COMPUTATIONAL HYPOTHESIS GENERATION WITH GENOME-WIDE METABOLIC RECONSTRUCTIONS:

IN-SILICO PREDICTION OF METABOLIC CHANGES IN THE FRESHWATER MODEL ORGANISM *DAPHNIA* TO ENVIRONMENTAL STRESSORS

By

James Bradbury

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Computer Science

College of Engineering and Physical Sciences

University of Birmingham

September 2017

Abstract

Computational toxicology is an emerging, multidisciplinary field that uses *in-silico* modelling techniques to predict and understand how biological organisms interact with pollutants and environmental stressors. Genome-wide metabolic reconstruction (GWMR) is an *in-silico* modelling technique that aims to represent the metabolic capabilities of an organism at a genomic scale by representing a metabolome as a network of connected nodes. GWMRs provide a platform for analysis, visualisation and contextualisation of omics datasets and have untapped potential for use in computational toxicology.

Environmental metabolomics is the application of metabolomics to study how living organisms interact with their environment. *Daphnia* is an emerging model species for environmental omics whose underlying biology is still being uncovered. Creating a metabolic reconstruction of *Daphnia* and applying it in an environmental computational toxicology setting has the potential to aid in understanding its interaction with environmental stressors. Here, the fist GWMR of *D. magna* is presented, which is built using METRONOME, a newly developed tool for automated GWMR of new genome sequences.

Active module identification allows for omics data sets to be integrated into *in-silico* models and uses optimisation algorithms to find *hot-spots* within networks that represent areas that are significantly impacted based on a toxicogenomic transcriptomics dataset. Previous work has used the active module identification approach with metabolic networks to investigate underlying genomic mechanisms behind known metabolic responses. Here, a method that uses the active modules approach in a predictive capacity for computational hypothesis generation is introduced. Active module identification is

used with the *Daphnia* GWMR to predict unknown metabolic responses to environmentally relevant human-induced stressors.

A computational workflow is presented that takes a new genome sequence, builds a GWMR and integrates gene expression data to make predictions of metabolic effects. The aim is to introduce an element of hypothesis generation into the untargeted metabolomics experimental workflow. A study to validate this approach using *D. magna* as the target organism is presented, which uses untargeted Liquid-Chromatography Mass Spectrometry (LC-MS) to make metabolomics measurements. A software tool MUSCLE is presented that uses multi-objective closed-loop evolutionary optimisation to automatically develop LC-MS instrument methods and is used here to develop the analytical method. Some positive results are obtained, but difficulties with the dataset make it hard to draw any concrete conclusions.

Acknowledgments

I would foremost like to thank Martinique, Nora and my family. Without their love and support over the past four years, this work would not have been possible.

Thank you to my supervisors Shan He and Mark Viant and their respective research groups. Particular thanks goes to Martin Jones, Tom Lawson for the crash course in metabolomics data processing. Thank you to Luisa Orsini and Ulf Sommer for their involvement in the experimental work.

This work would not have been possible without the generous funding from the National Environmental Research Council.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Systems biology

Systems biology endeavours to revolutionise our knowledge of how biological systems behave. Traditionally biological systems are studied using a top-down reductionist approach. This approach characterises, in high detail, individual components (such as genes, enzymes or metabolites) of a larger biological system to build up an understanding of the processes that these single elements are involved in. The systems biology approach differs in that it studies biological systems as a whole, striving toward developing a mechanistic understanding of biological systems at a systems level (Kitano, 2002).

The ultimate goal is to transform biology into a precise science by establishing a holistic mechanism for studying biological systems, following established practices in systems engineering (Laszlo, 1996; Weinberg, 2011). Central to this is the generation of *in-silico* models that allow for complex biological systems to be modelled. These models often take the form of network graphs which can be used to describe the interactions between genes (Hecker et al, 2009), transcripts (McClure et al, 2016), proteins (Vazquez, 2010) and metabolites (Thiele & Palsson, 2010).

This new approach to biology promises to modernise the field, allowing for complex biological behaviour to be studied and modelled. High throughput technologies have transformed biology into a big data science, which brings with it new and exciting challenges (Marx, 2013). In order to study and understand biological systems as a whole in this data rich environment, biologists must develop and apply new techniques through collaborations with computer scientists, statisticians and informaticians.

Systems biology has been made possible by the advent of high throughput omics technologies. These technologies allow for an unbiased, non-targeted snapshot of all elements at the level of interest (Figure 1.1) to be measured in a high throughput manner (Blankenburg et al, 2009). This fits with the holistic systems biology paradigm as it considers the biological systems being measured as a whole. These techniques provide a different starting point when conducting biological research; they are hypothesis generating rather than being hypothesis driven (Horgan & Kenny, 2011).

1.2. Omics science

Technologies exist for genomics, transcriptomics, proteomics and metabolomics for measurement of genes, mRNA, proteins and metabolites respectively. These technologies are relatively new are still experiencing rapid technological advancement (Scott & Treff, 2010). This technological advancement has enabled a greater intersection between technology and biology, necessitating the need for the rapid advancement in the field of bioinformatics (Ning & Lo, 2010).

Omics sciences are in an embryonic stage with a number of challenges being faced. Reproducibility is a big issue in omics science due to the complexity of the experimental designs and methods along with the sensitivity of the instruments used (Petricoin et al, 2002a; Ransohoff, 2005; Zhu et al, 2003). This sensitivity coupled to the fact that so much biological information is being measured also results in noisy data sets, where meaningful biological information can be lost or hidden (Amariei et al, 2014; Anderson, 2010). Omics science has also transformed biological data sets into a high dimensional space, requiring a vastly different approach to interpretation, both in the sense of understanding statistical relevancy and biological interpretation of the data sets (Clarke et al, 2008). This further underlines the importance collaborations with computer scientists, statisticians and

informaticians to help face these challenges posed by this new way of conducting biological research.

1.3. Computational methods and omics science

The transformation of biology into a data rich discipline through the application of omics science has opened up a number of opportunities for computational methods to be applied. A wide range of challenges have presented themselves. Data processing, data storage data retrieval and data standardisation is now key to conducting omics scientific research efficiently (Berger et al, 2013).

A large number of data processing tools exist for omics science (Benton et al, 2010; Cacciatore et al, 2017; Chambers et al, 2012; Dao et al, 2011; Gentleman et al, 2004; Kuhl et al, 2012; Kurczy et al, 2015; Libiseller et al, 2015; Overbeek et al, 2005; Parsons et al, 2007; Scheltema et al, 2011; Selivanov et al, 2017; Smith et al, 2006; Tautenhahn et al, 2008; Xia et al, 2015). A number of these make use of clustering, optimisation, network modelling, data mining, signal processing and software engineering disciplines. Recently there has been a large effort to increase data standardisation (Rocca-Serra et al, 2016; Salek et al, 2015) and management practices (Hastings et al, 2016; Haug et al, 2017; Kale et al, 2016). Platforms such as galaxy (Goecks et al, 2010) and Worflow4Metabolomics (Giacomoni et al, 2015) are also available for standardising data processing (Davidson et al, 2016; Giacomoni et al, 2015; Karaman et al, 2016; Weber et al, 2017).

1.4. Computational toxicology

Computational toxicology is an emerging field that uses *in-silico* modelling techniques to predict and understand how biological organisms interact with pollutants,

environmental stressors and pharmaceuticals based on biological and chemical datasets (Rusyn & Daston, 2010). The field is inherently multidisciplinary, incorporating computer science, systems biology, biostatistics, toxicology, biochemistry, and medicine. The field has been accelerated by the advancement in *in-silico* modelling techniques for biological systems and the stream of high dimensional datasets produced by omics science. The natural progression of these techniques is to move toward a capability of using these models in a predictive capacity, to predict the biological response to chemicals, akin to traditional toxicology. Large scale omics studies involve a high number of samples and are costly, both in terms of time, personnel and the costs of procuring and maintaining complex analytical platforms, computational toxicology seeks to reduce this burden.

Applications of computational toxicology include understanding the hazards and risks of chemicals, environmental science and drug safety (Reisfeld & Mayeno, 2012), and is also seen as a key to environmental health protection and regulatory decision making (Kavlock et al, 2009; Rusyn & Daston, 2010). There is a clear benefit to the use of computational toxicology in the field of environmental toxicology and metabolomics.

1.5. Biological systems

Through the use of omics technologies, the field of systems biology aims to generate detailed lists of biological components and ultimately reconstruct *in-silico* models of the comprehensive functional network of a living organism. This complex and multi-layered network (also known as the interactome) represents the complex biological interactions that take place within all living organisms, with layers representing the genome, transcriptome, proteome and metabolome (Figure 1.1).

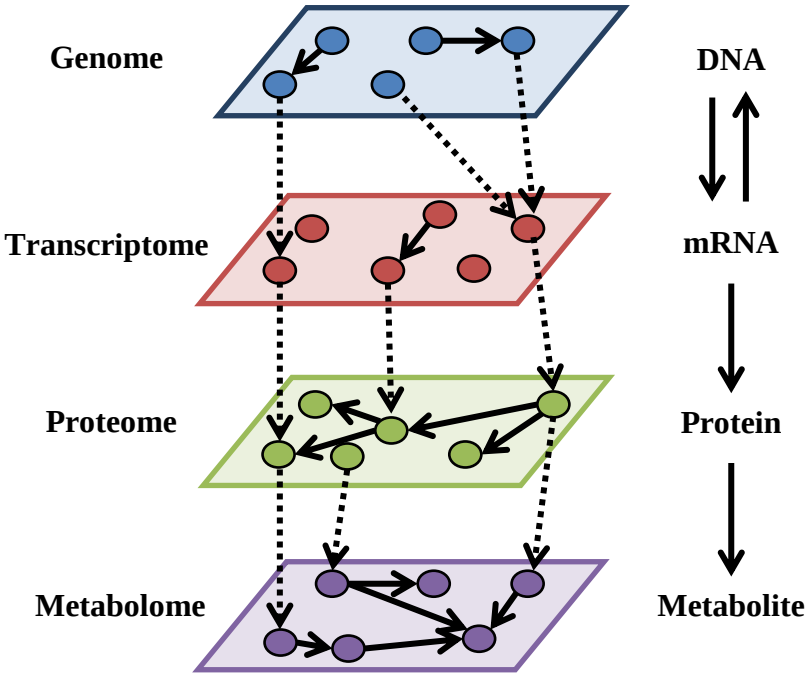


Figure 1.1: Multi-layered biological network. The interaction between the genome, transcriptome, proteome and metabolome is also known as the central dogma.

All of the information for each of these layers can be traced back to an organism's genome sequence. A genome sequence consists of DNA, which constitutes the passive part of the biochemistry of the cell, with the active part being achieved through transcription which leads to proteins catalysing biochemical reactions as well as many other cell mechanisms. This is also known as the central dogma, Figure 1.1 shows how the different layers of the interactome interact. Each layer of the interactome can be measured using omics technologies, and has a dedicated field associated with it; genomics, transcriptomics, proteomics and metabolomics.

1.6. Metabolomics

The metabolome layer is the most downstream of the layers of the interactome. It is directly affected by the proteome, which is in turn affected by gene expression in the

transcriptome layer. The metabolic profile, or metabolome, of an organism can therefore be seen as a combination of its gene expression and cellular metabolism. In other words, the metabolome can be seen as the phenotype, or observable manifestation, of the changes in its gene expression (Fiehn, 2002) caused by interaction with an organism and its environment.

