for large-scale protein quantification," *Proteomics*, vol. 6, no. 6, pp. 1770–1782, 2006.

[75] M. Hermansson, A. Uphoff, R. Käkelä, and P. Somerharju, "Automated quantitative analysis of complex lipidomes by liquid chromatography/mass spectrometry," *Analytical chemistry*, vol. 77, no. 7, pp. 2166–2175, 2005.

[76] F. E. Lytle and R. K. Julian, "Automatic Processing of Chromatograms in a High-Throughput Environment.," *Clinical chemistry*, vol. 62, pp. 144–53, jan 2016.

[77] E. Grushka, "Characterization of exponentially modified Gaussian peaks in chromatography.," *Analytical chemistry*, vol. 44, pp. 1733–8, sep 1972.

[78] M. Kempka, J. Sjödahl, A. Björk, and J. Roeraade, "Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.," *Rapid communications in mass spectrometry : RCM*, vol. 18, pp. 1208–1212, jan 2004.

[79] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," *Bioinformatics*, vol. 22, no. 17, pp. 2059–2065, 2006.

[80] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis.," *BMC bioinformatics*, vol. 10, p. 4, 2009.

[81] M. Holčapek, R. Jirásko, and M. Lísa, "Basic rules for the interpretation of atmospheric pressure ionization mass spectra of small molecules," *Journal of Chromatography A*, vol. 1217, no. 25, pp. 3908–3921, 2010.

[82] G. Alves, A. Y. Ogurtsov, and Y. K. Yu, "Molecular Isotopic Distribution Analysis (MIDAs) with adjustable mass accuracy," *Journal of the American Society for Mass Spectrometry*, vol. 25, no. 1, pp. 57–70, 2014.

[83] M. W. Senko, S. C. Beu, and F. W. McLaffertycor, “Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions,” *Journal of the American Society for Mass Spectrometry*, vol. 6, no. 4, pp. 229–233, 1995.

[84] N. Jaitly, A. Mayampurath, K. Littlefield, J. N. Adkins, G. a. Anderson, and R. D. Smith, “Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data.,” *BMC Bioinformatics*, vol. 10, p. 87, 2009.

[85] D. M. Horn, R. a. Zubarev, and F. W. McLafferty, “Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules,” *Journal of the American Society for Mass Spectrometry*, vol. 11, no. 4, pp. 320–332, 2000.

[86] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard, “Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping,” *Journal of Chromatography A*, vol. 805, pp. 17–35, may 1998.

[87] J. Listgarten, R. M. Neal, S. T. Roweis, P. Wong, and A. Emili, “Difference detection in LC-MS data for protein biomarker discovery.,” *Bioinformatics (Oxford, England)*, vol. 23, pp. e198–204, jan 2007.

[88] M. Katajamaa and M. Oresic, “Processing methods for differential analysis of LC/MS profile data.,” *BMC Bioinformatics*, vol. 6, pp. 179–192, 2005.

[89] C. Salazar, J. Schütze, and O. Ebenhöh, “Bioinformatics meets systems biology,” *Genome Biology*, vol. 7, no. 1, p. 303, 2006.

[90] E. Lange, R. Tautenhahn, S. Neumann, and C. Gröpl, “Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements,” *BMC Bioinformatics*, vol. 9, no. 1, p. 375, 2008.

[91] S. E. Stein and D. R. Scott, “Optimization and Testing of Mass-Spectral Library Search Algorithms for Compound Identification,” *Journal of the American Society for Mass Spectrometry*, vol. 5, no. 9, pp. 859–866, 1994.

[92] T. Kind and O. Fiehn, “Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm.,” *BMC bioinformatics*, vol. 7, no. 1, p. 234, 2006.

[93] T. Kind and O. Fiehn, “Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry.,” *BMC bioinformatics*, vol. 8, p. 105, jan 2007.

[94] G. T. Gipson, K. S. Tatsuoka, B. A. Sokhansanj, R. J. Ball, and S. C. Connor, “Assignment of MS-based metabolomic datasets via compound interaction pair mapping,” *Metabolomics*, vol. 4, no. 1, pp. 94–103, 2008.

[95] S. Rogers, R. A. Scheltema, M. Girolami, and R. Breitling, “Probabilistic assignment of formulas to mass peaks in metabolomics experiments,” *Bioinformatics*, vol. 25, no. 4, pp. 512–518, 2009.