Metabolomics is the application of omics technologies to measure all naturally-occurring low weight biological compounds, metabolites, within a given sample (Harrigan & Goodacre, 2012; Lindon et al, 2011; Wang et al, 2007) to study the interaction of an organism with its natural environment (Viant, 2008). Metabolites are the smallest components of the interactome, and are responsible for many biological functions such as producing energy and producing basic materials required for important life processes (Gancedo & Serrano, 1989; Meurant, 2012). Metabolism is the biochemical modification of these chemical compounds in cells and organisms.

The process in which two or more metabolites (reactants) interact and produce other metabolites (products), is called a biochemical reaction, and are typically catalysed by at least one enzyme. Metabolic processes consist of a number of biochemical reactions chained together to form a metabolic pathway (Figure 1.2). These pathways achieve either the formation of another metabolic product to be used or stored in a cell, or the initiation of another metabolic pathway. Examples of metabolic pathways include the glycolysis pathway, which converts sugars into energy to be stored in cells, and the citric acid cycle, which releases stored energy. The collection of metabolic pathways within an organism is known as a metabolic network.

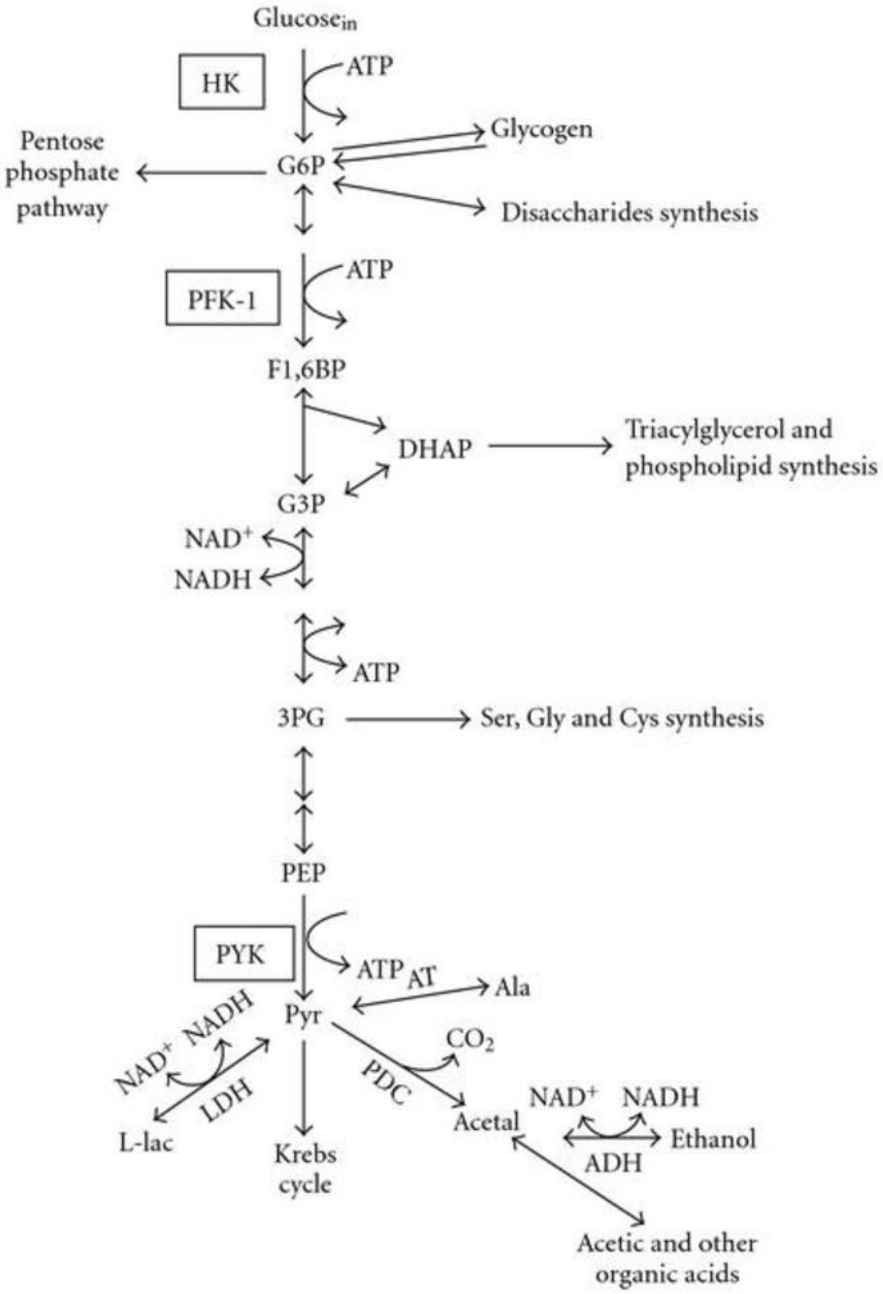


Figure 1.2: The glycolysis pathway (Moreno-Sanchez et al, 2008). This primary metabolic pathway is responsible for converting sugars into energy to be stored in cells. The arrows represent biochemical reactions that take place within a cell.

1.6.1. Environmental metabolomics

Environmental metabolomics is the application of metabolomics to study how living organisms interact with their environment. This application of metabolomics can occur in the natural environment of the target organism, or more commonly be applied under controlled laboratory conditions (Morrison et al, 2007).

Environmental metabolomics is often applied in the context of environmental toxicology. Environmental toxicology is a multi-disciplinary field involving biology, chemistry, environmental sciences as well as computational and mathematical disciplines. The goal is to measure the effect that harmful toxic chemicals have on biological organisms (Ankley et al, 2007) for the purpose of environmental monitoring or ecological risk assessment (Bundy et al, 2009). The field has seen substantial interest in recent years as stricter environmental protection regulations and subsequent testing methods have been developed and enforced (McCarty, 2012; McCarty, 2013). Aquatic environmental and ecological toxicology studies are deemed vital to this effort and have thus seen significant research activity (Brooks et al, 2016; Cedergreen, 2014; Hodson et al, 2007; McCarty et al, 2013; Sumpter & Jobling, 2013; Valavanidis et al, 2006).

An environmental toxicology study usually involves the comparison of treated and untreated, or control, samples using one or many analytical techniques or platforms. Commonly these studies involve investigating the effects of human induced environmental stressors such as pesticides (Pestana et al, 2010), insecticides (Jansen et al, 2015), fertilisers (Pivato et al, 2016) and various by-products of industry such as heavy metals e.g. lead (Offem & Ayotunde, 2008), mercury (Tsui & Wang, 2006) and cadmium (Qu et al, 2013). Model organisms are commonly used in environmental toxicology studies.

1.6.2. Analytical platforms for metabolomics

Metabolites are typically very low in molecular weight. As a result, in order to measure and characterise an organism's metabolome, highly sensitive analytical techniques are required for metabolomics measurements (Lenz & Wilson, 2007). There are two principal technologies for making metabolomics measurements, Mass Spectrometry (MS) and

Nuclear Magnetic Resonance (NMR). NMR is very high throughput technology and has the major advantage of analyses being very reproducible amongst laboratories (Viant et al, 2009). NMR however is not very sensitive when compared to MS, meaning that low concentration metabolites are difficult to detect. This drawback can limit the suitability of using NMR for some metabolomics studies.

MS is a high throughput technology and is extremely sensitive. The high sensitivity allows for low-concentration metabolites in a given biological sample to be measured in large numbers (Lenz & Wilson, 2007), making MS analysis the principal analytical platform for both general metabolomics studies (Gowda & Djukovic, 2014) as well as environmental metabolomics studies (Viant & Sommer, 2013). MS involves measuring the mass/charge (m/z) ratios of ionized metabolites. The output of a MS analysis is a spectrum of m/z values, with the relative abundancies of each feature (or metabolite) in the spectrum.

A number of different type of MS based analytical platforms exist. Direct Infusion MS (DIMS) involves injecting samples directly into the ion source of a MS. DIMS has the advantage that it is extremely high throughput, and requires less biological mass per sample. The disadvantage of DIMS is that due to there being no chromatography, ion suppression and peak overlapping is likely to occur. Ion suppression is where more easily ionised metabolites reduce the ability of less ionisable metabolites from being ionised as they enter the MS. This results in the prevention of less ionisable metabolites from being detected and can also result in detected metabolites having a lower quantification accuracy (Hop et al, 2005).

Chromatographic pre separation can be used to better separate peaks in a MS spectra and reduce peak overlapping. This has the advantage of superior quantification and identification of metabolites when compared to DIMS, but with the disadvantage of increased analysis times (Allwood & Goodacre, 2010). Two principal chromatographic separation technologies exist, Gas-Chromatography (GC) and High-performance Liquid Chromatography (HPLC). GC-MS provides the best separation but with the disadvantage that that chemical derivatization of metabolites is required before analysis (Dunn et al, 2008; Fiehn et al, 2000). HPLC analyses allow for a greater separation of a wider range of biological compounds, but it is a slower chromatographic technique than GC, and tends to result in lower peak resolution (Rohrs, 2006). HPLC is also more prone to ion suppression than GC. HPLC-MS is the most commonly used MS based analytical technique for metabolomics experiments as it is capable of good separation, is relatively rapid and requires less complex sample preparation (Gowda & Djukovic, 2014). All MS analysis presented in this thesis makes use of HPLC-MS analytical platforms.

Regardless of the MS platform or the type of sample being studied, a MS method (or assay) must be used. Developing these methods is both time consuming and challenging. A HPLC-MS method consists of a set of values for a number of MS and LC instrument control parameters. MS parameters exist to control a number of voltages, temperatures and pressures associated with the ionisation processes. Generally, these parameters effect the sensitivity of the MS method, the metabolite adducts that are produced and ion suppression (Unger et al, 2013). LC parameters control the gradient of mobile phases applied passed through the stationary phase over the course of the run, as well as the column temperature and flow rate. These LC parameters mainly effect the separation of

compounds but also effect peak height, sensitivity and indirectly effect ion suppression (Meyer, 2013).

HPLC-MS method development requires a high level of expertise, varying the LC and MS parameters systematically to optimise the analysis of a particular sample is impossible due to the sheer number of possible combinations of parameter values, and the time it takes to evaluate them. Typically, method optimisation is a manually performed task undertaken by an expert analytical chemist with high levels of expertise knowledge and experience in the analytical platform being used.

1.6.3. Statistical analysis of metabolomics data sets

Metabolomics data sets are multivariate, with each peak being treated as a variable. Metabolomics measurement typically produce peak tables in the form of a matrix $\boldsymbol{X}$ containing peak intensities**,** with N rows representing observations or samples and K variables representing detected peaks (Beckonert et al, 2007; Wold et al, 2001). Unsupervised dimensionality reduction techniques such as principal components analysis (PCA) are often used to decompose the matrix $\boldsymbol{X}$ to identify differences between classes and is widely used in metabolomics studies (Worley & Powers, 2013).

PCA performs a linear transformation of the matrix $\boldsymbol{X}$ into a lower dimension that preserves as much variance as possible from the original data (Jolliffe, 2002). PCA generates a set of principal components (PCs) which each describe a proportion of the datasets variance. The first PC, PC1 accounts for the highest amount of variance, with each subsequent PC accounting for a smaller amount of variance. Each sample data point has a vector V of n dimensions assigned to it, where n is the total number of PCs. The

score for each element of the vector corresponds to how much each sample reflects the variation described by the associated PC (Robertson, 2005).

A PC scores plot plots *N* samples based on the values in the PC scores for each sample in the vector *V*. These plots allow for the similarities between samples to be visually observed. It is preferable for the samples in each class to be clustered with a clear inter-class separation on the scores plots. The clusters allow for samples with comparable or disparate metabolome responses to be identified. These differences in metabolome responses can be indicative of a common response (Keun, 2006) i.e. to a given treatment in a toxicology study.

Metabolomics datasets can also be interrogated using univariate statistical methods to interrogate pairwise differences between two classes in the data set and identify the individual peaks that contribute to these differences. Commonly applied tests include student's t-tests, ANOVA and fold change analysis (Vinaixa et al, 2012).

1.6.4. Metabolite annotation

Metabolite annotation is the process of assigning chemical formulas and thus chemical identities to MS spectra. The annotation of metabolites is crucial to extracting biological meaning and interpretation from analytical metabolomics data sets (Creek et al, 2014). Metabolite annotation is far from trivial however and is a major bottleneck in metabolomics research.

Many metabolites can have similar chemical structures, the same molecular formula and the same mass-to-charge ratio, making it difficult to confidently annotate a given MS signal. For example, the sugars Glucose and Fructose have the same molecular formula ($C_6H_{12}O_6$) and therefore the same monoisotopic mass (180.063385 Da), with similar but

not identical chemical structures (Figure 1.3). Such similarities between measured metabolites can result in multiple structurally similar metabolites being assigned to MS peaks. Structurally similar metabolites can have a different metabolic and biological functions, so identifying them uniquely is important to accurately interpret any MS datasets biologically.

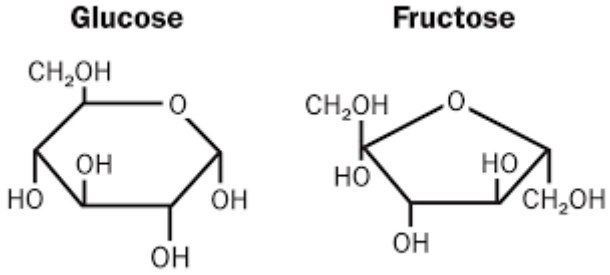


Figure 1.3: The sugars Glucose and Fructose have similar but different structures but share the same mass and chemical formula.

A typical untargeted MS study can contain hundreds or thousands of metabolites that all have varying concentration levels. This coupled to the fact that the expected metabolites in a given sample is unknown further complicates the annotation process (Dunn et al, 2013).

A four level system for assigning confidence to metabolite annotations exists (Sumner et al, 2007), with the highest annotation confidence level, level 1, being labelled as confidently identified compounds. Level 2 confidence is known as putatively annotated compound, and level 3 as putatively annotated compound classes, Level 4 compounds are unknown. Level 1 confidence is difficult to achieve, and level 2 putative annotation is a minimum level of confidence needed for MS peaks to be considered for biological interpretation of metabolomics data sets (Dunn et al, 2013).

A number of software packages exist for automated annotation of metabolites in LC-MS data sets. MI-Pack (Weber & Viant, 2010), CAMERA (Kuhl et al, 2012),

PeakML/mzMatch (Scheltema et al, 2011), PUTMEDID-LCMS (Brown et al, 2011) and IDEOM (Creek et al, 2012) are some examples of freely available automated metabolite annotation software. All of these packages work in a similar manner. The MS spectra are analysed to find correlations between them. Because of how an MS system ionises compounds as they enter the instrument, many MS features or responses can relate to the same metabolite or compound, these are known as adducts. These adduct features should correlate in terms of retention time similarity, accurate m/z values within a permissible range and chromatographic peak shape (Dunn et al, 2013). These correlations can be used to assign molecular formula to MS features with increased confidence. Once molecular formula are identified, databases such as KEGG (Kanehisa et al, 2000), LipidMaps (Sud et al, 2007), HMDB (Wishart et al, 2012) and PubChem (Kim et al, 2016) are queried to see if there are any compounds with matching formulas.

1.7. Genome-wide metabolic reconstruction

Genome-Wide Metabolic Reconstructions (GWMRs) have the potential for use in environmental metabolomics based computational toxicology (Kesari, 2017). However, to date very little work has been published in this area (Blais et al, 2017; Kotera & Goto, 2016; Topfer et al, 2015). GWMR is an *in-silico* modelling technique that seeks to represent the metabolic capabilities of an organism at a genomic scale. Reconstructed networks are a powerful tool that can prove extremely useful for linking experimental data with computational systems biology (Feist et al, 2009). GWMRs also provide a platform for analysis, visualisation and contextualisation of high throughput omics data sets (Francke et al, 2005), can help in understanding the global properties of metabolic networks (Ma & Zeng, 2003), and can guide metabolic engineering and hypothesis driven discovery (Oberhardt et al, 2009).

GWMRs are built using a bottom-up approach using an organism's genome sequence to infer enzymatic metabolic reactions. GWMRs represent metabolic networks as a bipartite graph of nodes and edges, with reactions and metabolites each being represented as a different class of node, and edges linking reactions to the metabolites participating in them. Each reaction is linked to one or many catalysing enzymes which are in turn linked to one or many encoding genes (Figure 1.4).

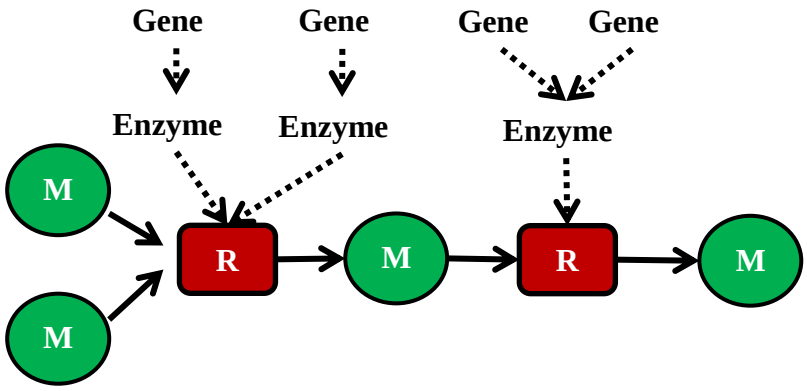


Figure 1.4: Relationship between genes, enzymes, reactions and metabolites in a metabolic reconstruction. Each reaction (R) is linked to one or many enzymes, which are in turn inferred from genes. Metabolites (M) participate in each reaction and are represented as a different class of node. Metabolites that are substrates or products to reactions are linked with an edge.

The procedure for generating a GWMR is well defined by (Thiele & Palsson, 2010). Their highly detailed 96 step protocol can be broadly categorised into two stages; draft reconstruction and model curation. Draft reconstruction involves collecting a large set of biochemical reactions that are encoded in a genome sequence via enzymes and assembling them into a network along with the metabolites that participate in the reactions. Model curation involves manually checking the draft reconstruction to assure model accuracy.

1.7.1. Constraint-based modelling

Constraint-based modelling is a technique that applies physiochemical constraints or limitations on a GWMR to describe possible behaviours of the target organism (Moreno-Sanchez et al, 2008; Ramakrishna et al, 2001). Examples of constraints include; flux limitations, mass balance and energy balance. Constraint-based modelling assumes that the target organism can reach a steady state that satisfies the constraints for a specified environmental condition. Numerous steady-state solutions are conceivable, as the entire constraints on a system can never be fully known. Therefore, an optimisation is performed to find the optimal value for a given objective function that relates to the constraints, thus finding a physiologically meaningful steady state solution (Segre et al, 2002).

Flux balance analysis (FBA) is a constraint-based modelling technique that uses the stoichiometry of the reactions within a GWMR to constrain possible solutions and to analyse the flow of metabolites through a GWMR. Linear programming is used to calculate optimal solutions with respect to an objective function (Edwards et al, 2002). The relationship amongst metabolite concentrations, x, and reaction activities, v, is described by the dynamic mass balance equation:

$$\frac{dx}{dt} = S.v$$

S is the stoichiometric matrix, which is formed from the stoichiometric coefficients of the biochemical reactions that make up the GWMR. Each column of S corresponds to a reaction, and each row corresponds to a metabolite. The values in the matrix are stoichiometric coefficients, which are always integers. Each column of the S matrix describes a reaction and is constrained by its elemental balancing. Each row describes the set of reactions that the corresponding metabolite participates in, and also describes how

the reactions are interconnected. v is a vector of reaction fluxes, or reaction activities. Under steady-state conditions, the above equation becomes:

$$S.v = 0$$

FBA seeks to find a set of steady-state values for the vector v by defining an objective function, useing linear programming to find the optimal set of fluxes through the GWMR for the objective function (Segre et al, 2002)

A commonly used objective function is the biomass objective function, which aims to find a flux distribution that results in the biomass precursors of the target organism being created through the metabolic reactions in the network in the correct proportions (Feist & Palsson, 2010). The biomass objective function can be used to predict cellular growth of an organism when placed in a particular growth medium (Orth et al, 2010). Examples of other FBA objective functions are: The minimisation of ATP production (Ramakrishna et al, 2001; Vo et al, 2004), the minimisation of a particular nutrient uptake (Famili et al, 2003) and the maximisation of the production of particular metabolite(s) (Varma et al, 1993).

Steady-state flux distributions found with FBA can give insight into how the organism of interest responds under certain environmental conditions. If an unexpected result is obtained from FBA, i.e. predicted cellular growth under strictly controlled conditions is not as expected, it can provide insights into the accuracy of the GWMR and highlight potential errors in the network. It is important to note however that to accurately model cellular growth of an organism in a particular medium, the metabolic reactions describing the uptake of nutrients from the medium must be defined and included within the network (Cuevas et al, 2016). This is straightforward for organisms such as *Saccharomyces*

cerevisiae, whose uptake of glucose are typically modelled (Garcia Sanchez et al, 2012). FBA becomes much more challenging for organisms whose growth media are more complex, e.g. aquatic organisms such as *Danio rerio* (Bekaert, 2012).

1.7.2. Automated draft GWMR

A number of platforms exist for generating draft GWMRs in an automated way (Devoid et al, 2013; Karp et al, 2016; Moretti et al, 2016; Pinney et al, 2005; Swainston et al, 2011; Wrzodek et al, 2011), but usually they require either specific genome annotations or the target organism to have some presence in a genome or pathway database. For organisms with newly sequenced genomes, this is not always the case. Therefore, the use of automated tools for generating GWMRs for newly sequenced genomes can be problematic. For example, the Pathologic tool (Karp et al, 2011) requires genome annotations in the GenBank format (Benson et al, 2008), a NCBI maintained database of nucleotide sequences. Each sequence must be submitted to GenBank for inspection to generate a GenBank file. This extra process limits the suitability of Pathologic for generating GWMRs of new genome sequences when there is no immediate access to an appropriate GenBank file. Model SEED is a web based tool for automated GWMR based on genome annotations performed using the RAST algorithm (Overbeek et al, 2014). Model SEED can generate GWMRs for RAST genome annotations that are available on the Model SEED web page/database. Users can also upload their own sequences as FASTA files, but only plant and microbe sequences are accepted.