[96] R. J. M. Weber and M. R. Viant, “MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways,” *Chemometrics and Intelligent Laboratory Systems*, vol. 104, no. 1, pp. 75–82, 2010.

[97] H. Doerfler, X. Sun, L. Wang, D. Engelmeier, D. Lyon, and W. Weckwerth, “mzGroupAnalyzer-predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data,” *PLoS ONE*, vol. 9, no. 5, 2014.

[98] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, and Y. Liang, “Chemometric methods in data processing of mass spectrometry-based metabolomics: A review,” *Analytica Chimica Acta*, vol. 914, pp. 17–34, feb 2016.

[99] R. A. Dromms and M. P. Styczynski, “Systematic applications of metabolomics in metabolic engineering.,” *Metabolites*, vol. 2, pp. 1090–122, dec 2012.

[100] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, “Metabolomics by numbers: acquiring and understanding global metabolite data.,” *Trends in biotechnology*, vol. 22, pp. 245–52, may 2004.

[101] I. Narsky and F. C. Porter, “Methods for Variable Ranking and Selection,” in *Statistical Analysis Techniques in Particle Physics*, ch. 18, pp. 385–415, Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, nov 2013.

[102] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, “A review of variable selection methods in Partial Least Squares Regression,” *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012.

[103] S. Wold, M. Sjöström, and L. Eriksson, “Partial Least Squares Projections to Latent Structures (PLS) in Chemistry,” in *Encyclopedia of Computational Chemistry*, pp. 523–550, Chichester, UK: John Wiley & Sons, Ltd, apr 2002.

[104] S. Favilla, C. Durante, M. L. Vigni, and M. Cocchi, “Assessing feature relevance in NPLS models by VIP,” *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 76–86, 2013.

[105] S. Wold, M. Sjöström, and L. Eriksson, “PLS-regression: a basic tool of chemometrics,” *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[106] O. M. Kvalheim, “Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots,” *Journal of Chemometrics*, vol. 24, pp. 496–504, jul 2010.

[107] L. Yi, N. Dong, S. Shi, B. Deng, Y. Yun, Z. Yi, and Y. Zhang, “Metabolomic identification of novel biomarkers of nasopharyngeal carcinoma,” *RSC Adv.*, vol. 4, pp. 59094–59101, oct 2014.

[108] Y.-H. Yun, B.-C. Deng, D.-S. Cao, W.-T. Wang, and Y.-Z. Liang, "Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery.," *Analytica chimica acta*, vol. 911, pp. 27–34, mar 2016.

[109] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[110] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, and D.-S. Cao, "Model population analysis for variable selection," *Journal of Chemometrics*, vol. 24, pp. 418–423, jul 2010.

[111] Z. Huang, Y. Chen, W. Hang, Y. Gao, L. Lin, D. Y. Li, J. Xing, and X. Yan, "Holistic metabonomic profiling of urine affords potential early diagnosis for bladder and kidney cancers," *Metabolomics*, vol. 9, pp. 119–129, feb 2013.

[112] N. Bhattacharya, A. Singh, A. Ghanate, G. Phadke, D. Parmar, D. Dhaware, T. Basak, S. Sengupta, and V. Panchagnula, "Matrix-assisted laser desorption/ionization mass spectrometry analysis of dimethyl arginine isomers from urine," *Analytical Methods*, vol. 6, no. 13, pp. 4602–4609, 2014.

[113] J. F. Moxley, M. C. Jewett, M. R. Antoniewicz, S. G. Villas-Boas, H. Alper, R. T. Wheeler, L. Tong, A. G. Hinnebusch, T. Ideker, J. Nielsen, and G. Stephanopoulos, "Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p," *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6477–6482, 2009.