1.7.3. Linking transcriptomics data to GWMRs

Linking transcriptomics data to GWMRs is an area of research that has seen some attention recently. Transcriptomics datasets can be integrated with GWMRs, by using the

data to score nodes in the network. Transcriptomics data sets reveal genes which are over or under expressed, these genes can subsequently be linked to reactions in a GWMR (Figure 1.4). The active module identification approach, is a generalised method for searching a network to find connected sets of nodes that are deemed to be highly active under a certain condition (Ideker et al, 2002). The approach was originally used with protein interaction networks to find active modules representing connected sets of genes with higher levels of differential expression than the overall network, and has been successfully applied to metabolic networks (Bryant et al, 2013b; Cho et al, 2014; Deo et al, 2010; Wang et al, 2013; Wang et al, 2014).

The active modules approach can be used to detect actively changing areas of the metabolic network effected by an external influence. Studies have used this approach to better understand known organism responses to chemicals (Bryant et al, 2013b; Cho et al, 2014; Deo et al, 2010), or to optimise production of industrially important metabolites (Wang et al, 2013; Wang et al, 2014). No work has used this approach to predict unknown organism response to chemical perturbation.

Applying a computational toxicology approach to environmental metabolomics using GWMRs to predict how an organism will respond metabolically to an environmental stressor is of interest as it would allow for organism response to environmental perturbations to be hypothesised *in-silico*. The requirements for this would be a GWMR of the target organism, and a toxicogenomic transcriptomics dataset describing the transcriptional response of the target organism to some environmental stressor.

1.8. Daphnia

Species of the genus *Daphnia* (often referred to as a water flea) are small aquatic crustaceans that inhabit many types of freshwater ecosystems such as ponds and lakes (Ebert, 2005; Lampert & Kinne, 2011), the most common species being *Daphnia magna* and *Daphnia pulex*. *Daphnia* is an extremely sensitive species within these ecosystems and as a result has been widely used as a model species for environmental toxicology studies (Iampolskii & Galimov Ia, 2005; Jansen et al, 2015; Lampert & Kinne, 2011; Martins et al, 2007), ecogenomic studies (Eads et al, 2008; Miner et al, 2012; Orsini et al, 2012), and evaluating the impact of environmental change (Martins et al, 2007; Shaw et al, 2008). *Daphnia* are widely used as indicators of water quality and environmental health, and are also key models in evolutionary biology and the study of adaptive responses to environmental change (Frisch et al, 2014).

Figure 1.5: Adult female Daphnia magna with eggs in its brood chamber.

Daphnia has an adaptable life cycle (Figure 1.6) that provides interesting mechanisms for coping with environmental changes. If there are no environmental stresses, Daphnia reproduce asexually and parthenogenetically. This means that all offspring are female and are genetically identical, which is extremely beneficial trait for conducting omics studies. Reproduction mechanisms can be altered in the presence of environmental stresses such as predation or chemical stress, sexual reproduction can occur which produces eggs which are in a protective structure called an ephippium, capable of surviving hundreds of years (Carvalho & Hughes, 1983; Doma, 1979; Frisch et al, 2014).

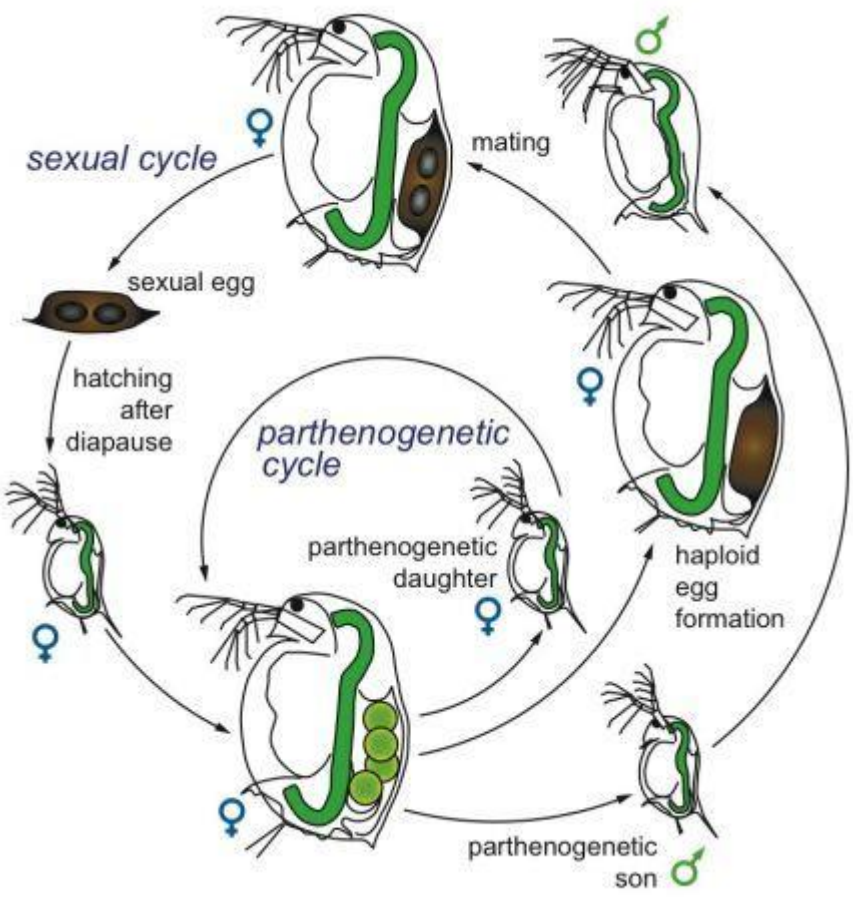


Figure 1.6: *Daphnia* life cycle (Ebert, 2005).

Daphnia has been used in a large number of toxicology omics studies (Altshuler et al, 2011), including transcriptomics (Campos et al, 2013; David et al, 2011; Garcia-Reyero

et al, 2009; Orsini et al, 2016; Poynton et al, 2012; Rivetti et al, 2015), proteomics (Le et al, 2013; Otte et al, 2014; Rainville et al, 2014) and metabolomics (Bunescu et al, 2010; Nagato et al, 2013; Poynton et al, 2011; Taylor et al, 2008; Taylor et al, 2010).

Daphnia's is an important species in ecotoxicology, ecology and environmental studies. Its short parthogenetic life cycle that can develop genetically identical offspring is clearly an asset for a range of omics fields. It is no surprise that *Daphnia* is a quickly becoming a leading model species for environmental omics research, and *Daphnia* is an ideal candidate for environmental computational toxicology.

2. Research Objectives

Daphnia is an extremely sensitive species in freshwater ecosystems and is widely used as a model for ecotoxicological studies (Iampolskii & Galimov Ia, 2005; Jansen et al, 2015; Lampert & Kinne, 2011; Martins et al, 2007; Shaw et al, 2008) and is therefore a key model species for evaluating ecological impact of environmental change. Increasingly, Daphnia is used as a surrogate species to understand genomic responses to environmental stressors that are important factors in human health and wellbeing, and is an emergent ecological model species (Harris et al, 2012; Stollewerk, 2010). The National Institutes of Health list Daphnia as one of 13 key model organisms for biomedical research (Ebert, 2011).

Omics science has transformed biological science into data rich discipline. A number of opportunities exist for computational methods to take advantage of this and can be applied to gain insight from these datasets or to construct *in-silico* models. *Daphnia* is clearly an important model species in environmental research, and there is still a lot to understand at a metabolomics and systems biology level. There is a clear case for using *Daphnia* as a target organism in an environmental computational toxicology context. GWMRs can be used in such a setting to gain insight into an organism's metabolic response to a chemical perturbation (see Section 1.7). Daphnia is an established organism for omics science, and the required datasets for this type of approach; genome assemblies and transcriptomics datasets have recently become available (Colbourne et al, 2011; Orsini et al, 2016).

This thesis will investigate the use of GWMRs as predictive models using a computational environmental toxicology approach. I*n-silico* computationally-generated hypothesis of the unknown metabolic response of Daphnia to environmental stressors will be generated

followed by an attempt to validate them biologically. To achieve this, a workflow consisting of a number of distinct steps is developed. Figure 2.1 shows what this workflow will look like. The workflow is designed so that it can be used with any organism of interest provided there is a genome sequence and transcriptomic data set available and is also designed to be as automated as possible. The ultimate goal of this approach is to be able to make *in-silico* predictions of an organism's metabolic response based upon a transcriptomics study.

The overall aim of this thesis can be broken down into the following sub aims:

- The development of a computational tool for the draft GWMR for organisms with newly sequenced genomes (Chapter 3).
- The use of this tool to generate a draft Genome-Wide Metabolic Reconstruction of *Daphnia magna* (Chapter 4).
- Computational hypothesis generation of *D. manga's* metabolic response to environmental stressors using the draft GWMR (Chapter 5).
- A computational tool for automated LC-MS method development, and its subsequent use to develop an LC-MS assay that will detect as many of the predicted metabolites as possible (Chapter 6).
- Traditional metabolomics study to validate these hypotheses (Chapter 7).

Omics science has shifted the biological sciences paradigm. Traditionally research was hypothesis testing, with the ability to measure vastly more entities at the same time using omics, this has enabled data-driven hypothesis generation. A criticism of this approach has seen omics science being compared to a fishing expedition, or blindly looking for interesting features in the data (Ning & Lo, 2010). The approach proposed in this thesis

is to computationally generate hypotheses in an unbiased way using genomics and transcriptomics data to predict metabolic responses. This allows for a metabolomics dataset to become hypothesis testing whilst at the same time be hypotheses generating.