[114] J. M. Buescher, M. R. Antoniewicz, L. G. Boros, S. C. Burgess, H. Brunengraber, C. B. Clish, R. J. DeBerardinis, O. Feron, C. Frezza, B. Ghesquiere, E. Gottlieb, K. Hiller, R. G. Jones, J. J. Kamphorst, R. G. Kibbey, A. C. Kimmelman, J. W. Locasale, S. Y. Lunt, O. D. K. Maddocks, C. Malloy, C. M. Metallo, E. J. Meuillet, J. Munger, K. Nöh, J. D. Rabinowitz, M. Ralser, U. Sauer, G. Stephanopoulos, J. St-Pierre, D. A. Tennant, C. Wittmann, M. G. Vander Heiden, A. Vazquez, K. Vousden, J. D. Young, N. Zamboni, and S.-M. Fendt, "A roadmap for interpreting (13)C metabolite labeling patterns from cells.," *Current opinion in biotechnology*, vol. 34, pp. 189–201, aug 2015.

[115] G. Stephanopoulos, "Metabolic Fluxes and Metabolic Engineering," *Metabolic Engineering*, vol. 1, no. 1, pp. 1–11, 1999.

[116] S. Christen and U. Sauer, "Intracellular characterization of aerobic glucose metabolism in seven yeast species by13C flux analysis and metabolomics," *FEMS Yeast Research*, vol. 11, no. 3, pp. 263–272, 2011.

[117] T. Hasunuma, T. Sanda, R. Yamada, K. Yoshimura, J. Ishii, and A. Kondo, "Metabolic pathway engineering based on metabolomics confers acetic and formic acid tolerance to a recombinant xylose-fermenting strain of Saccharomyces cerevisiae," *Microbial Cell Factories*, vol. 10, 2011.

[118] M. Kogadeeva and N. Zamboni, "SUMOFLUX: A Generalized Method for Targeted 13C Metabolic Flux Ratio Analysis," *PLoS Computational Biology*, vol. 12, no. 9, 2016.

[119] Y. Morales, G. Bosque, J. Vehí, J. Picó, and F. Llaneras, “PFA toolbox: A MATLAB tool for Metabolic Flux Analysis,” *BMC Systems Biology*, vol. 10, no. 1, 2016.

[120] D. Segre, D. Vitkup, and G. M. Church, “Analysis of optimality in natural and perturbed metabolic networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 15112–15117, 2002.

[121] H. Alper, K. Miyaoku, and G. Stephanopoulos, “Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets,” *Nature Biotechnology*, vol. 23, no. 5, pp. 612–616, 2005.

[122] C. M. Ghim, K. I. Goh, and B. Kahng, “Lethality and synthetic lethality in the genome-wide metabolic network of Escherichia coli,” *Journal of Theoretical Biology*, vol. 237, no. 4, pp. 401–411, 2005.

[123] O. Folger, L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi, “Predicting selective drug targets in cancer through metabolic networks.,” *Molecular systems biology*, vol. 7, p. 501, jun 2011.

[124] I. Apaolaza, E. San José-Eneriz, L. Tobalina, E. Miranda, L. Garate, X. Agirre, F. Prósper, and F. J. Planes, “An in-silico approach to predict and exploit synthetic lethality in cancer metabolism,” *Nature Communications*, vol. 8, p. 459, dec 2017.

[125] J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. Palsson, “Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0,” *Nature Protocols*, vol. 6, no. 9, pp. 1290–1307, 2011.

[126] A. Römpp and B. Spengler, “Mass spectrometry imaging with high resolution in mass and space,” *Histochemistry and Cell Biology*, vol. 139, no. 6, pp. 759–783, 2013.

[127] M. Strohalm, M. Hassman, B. Košata, and M. Kodíček, “mMass data miner: An open source alternative for mass spectrometric data analysis,” *Rapid Communications in Mass Spectrometry*, vol. 22, no. 6, pp. 905–908, 2008.

[128] M. Strohalm, D. Kavan, P. Novák, M. Volný, and V. Havlícek, “mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data.,” *Analytical Chemistry*, vol. 82, pp. 4648–4651, jun 2010.

[129] M. Hirosawa, M. Hoshida, M. Ishikawa, and T. Toya, “MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming.,” *Computer applications in the biosciences : CABIOS*, vol. 9, no. 2, pp. 161–167, 1993.

[130] M. Sturm and O. Kohlbacher, “TOPPView: An open-source viewer for mass spectrometry data,” *Journal of Proteome Research*, vol. 8, no. 7, pp. 3760–3763, 2009.

[131] A. Singh and V. Panchagnula, “High throughput quantitative analysis of melamine and triazines by MALDI-TOF MS,” *Analytical Methods*, vol. 3, no. 10, pp. 2360–2366, 2011.

[132] S. R. C. Immanuel, A. D. Ghanate, D. S. Parmar, F. Marriage, V. Panchagnula, P. J. Day, and A. Raghunathan, “Integrative analysis of rewired central metabolism in temozolomide resistant cells.,” *Biochemical and biophysical research communications*, vol. 495, pp. 2010–2016, jan 2018.

[133] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, “ProteoWizard: Open source software for rapid proteomics tools development,” *Bioinformatics*, vol. 24, no. 21, pp. 2534–2536, 2008.

[134] S. G. Roussis and R. Proulx, “Reduction of chemical formulas from the isotopic peak distributions of high-resolution mass spectra.,” *Analytical Chemistry*, vol. 75, pp. 1470–1482, mar 2003.

[135] A. L. Rockwood and P. Haimi, “Efficient calculation of accurate masses of isotopic peaks,” *Journal of the American Society for Mass Spectrometry*, vol. 17, no. 3, pp. 415–419, 2006.

[136] S. Sibani, S. Melnyk, I. P. Pogribny, W. Wang, F. Hiou-Tim, L. Deng, J. Trasler, S. J. James, and R. Rozen, “Studies of methionine cycle intermediates (SAM, SAH), DNA methylation and the impact of folate deficiency on tumor numbers in Min mice.,” *Carcinogenesis*, vol. 23, no. 1, pp. 61–65, 2002.

[137] S. J. James, P. Cutler, S. Melnyk, S. Jernigan, L. Janak, D. W. Gaylor, and J. A. Neubrander, “Metabolic biomarkers of increased oxidative stress and impaired methylation capacity in children with autism,” *American Journal of Clinical Nutrition*, vol. 80, no. 6, pp. 1611–1617, 2004.

[138] P. Du, G. Stolovitzky, P. Horvatovich, R. Bischoff, J. Lim, and F. Suits, “A noise model for mass spectrometry based proteomics,” *Bioinformatics*, vol. 24, no. 8, pp. 1070–1077, 2008.

[139] A. Makarov, E. Denisov, A. Kholomeev, W. Balschun, O. Lange, K. Strupat, and S. Horning, “Performance Evaluation of a Hybrid Linear Ion Trap / Orbitrap Mass Spectrometer,” *Analytical Chemistry*, vol. 78, no. 7, pp. 2113–2120, 2006.

[140] R. a. Zubarev and A. Makarov, “Orbitrap mass spectrometry,” *Analytical Chemistry*, vol. 85, no. 11, pp. 5288–5296, 2013.

[141] K. Busch, “Chemical Noise in Mass Spectrometry III,” *Spectroscopy*, vol. 17(10) Oct, no. 5, pp. 32–37, 2002.

[142] Greg Wells, Harry Prest, and Charles William Russ IV, “Why Use Signal-To-Noise As a Measure of MS Performance When It Is Often Meaningless?,” *Current Topics in Mass Spectrometry*, pp. 28–33, 2011.

[143] Z. Zhang and a. G. Marshall, “A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra.,” *Journal of the American Society for Mass Spectrometry*, vol. 9, no. 3, pp. 225–233, 1998.

[144] A. Makarov, E. Denisov, O. Lange, and S. Horning, "Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer," *Journal of the American Society for Mass Spectrometry*, vol. 17, no. 7, pp. 977–982, 2006.

[145] S.-Y. Lee, E. Son, J.-Y. Kang, H.-S. Lee, M.-K. Shin, H.-S. Nam, S.-Y. Kim, Y.-M. Jang, and G.-S. Rhee, "Development of a Quantitative Analytical Method for Determining the Concentration of Human Urinary Paraben by LC-MS/MS," *Bulletin of the Korean Chemical Society*, vol. 34, pp. 1131–1136, apr 2013.

[146] USP-NF, "First Supplement to USP 40–NF 35," *United States Pharmacopeia and National Formulary (USP 40-NF 35)*, pp. 621:1–12, 2017.

[147] K. Yizhak, E. Gaude, S. Le Dévédec, Y. Y. Waldman, G. Y. Stein, B. van de Water, C. Frezza, and E. Ruppin, "Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer," *eLife*, vol. 3, no. November, pp. 1–23, 2014.