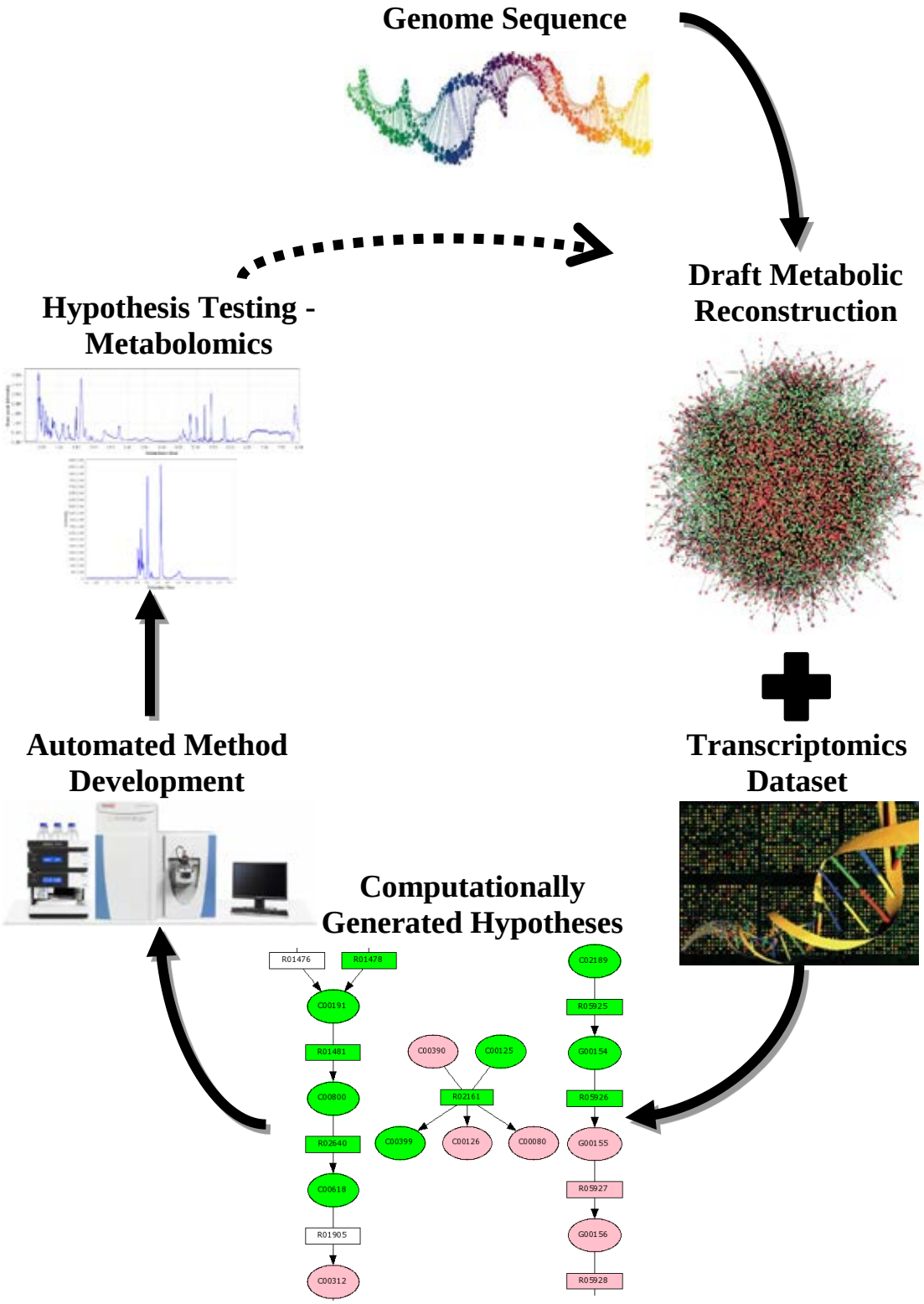


Figure 2.1: Workflow for computational hypothesis generation using GWMR and transcriptomics data. A genome sequence is used to construct a draft GWMR. Transcriptomics data is then integrated into this network and the active module approach is used to generate computationally generated hypotheses of metabolic response. The metabolites within these predictions can be used with a closed-loop optimisation approach to automatically develop an analytical method for metabolite measurement. The hypotheses can then be experimentally verified. This experimental data can feed back into the GWMR to improve the model and hence is predictive capability.

2.1. Thesis organisation

Figure 2.1 visualises the workflow that will be developed to address the research question posed. The first task is to build a draft Genome-Wide Metabolic Reconstruction (GWMR) of *Daphnia*. The genome sequences of *D. pulex* and *D. magna* are relatively new and thus *Daphnia* has very little presence in genome and pathway databases. A number of platforms exist for automated generation of draft GWMRs, however they are often not suitable for *new* genome sequences such as *Daphnia* (see section 1.7). In order to build a draft GWMR of *Daphnia* a software package, METRONOME, is developed for automated draft GWMR for new genome sequences which is documented in Chapter 3. Chapter 4 details the draft GWMR of *D. magna* using METRONOME and includes some analysis of the quality of the resulting network.

Chapter 5 describes the use of the active modules approach to generate computational hypotheses of how the metabolic response of *D. magna* is effected by two chemical perturbations. Transcriptomics data from a *D. magna* environmental genomic toxicology study is used along with the draft GWMR generated in Chapter 4 to generate these predictions. As the predicted metabolic responses generated in Chapter 5 are unknown, a metabolomics study is also performed to attempt to validate the predictions. Chapter 6 describes the software package MUSCLE, which performs closed-loop multi-objective evolutionary optimisation of LC-MS analyses. MUSCLE is then used to optimise a HLPC-MS method that is used in the metabolomics study. Chapter 7 describes the experimental design, data processing and statistical interpretation of the metabolomics study used to validate the computationally generated hypotheses from Chapter 5.

3. METRONOME: METabolic Reconstruction Of New genOMe sEquences

Genome-wide metabolic reconstructions (GWMRs) model the metabolic capabilities of a target organism by representing biochemical reactions and metabolites in a network of connected nodes. GWMRs are a potent tool that can prove extremely useful for linking experimental data with computational systems biology and for computational hypothesis generation of metabolic response (Chapter 5).

The highly detailed process for generating a GWMR is well defined. The first part of the process can be automated and a number of tools exist for this purpose. These tools however are limited when used to generate GWMR s for newly sequenced genome sequences, as they require detailed genome annotations or data curation.

Here the flexible METRONOME platform is introduced for automated reconstruction of metabolic networks for new genome sequences that addresses these limitations. METRONOME is capable of finding enzyme encoding genes in a genome sequence and infer metabolic reactions without the need for the target organism to have a presence in genome/pathway databases.

METRONOMEs effectiveness is demonstrated by comparing its performance at building draft GWMRs of model organisms with two popular automated tools which are not very suitable for use with new genome sequences.

3.1. Introduction

GWMRs seek to represent the metabolic capabilities of an organism at a genomic scale and are built using a bottom-up approach, using an organism’s genome sequence to infer

enzymatic metabolic reactions. Reconstructed networks are a powerful tool that can prove extremely useful for linking experimental data with computational systems biology (Feist et al, 2009). GWMRs also provide a platform for analysis, visualisation and contextualisation of high throughput omics data sets (Francke et al, 2005), can help in understanding the global properties of metabolic networks (Ma & Zeng, 2003), and can guide metabolic engineering and hypothesis driven discovery (Oberhardt et al, 2009).

The procedure for generating a GWMR is well defined by (Thiele & Palsson, 2010). Their highly detailed 96 step protocol can be broadly categorised into two stages 1) draft reconstruction; and 2) model curation. Draft reconstruction involves collecting a large set of biochemical reactions that are encoded in a genome sequence via enzyme encoding genes and assembling them into a network along with the metabolites that participate in the reactions. Model Curation involves manually checking the draft reconstruction to assure model accuracy.

A number of platforms (Devoid et al, 2013; Karp et al, 2016; Moretti et al, 2016; Pinney et al, 2005; Swainston et al, 2011; Wrzodek et al, 2011) exist for generating draft GWMRs in an automated way, but they all have limitations when being used with new genome sequences. Model SEED (Devoid et al, 2013) requires the organisms genome sequence to be included in their maintained database. The Pathway Tools software package (Karp et al, 2016) requires the target organism to have an annotated genome sequence to have been submitted to the GenBank database (Benson et al, 2008). MetaSHARK (Pinney et al, 2005), the SuBliMinaL (Swainston et al, 2011) toolbox and KEGGtranslator (Wrzodek et al, 2011) all require the target organism's annotated genome sequence to be included in the KEGG collection of databases (Kanehisa et al, 2004).

Generating draft reconstructions for organisms with newly sequenced genomes using these automated tools is therefore not straightforward. For example, *D. magna*, a model species in environmental toxicology studies, has had its genome sequenced recently, but it is not yet publicly available. As a result the organisms annotated genome sequence has not been submitted to the GenBank database (Benson et al, 2008) and there are no organism specific database entries for *D. magna* in any of the key genomics or biochemical reaction databases. This rules out using any of the above-mentioned tools for generating a draft GWMR. These limitations of the currently available automated draft GWMR tools severely impedes the ability to conduct research using GWMRs for organisms with newly sequenced genomes. Here the METRONOME platform for the automated generation of draft GWMRs for new genome sequences is presented. METRONOME is a modular platform consisting of three key sub-modules. First, enzymes are assigned to the genome sequence by assessing the similarity of the input sequence with known annotations (Section 3.2.1). Secondly, the assigned enzymes are used to mine multiple databases/data sources for associated biochemical reactions (Section 3.2.2). Thirdly, information from different genome/reaction databases are merged (Section 3.2.3). The effectiveness of the pipeline is demonstrated by reconstructing draft GWMRs for two model organisms using a variety of tools (Section 3.3).

3.2. Methods

Generating draft GWMRs involves transforming a genome sequence into a set of connected biochemical reactions. This is achieved by deciphering which genes encode enzymes. Enzymes act as catalysts to biochemical reactions which can be inferred to be present in the metabolome of the target organism. The METRONOME platform follows

this principle, and contains three main modules (Figure 3.1). The first module assigns enzymes to genes (Section 3.2.1). This module generates a set of enzyme-gene pairs, which can either be provided explicitly or generated from a genome sequence directly using the enzyme assignment sub-module. The second module is a biochemical reaction data mining module (Section 3.2.2). This module takes the enzyme-gene pairs from the enzyme assignment module as its input, and outputs a draft GWMR for each of the databases/sources that are selected to be data mined. The third module combines reactions and metabolites from multiple data sources into a single coherent network (Section 3.2.3). All networks are represented using the SBML format (Hucka et al, 2003), which is an open XML based file format for representing biological networks and is the most common format for representing metabolic reconstructions. METRONOME is written in Python and can either be run from the command line or using a GUI

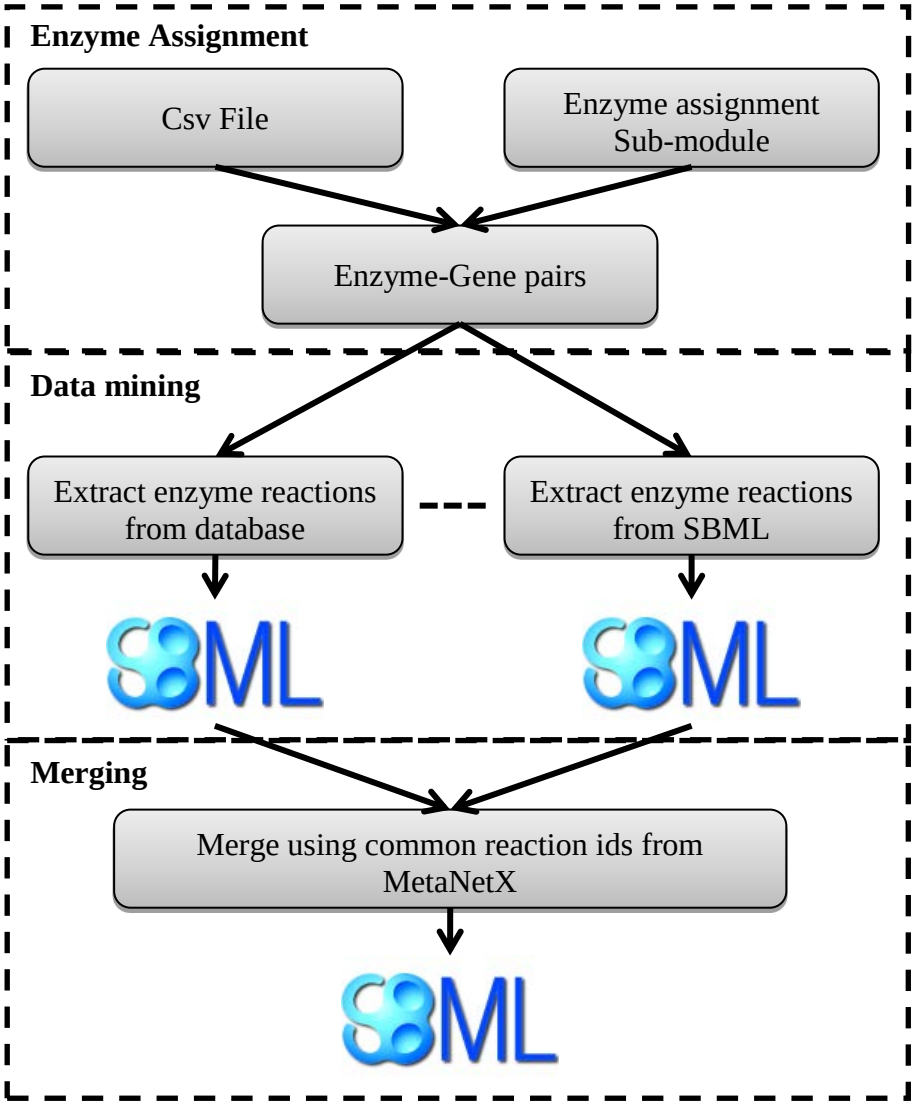


Figure 3.1: The METRONOME pipeline consists of three main modules, enzyme assignment, data mining and merging. The enzyme assignment module generates a list of enzyme-gene pairs which is used by the data mining module to extract biochemical reactions from either a database or a SBML file. Each extraction produces a SBML file, which is then merged using the MetaNetX reconciliation database.

3.2.1. Enzyme assignment module

The starting point of the draft reconstruction process using METRONOME is a list of gene and enzyme pairs, with enzymes being represented using the E.C. number nomenclature (Bairoch, 1994). The pairs can be directly provided as a CSV file, or can be generated using the enzyme assignment module. The enzyme assignment module is designed to be flexible so that any algorithm, tool or technique for assigning enzymes

(Claudel-Renard et al, 2003; Curtis et al, 2013; Devoid et al, 2013; Li et al, 2003; Romero et al, 2005; Waterhouse et al, 2013; Zhao et al, 2013) can be applied. In this implementation, the OthoMCL (Li et al, 2003) algorithm is incorporated as detailed below.

OrthoMCL based enzyme assignment

OrthoMCL (Li et al, 2003) is a widely used (>2,500 citations) algorithm for assigning orthologous groups across a wide range of eukaryotic organisms using a Markov Cluster algorithm. OrthoMCL takes a protein sequence of an organism and finds similar sequences in other well characterised organisms in a protein by protein basis. E.C. numbers can be assigned reliably, providing an automated genome annotation that works well for unannotated genomes.

An OrthoMCL enzyme assignment sub-module has been written that wraps up the algorithm and outputs a list of gene and enzyme pairs. The algorithm takes a sequence as a FASTA file, a text based representation of a nucleotide or peptide sequence, and returns files that link the input sequence to ortholog or paralog groups in an OrthoMCL namespace. The sub-module takes these output files and uses them to programmatically access the OrthoMCL REST web service to achieve the E.C. number assignment.

This returns an XML object that contains a list of E.C. numbers that are assigned to the OrthoMCL group. The E.C. numbers can then be linked with the relevant gene ids consequently providing the gene-id enzyme pairs that are required for the data mining module.

3.2.2. Data mining module

GWMRs are made up of nodes representing biochemical reactions and the metabolites that are associated with them. A number of databases exist for genome, pathway and metabolic information such as chemical structure, molecular weight and reaction stoichiometry. Each database has a different schema and contains some unique as well as common information (Altman et al, 2013). It is therefore beneficial to include as much information from the available sources as possible.

The METRONOME data-mining module is designed so that information from a number of sources can be used for draft GWMR generation. Most sources of data are hosted online and have APIs for querying and accessing data from them. In order to extract data from a given database, a wrapper must be implemented that defines the methodology for extracting reactions and metabolites. Three methods must be written for each wrapper, ExtractReaction, ExtractMetabolite and BuildSMBL. The ExtractReaction method takes as input an Enzyme-Gene pair and appends the reaction ids linked with the given E.C number. The reaction id appended will be native to the given database that is being queried. The ExtractMetabolite method takes a reaction id and finds the ids of any metabolite that acts as either a substrate or product of the given reaction. METRONOME makes extensive use of Python dictionary data structures to assign information during the reaction and metabolite extraction process, as each database represents information in a different way, the flexibility proves useful as it allows for a wide range of information to be stored, minimising the required number web service calls.

The BuildSBML method takes all of the reaction and metabolite information and uses it to construct an SBML object. This method is required for each sub-module, as the content of each database can vary greatly both in terms of content and representation.

SBML files contain lists of *Species* and *Reactions*. When representing metabolic networks in the SBML format, species represent metabolites, and reactions represent biochemical reactions. A reaction SBML object contains fields called *listOfReactants* and *listOfProducts*, which contain references to the relevant species objects. Reactants and products can be seen as input and output metabolites for a reaction respectively and each species can be a reactant and product in many reactions. Both species and reaction SBML objects contain a field called *Notes* which contains metadata. The Notes field is used to store a large range of useful information, including database references (for both Species and Reactions), enzyme and gene ids (for Reactions), chemical formulas, and structural chemical information using the SMILES (Weininger, 1988) and InChI (Heller et al, 2015) formats (for Species).

Sub-modules can be written for extraction from any data source, but they can only be merged (Section 3.2.3) if that data source is represented in the MetaNetX database (Moretti et al, 2016). Table 3.1 contains the databases that are represented in MetaNetX. Some of these databases can be downloaded as flat files. Consequently sub-modules can be written that extract the reactions and metabolites from these files directly. This removes the need for using web-service calls but has the disadvantage that the download files can be too large to download.

SBML Extraction

METRONOME also provides a mechanism for extracting biochemical reactions from SBML files directly. SBML extraction is useful as some databases can be downloaded directly as SBML files, or previously constructed GWMRs can be used as a basis for a new draft GWMR. SBML extraction requires the same three methods to be defined. As

no web-service calls are needed using this method, extraction of reactions as metabolites can be performed quickly without the need to download large database flat files.

KEGG Extraction Sub-Module

METRONOME provides a sub-module for extracting biochemical reactions from the KEGG collection of databases (Kanehisa et al, 2014). KEGG is a comprehensive collection of databases that contain resources for understanding biological systems with databases for genomes, biological pathways, enzymes, reactions, chemical substances and much more. KEGG provides a REST API for accessing resources from any of the collection of databases, with a flat file database format or a tab delimited file being returned for each record.

The KEGG extraction sub-module contains a parent class KEGGObject with sub-classes for each of the KEGG databases that are needed KEGGEnzyme, KEGGReaction and KEGGCompound. Each KEGGObject contains the flat file database file that is returned using the KEGG REST API. Each sub-class of KEGGObject contains some unique information such as substrate/product compounds for reactions and chemical formulas for compounds.

The ExtractReactions method first extracts a KEGGEnzyme object by passing an E.C. number as an argument. The object is then examined for reaction ids, which are subsequently used to extract KEGGReaction objects. The Gene Id and E.C. numbers are stored in the KEGGReaction object before it is examined for compound ids. The ExtractMetabolites method is then called using these compound ids to extract KEGGCompound objects, with references to these objects being stored in the relevant KEGGReaction object. If at any point a duplicate E.C. number is passed, the reactions

associated with will have the Gene ID field updated to include the current gene id to avoid time-consuming web service calls.

```
1: function KEGG EXTRACT(Gene-Enzyme pairs[])
2:     for each Gene-Enzyme pair p in Gene-Enzyme pairs do
3:         Extract KEGGEnzyme object
4:         for each Reaction in KEGGEnzyme.Reactions do
5:             Extract KEGGReaction object
6:             Store Gene ID in KEGGReaction.GeneID
7:             Store E.C. Number in KEGGReaction.Enzyme
8:             for each Compound in KEGGReaction.Substrates do
9:                 Extract KEGGCompound
10:                Store KEGGCompound object(s) in KEGGReaction.Substrates
11:            end for
12:            for each Compound in KEGGReaction.Products do
13:                Extract KEGGCompound
14:                Store KEGGCompound object(s) in KEGGReaction.Products
15:            end for
16:        end for
17:    end for
18: end function
```

Algorithm: 3.1: KEGG extraction procedure.

The BuildSBML method takes all of the KEGGReaction and KEGGCompound object and constructs a SBML object.

MetaCyc Extraction Sub-Module

A sub-module for extracting biochemical reactions from the MetaCyc (Caspi et al, 2014) database has also been written. MetaCyc is a database of metabolic pathways that covers all domains of life and contains a number of highly curated organism specific biological pathway resources. The SBML extraction mechanism has been used here as a SBML file containing the entire MetaCyc database can be easily exported using Pathway Tools, a free (for academic users) software package written and maintained by the MetaCyc team (Karp et al, 2016).

The ExtractReactions method takes an E.C. number, iterates through all of the reactions in the MetaCyc SBML file and examines the metadata stored in the reactions notes field.

If the notes field contains the corresponding E.C. number, the reaction id is stored, along with the ids of the substrate and product metabolites. The BuildSBML method then takes the stored reaction and metabolite ids and constructs a SBML file by copying the relevant SBML objects to a new SBML file and adding the appropriate gene ids to each of the reactions.

3.2.3. Network merging module

Each data mining sub-module extracts biochemical reactions from a given source and generates a SMBL file containing a metabolic network. There is a large amount of overlap between various reaction and metabolite databases, but often databases either lack cross references, or contain duplicate or incomplete information. MetaNetX (Moretti et al, 2016) is a repository that has reconciled a number of resources into a common namespace.

3.2.3.1. MetaNetX metabolite and reaction reconciliation

A large number of databases have been reconciled (Table 3.1) by comparing chemical structure, shared nomenclature, cross-references and reaction context (Bernard et al, 2014).

MetaNetX reconciliation of metabolites and reactions is achieved through several techniques. Metabolites are first reconciled based on their chemical structures. This is achieved by comparing two standardised string representations of chemical structures, SMILES (Weininger, 1988) and InChI (Heller et al, 2015), and merging where appropriate. In the case of stereoisomers, all reactions where the metabolite is present are inspected. If different reactions include different stereoisomers of a metabolite, then the reconciliation process assumes that they are biologically distinct and does not merge the metabolites. The second reconciliation step uses string-matching algorithms to merge

metabolites that have shared nomenclature. This is used as for some metabolite databases structural information is not present. Merging metabolite entries based on names can be problematic as many synonyms exist for metabolites and their classes between and sometime within different databases. Therefore, the MetaNetX reconciliation process only merges metabolites if their names exactly match.

Reactions are first reconciled by looking at their participating metabolites and their stoichiometry and are merged if there is a match. Reactions are then reconciled based on shared cross-references from the source databases (Table 3.1). An iterative procedure of reconciliation of metabolites is then performed by looking at their reaction context. Reactions that share at least a single metabolite or cross reference are inspected. If two reactions share several reconciled metabolites, but some of the other metabolites in each of the reactions are unreconciled, the remaining metabolites are considered to possibly be the same. A rule based system is then applied to reconcile the metabolites if enough evidence, such as chemical formula or charge is present (Bernard et al, 2014).

The MetaNetX reconciliation process is extensive and is proven to perform well (Bernard et al, 2014; Moretti et al, 2016). There are however some instances where MetaNetX is unable to reconcile metobiltes correctly, such as where metabolites are present in a database, but are not well represented in reactions, or where chemical structure information is unavailable. However, by using MetaNetX to merge the networks, METRONOME is able to assign a large number of database identifiers to species and reaction metadata fields in the SBML file, as well as avoiding duplicate information in the draft GWMR.