[148] D. C. Zielinski, N. Jamshidi, A. J. Corbett, A. Bordbar, A. Thomas, and B. O. Palsson, "Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism," *Scientific Reports*, vol. 7, p. 41241, jan 2017.

[149] P. Maia, M. Rocha, and I. Rocha, "In Silico Constraint-Based Strain Optimization Methods: the Quest for Optimal Cell Factories.," *Microbiology and molecular biology reviews : MMBR*, vol. 80, pp. 45–67, mar 2016.

[150] S. A. Becker and B. O. Palsson, "Context-specific metabolic networks are consistent with experiments," *PLoS Computational Biology*, vol. 4, no. 5, 2008.

[151] T. Shlomi, M. N. Cabili, M. J. Herrgård, B. Ø. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nature Biotechnology*, vol. 26, pp. 1003–1010, sep 2008.

[152] H. Zur, E. Ruppin, and T. Shlomi, "iMAT: An integrative metabolic analysis tool," *Bioinformatics*, vol. 26, no. 24, pp. 3140–3142, 2010.

[153] P. A. Jensen and J. A. Papin, "Functional integration of a metabolic network model and expression data without arbitrary thresholding," *Bioinformatics*, vol. 27, pp. 541–547, feb 2011.

[154] C. Colijn, A. Brandes, J. Zucker, D. S. Lun, B. Weiner, M. R. Farhat, T.-Y. Cheng, D. B. Moody, M. Murray, and J. E. Galagan, "Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production.," *PLoS computational biology*, vol. 5, p. e1000489, aug 2009.

[155] S. Chandrasekaran and N. D. Price, "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 17845–50, oct 2010.

[156] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. Ø. Palsson, "Global reconstruction of the human metabolic network based on genomic and bibliomic data.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 1777–82, feb 2007.

[157] A. H. Ramos, L. Lichtenstein, M. Gupta, M. S. Lawrence, T. J. Pugh, G. Saksena, M. Meyerson, and G. Getz, "Oncotator: Cancer Variant Annotation Tool," *Human Mutation*, vol. 36, pp. E2423–E2429, apr 2015.

[158] K. R. Tanaka and C. R. Zerez, "Red cell enzymopathies of the glycolytic pathway.," *Seminars in hematology*, vol. 27, pp. 165–85, apr 1990.

[159] G. Jacobasch and S. M. Rapoport, "Hemolytic anemias due to erythrocyte enzyme deficiencies.," *Molecular aspects of medicine*, vol. 17, pp. 143–70, apr 1996.

[160] N. D. Price, J. Schellenberger, and B. O. Palsson, "Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies.," *Biophysical journal*, vol. 87, pp. 2172–86, oct 2004.

[161] A. Bar-Even, E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik, and R. Milo, "The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters," *Biochemistry*, vol. 50, pp. 4402–4410, may 2011.

[162] P. S. Ward and C. B. Thompson, "Metabolic reprogramming: a cancer hallmark even warburg did not anticipate.," *Cancer cell*, vol. 21, pp. 297–308, mar 2012.

[163] C. Wagner, "BIOCHEMICAL ROLE OF FOLATE IN CELLULAR METABOLISM*," *Clinical Research and Regulatory Affairs*, vol. 18, pp. 161–180, jan 2001.

[164] M. Krebs, A. Bellon, G. Mainguy, T. Jay, and H. Frieling, "One-carbon metabolism and schizophrenia: current challenges and future directions," *Trends in Molecular Medicine*, vol. 15, pp. 562–570, dec 2009.

[165] M. Jain, R. Nilsson, S. Sharma, N. Madhusudhan, T. Kitami, A. L. Souza, R. Kafri, M. W. Kirschner, C. B. Clish, and V. K. Mootha, "Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation," *Science*, vol. 336, pp. 1040–1044, may 2012.

[166] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, and C. Magnes, "IPO: a tool for automated optimization of XCMS parameters.," *BMC bioinformatics*, vol. 16, p. 118, 2015.

[167] X. Domingo-Almenara, J. R. Montenegro-Burke, H. P. Benton, and G. Siuzdak, "Annotation: A Computational Solution for Streamlining Metabolomics Analysis.," *Analytical chemistry*, vol. 90, pp. 480–489, jan 2018.

Index