MetaNetX consists of a number of web-accessible tab delimited files which are loaded into memory at the start of the METRONOME merging process. The merging module looks at the SMBL files from each data mining sub-module and only adds each reaction or metabolite if it is not already in the network based on its MetaNetX id. This has the effect of consolidating the information contained within the SBML files.

Table 3.1: Databases represented in MetaNetX.

Database	**Description**	**Reference**
BiGG	Knowledgebase of over 70 published GWMR with standardised identifiers. Genomes are mapped to NCBI genome annotations and metabolites contain cross-references with KEGG, ChEBI, PubChem and more.	(Schellenberger et al, 2010)
BioPath	Database of biochemical pathways that is based on the Roche *Biochemical Pathways* wall chart as well as metabolic reactions that have been reported in primary literature.	(Forster et al, 2002)
ChEBI	Database/dictionary of molecules with a focus on small, or low-weight, compounds	(Hastings et al, 2012)
HMDB	Database containing detailed information about metabolites found in the human body.	(Wishart et al, 2012)
KEGG	Collection of databases that form a resource for interpreting genome sequence data. Databases exist for genomes, biological pathways, enzymes, chemical substances and many more.	(Kanehisa et al, 2014)
LIPIDMAPS	Database that contains information about lipid species measured in mammalian cells.	(Sud et al, 2012)
MetaCyc	Database of metabolic pathways that cover all domains of life. A number of highly curated organism specific database exist that all use the Cyc postfix i.e. EcoCyc and YeastCyc.	(Caspi et al, 2014)
Reactome	Pathway database that contains many cross-references.	(Croft et al, 2014)
Rhea	Manually annotated database of biochemical reactions.	(Morgat et al, 2015)
The SEED	Database of genome annotations that also includes genome linked biochemical reactions.	(Overbeek et al, 2014)
UniProt	Database of protein sequences and protein functional information.	(Consortium, 2015)

3.2.4. Output

A number of files are generated by METRONOME. Intermediate SBML files are generated for each data mining process run, as well as a merged SBML file which represents the union of the SBML files. It is this merged SBML file that represents the final draft GWMR.

Two csv files are also generated that contain lists of all of the reactions and metabolites along with the chemical formulas and structures, and all of the reconciled database identifiers generated using the merge module. These files provide a useful resource for inspecting the reactions and metabolites that are contained within the draft GWMR.

3.3. Results

In order to assess the effectiveness of the METRONOME pipeline, draft GWMRs are constructed for the model species *Escherichia coli* and *Saccharomyces cerevisiae*. Although the METRONOME is specifically designed for new genome sequences, these model species are chosen because they have well curated GWMRs that can be used as a benchmark to test the effectiveness of the pipeline. Two other automated tools for draft reconstruction are used to generate GWMRs of *E. coli* and *S. cerevisiae*; Pathologic (Karp et al, 2011) and Model SEED (Devoid et al, 2013).

Pathologic is the algorithm that is used by the Pathway Tools software (Karp et al, 2016) and predicts metabolic pathways by inferring enzymatic reactions based on an annotated genome sequence and comparing to previously known information from the MetaCyc database (Caspi et al, 2014). Pathologic requires genome annotations in the GenBank format (Benson et al, 2008). GenBank is a NCBI maintained database of nucleotide sequences which need to be submitted to GenBank for inspection in order to generate the

required files. As a result, Pathologic can be unsuitable for generating GWMRs of new genome sequences due to the effort involved in generating a GenBank file.

Model SEED is a web based tool for automated GWMR based on genome annotations performed using the RAST algorithm (Overbeek et al, 2014). Model SEED is only able to generate GWMRs for RAST genome annotations that are available on the Model SEED web page/database, or based on user uploaded FASTA files, for plant and microbe species. It is therefore not a suitable tool for generating GWMRs for new genome sequences that are not plants or microbes.

Draft GWMRs for both *E. coli* and *S. cerevisiae* are built using METRONOME, Pathologic and Model SEED and the reactions contained in the resulting SBML files are compared with well curated GWMRs/database information about the species. For the GWMRs generated with Pathologic, the default parameters are used: Taxonomic pruning is enabled, the pathway prediction score cut-off is set to 0.15 and name-matching is set to enabled. For the GWMRs generated with Model SEED, the default parameters were also used: No public media formulation is selected, and the select template model parameter is set to *Automatically select*. For the METRONOME GWMRs, the OrthoMCL enzyme assignment sub-module (Section 3.2.1) is used along with the KEGG and MetaCyc data mining sub modules (Section 3.2.2).

3.3.1. E. coli

The *E. coli* K-12 MG1655 genome sequence (Blattner et al, 1997) is used (NCBI Taxonomy: 511145, GenBank Accession: U00096.2) to generate all *E. coli* draft GWMRs. The curated set of reactions come from two sources, EcoCyc (Keseler et al, 2009) and KEGG (Kanehisa et al, 2004). EcoCyc is a well curated comprehensive database of *E. coli* biology and is part of the MetaCyc collection of databases. KEGG

contains detailed biological information for a number of organisms, including *E. coli*. Figure 3.2 and Table 3.2 show the overlap between the draft GWMRs generated using METRONOME, Pathologic and Model SEED with the curated set of reactions taken from EcoCyc and KEGG.

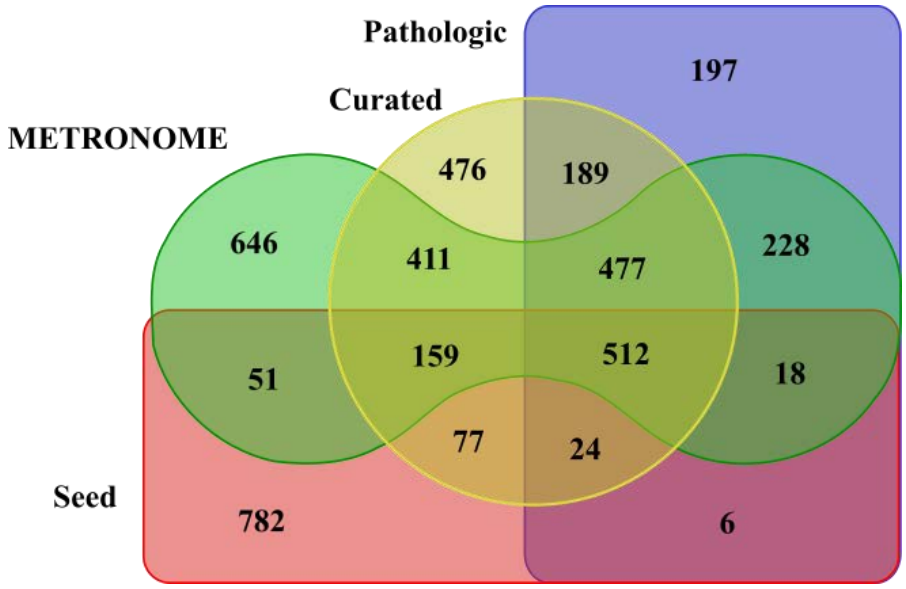


Figure 3.2: E. coli reaction overlap between the curated set of reactions and the reactions contained within the draft GWMRs generated using METRONOME, Pathologic and Model SEED.

In total there are 2,325 biochemical reactions in the *E. coli* curated set of which 512 (22%) are shared across all of the draft GWMRs, and 476 (20%) are not present in any of the draft GWMRs. METRONOME has the highest proportion of overlap with the curated set (67%), and Model SEED has the fewest (33%). METRONOME has the highest proportion of reactions that uniquely overlap with the curated set (16%), and Model SEED has the fewest (5%). Pathologic has the lowest proportion of un-curated reactions (27%), and also the lowest proportion of unique un-curated reactions (12%) with Model SEED having the highest (53% and 48%).

Table 3.2: Network stats for the *E. coli* draft GWMRs. The curated percentage is calculated as a proportion of the 2,325 reactions in the curated set. All of the other percentages are calculated relative to the total number of reactions within the respective draft GWMRs. The tool overlap describes the number of reactions that are common between the draft GWMR described by the row, and the draft GWMRs built using the other two tools.

GWMR	Total	Curated	Unique Curated	Un-curated	Unique Un-curated	Tool Overlap
METRONOME	2502	1559 (67%)	411 (16%)	943 (38%)	646 (26%)	1445 (58%)
Pathologic	1651	1202 (52%)	189 (11%)	449 (27%)	197 (12%)	1265 (77%)
Model SEED	1629	772 (33%)	77 (5%)	857 (53%)	782 (48%)	770 (47%)

3.3.2. S. cerevisiae

The *S. cerevisiae* S288C genome sequence (Goffeau et al, 1996) is used (NCBI Taxonomy: 559292, GenBank Accession: NC_001133.9) to generate all *S. cerevisiae* draft GWMRs. The curated set of reactions come from two sources, YeastCyc (Christie et al, 2004) and KEGG (Kanehisa et al, 2004). YeastCyc is a well curated comprehensive database of *S. cerevisiae* biology and is part of the MetaCyc collection of databases. KEGG contains detailed biological information for a limited number of organisms including *S. cerevisiae*. Figure 3.3 and Table 3.3 show the overlap between the draft GWMRs generated using METRONOME, Pathologic and Model SEED with the curated set of reactions taken from YeastCyc and KEGG.

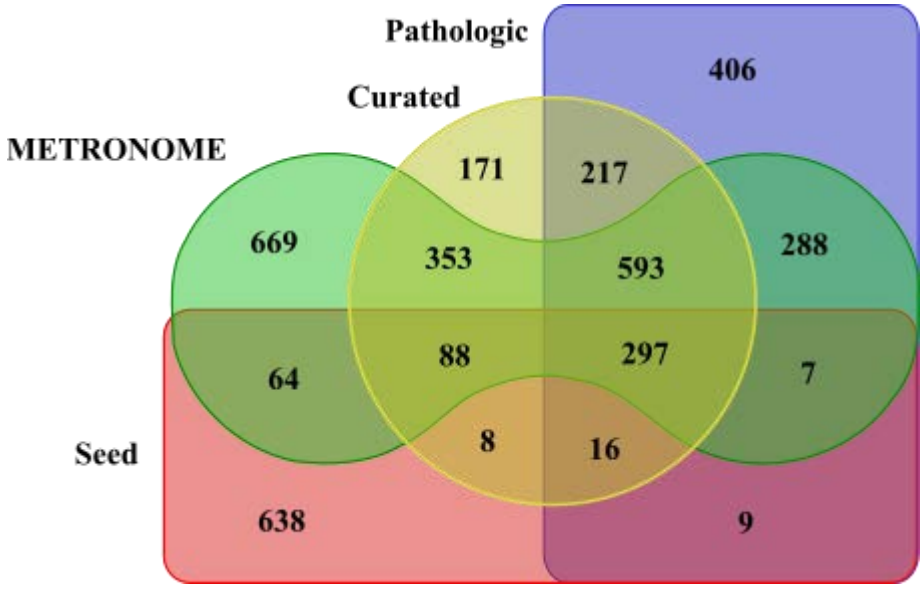


Figure 3.3: S. cerevisiae reaction overlap between the curated set of reactions and the reactions contained within the draft GWMRs generated using METRONOME, Pathologic and Model SEED.

In total there are 1,743 biochemical reactions in the *S. cerevisiae* curated set of which 297 (17%) are shared across all of the draft GWMRs, and 171 (10%) are not present in any of the draft GWMRs. METRONOME has the highest proportion of overlap with the curated set (76%) and Model SEED the fewest (23%). METRONOME has the highest proportion of reactions that uniquely overlap with the curated set (15%), and Model SEED has the fewest (1%). Pathologic has the lowest proportion of un-curated reactions (39%), and also the lowest proportion of unique un-curated reactions (22%) with Model SEED having the highest (64% and 57%).

Table 3.3: Network stats for the *S. cerevisiae* draft GWMRs. The curated percentage is calculated as a proportion of the 1,743 reactions in the curated set. All of the other percentages are calculated relative to the total number of reactions within the respective draft GWMRs. The tool overlap describes the number of reactions that are common between the draft GWMR described by the row, and the draft GWMRs built using the other two tools.

GWMR	Total	Curated	Unique Curated	Un-curated	Unique Un-curated	Tool Overlap
METRONOME	2339	1331 (76%)	353 (15%)	1028 (44%)	669 (28%)	1517 (65%)
Pathologic	1833	1123 (64%)	217 (12%)	710 (39%)	406 (22%)	1210 (66%)
Model SEED	1127	409 (23%)	8 (1%)	718 (64%)	638 (57%)	481 (43%)

3.4. Discussion

For a GWMR to be deemed of high quality, the 96-step protocol outlined in (Thiele & Palsson, 2010) should be followed. The protocol is divided into 4 key stages; draft reconstruction (steps 1-5), refinement of reconstruction (steps 6-37), conversion of reconstruction into computable format (steps 38-42) and network evaluation (steps 43-94). METRONOME is capable of automatically generating draft GWMRs, covering steps 1-5. Networks generated using METRONOME will require curation in order for them to be considered high quality. Curation is a largely manual process, but a number of the steps in the refinement process are made easier by the output files that are generated by METRONME (section 3.2.4), as they include metabolite formulas, reaction stoichiometry, reaction directionality, pathway information and metabolite identifiers when that information is available within the MetaNetX resource.

Draft GWMRs for *E. coli* and *S. cerevisiae* are built using METRONOME and two other automated draft GWMR tools, Pathologic and Model SEED producing six draft GWMRs in total. The reactions contained in each draft GWMR is compared to a curated set of reactions. For each species, the curated set came from two reliable and well curated data

sources within the KEGG (Kanehisa et al, 2014) and BioCyc (Christie et al, 2004; Keseler et al, 2009) resources. The networks generated using METRONOME also extracted reactions and metabolites from the KEGG and BioCyc databases, but they are extracted based on the enzyme associations of reactions, and not on any association based on the species. This could have introduced some bias and could perhaps explain why the Model SEED automated draft reconstructions performed badly.

Ideally well manually curated reconstructions of *E. coli* (Feist et al, 2007; Orth et al, 2011) and *S. cerevisiae* (Dobson et al, 2010; Heavner et al, 2013; Mo et al, 2009) that are independent from the KEGG and BioCyc databases would have been used for the comparisons to avoid any bias. Difficulties arise when taking this approach however as often manually curated reconstructions do not include database identifiers or enough metadata for the reactions and metabolites, making it hard to compare the contents of them with that of the draft reconstructions generated using the other tools. The relative performances of reconstruction tools should therefore be considered with this in mind.

The reactions from each draft GWMR is compared to the curated set by looking at: The relative number of reactions in the draft compared to the curated set, amount of the curated reactions that are in the GWMR and the precision of the GWMR. Precision is measured as the percentage of the total reactions in the draft GWMR that are also in the curated set. Table 3.4 and Table 3.5 contain these results.

Table 3.4: *E. Coli* draft GWMR accuracy information. The network size is calculated relative to the curated set of reactions. The curated reactions represent the total proportion of reactions within the network that are in the curated set. The precision is the total reactions in the draft GWMR that are also in the curated set.

	METRONOME	**Pathologic**	**Model SEED**
Network size	108%	71%	70%
Curated Reactions	67%	52%	33%
Precision	62%	73%	47%

The METRONOME *E. coli* draft GWMR is 108% the size of the size of the *E. coli* curated set capturing 67% of the reactions with a precision of 62%. 26% of the reactions in the METRONOME GWMR are not present in either the curated set or any other GWMR. The Pathologic *E. coli* GWMR is smaller than the curated set and is 11% more precise than the METRONOME GWMR but captures 15% less of the curated reactions. The Model SEED *E. coli* GWMR is slightly smaller than the Pathologic GWMR and it only covers one third of the curated reactions, has the worst precision and contains a large number of reactions that are not present in any other networks (48%).

In summary, the *E. coli* GWMRs generated by METRONOME and Pathologic are more accurate than that generated by Model SEED. Pathologic has generated a more compact and more precise network, whereas METRONOME has generated a larger and less precise network. METRONOME however, has recovered the most of the curated set of reactions.

Table 3.5: *S. cerevisiae* draft GWMR accuracy information. The network size is calculated relative to the curated set of reactions. The curated reactions represent the total proportion of reactions within the network that are in the curated set. The precision is the total reactions in the draft GWMR that are also in the curated set.

	METRONOME	**Pathologic**	**Model SEED**
Network size	134%	105%	65%
Curated captured	76%	64%	23%
Precision	52%	61%	36%

The METRONOME *S. cerevisiae* draft GWMR is 134% the size of the size of the *S. cerevisiae* curated set capturing 76% of the reactions with a precision of 52%. 29% of the reactions in the METRONOME GWMR are not present in either the curated set or any other GWMR. The Pathologic *S. cerevisiae* GWMR is 105% the size of the curated set and is 9% more precise than the METRONOME GWMR but captures 12% less of the

curated reactions. The Model SEED *S. cerevisiae* GWMR is the smallest GWMR at 65% the size of the curated set. It only covers 23% of the curated reactions, has the worst precision and contains a large number of reactions that are not present in any other networks (57%).

In summary, the S. *cerevisiae* GWMRs generated by METRONOME and Pathologic are more accurate than that generated by Model SEED. Pathologic has generated a more compact and more precise network, whereas METRONOME has generated a larger and less precise network. However, METRONOME has recovered the most of the curated set of reactions.

After looking at the results of the two species draft GWMRs, there is a clear pattern, METRONOME draft GWMRs are the largest and recover the most curated reactions, Pathologic draft GWMRs are smaller than METRONOMEs, covering less of the curated set but are more precise. MODEL Seed draft GWMRs are the smallest, recover the least curated reactions and are the least precise.

Table 3.6 shows how many MetaNetX cross references are present for various sets of overlapping reactions taken from the *S. cerevisiae* curated set and the *S. cerevisiae* draft GWMRs generated using METRONOME, Pathologic and Model SEED. The percentages show the proportion of reactions that have KEGG, MetaCyc or SEED ids in the MetaNetX reconciliation database. The proportion of reactions that have ids from the other databases represented in MetaNetX (Table 3.1) is also shown.

The reactions that are present in the curated set as well as all three of the draft GWMRs have a high proportion of references in MetaNetX across all data sources. The uniquely overlapping reactions between the curated set and each of the three GWMRs is inspected:

Uniquely shared between METRONOME and the curated set

The reactions have a higher proportion of KEGG ids (87%), followed by Model SEED (59%), suggesting that METRONOME mainly recovers its uniquely curated reactions from KEGG.

Uniquely shared between Pathologic and the curated set

As expected, the reactions have a far higher proportion of MetaCyc ids (96%), with 35% of reactions having KEGG ids and 41% Model SEED ids. This suggests that the uniquely curated reactions in the Pathologic draft GWMR come from MetaCyc

Uniquely shared between Model SEED and the curated set

No one database has a clearly higher proportion of ids, with 65% of reactions having KEGG ids, 75% having MetaCyc ids and 75% having Model SEED ids. This suggests that the Model SEED uniquely curated reactions are well cross-referenced across the different data sources.

Uniquely shared between METRONOME and other tools – not in curated set

The uniquely overlapping reactions between METRONOME and the other two tools is also inspected. The reactions that uniquely overlap between METRONOME and Pathologic have a far higher proportion of MetaCyc ids (96%). The overlapping reactions between METRONOME and Model SEED have a high number of Model SEED ids (92%), but also have a high number of KEGG ids (95%).

The reactions that are solely in the curated set have a high proportion of MetaCyc ids, suggesting that either Pathologic does not do a great job of recovering all reactions from

YeastCyc, or that the curated reactions that come from the KEGG *S. cerevisiae* data source are not recovered by Pathologic but well cross referenced in MetaNetX. The unique reactions in the METRONOME, Pathologic and Model SEED draft GWMRs have a higher proportion of KEGG ids, MetaCyc ids and Model SEED ids respectively.

The uniquely overlapping reactions between METRONOME and pathologic have a very high proportion of MetaCyc ids in MetaNetX, whereas the overlapping reactions between METRONOME and Model SEED have a high proportion of KEGG ids as well as Model SEED ids. This suggests that MetaNetX does a good job of reconciling KEGG and Model SEED reactions.

Table 3.6: MetaNetX cross references of the S. cerevisiae curated set of reactions and the reactions contained within the draft GWMRs generated using METRONOME, Pathologic and Model SEED.

Reaction set				***MetaNetX cross references***			
Curated	METRONOME	Pathologic	Seed	**KEGG**	**MetaCyc**	**SEED**	**Other**
X	**X**	**X**	**X**	97%	94%	94%	100%
X	**X**			87%	33%	59%	53%
X		**X**		35%	96%	31%	44%
X			**X**	63%	75%	75%	63%
X				50%	75%	34%	43%
	X			77%	56%	55%	43%
		X		30%	96%	29%	39%
			X	46%	37%	94%	55%
	X	**X**		16%	94%	15%	36%
	X		**X**	95%	27%	92%	38%

Pathologic is clearly the most precise draft GWMR tool however unlike METRONOME, a genome annotation is required before using the tool. In this instance well curated genome annotations have been taken from the NCBI database whereas METRONOME has used unchecked OrthoMCL genome annotations. METRONOMEs precision could be increased by inspecting the genome annotation prior to data mining or by including a GenBank parser. In reality this would not be feasible if METRONOME is being used for

its intended purpose of generating draft GWMRs for new genome sequences, which will not necessarily have accurate GenBank genome annotations available.

Pathologic is developed by the team behind the MetaCyc database (Caspi et al, 2014) and as a result exploits a great deal of information from within it. The algorithm consists of two key steps. The first extracts enzymatic reactions from MetaCyc in a similar manner to the data mining step in METRONOME. The second step infers metabolic pathways using a rule based approach, adding reactions to the model as complete pathways (Karp et al, 2011). If the reactions from a given pathway are mostly absent, then the pathway is not added. This is a form of automated curation, and results in a number of reactions being cut from the model is a likely factor in the higher precision rates in Pathologic.

Model SEED is part of the wider SEED (Overbeek et al, 2005) and RAST (Aziz et al, 2008) family of software for annotating and analysing genomes and is linked very closely to it. The approach used in Model SEED differs to that of METRONOME and Pathologic in that it relies on its own ontology generated using the RAST genome annotation to infer the reactions. This is different to the way that METRONOME and Pathologic operate, using the common E.C. number ontology to infer reactions. As a result the generated GWMR is heavily reliant on the specifics of RAST genome annotation, which could explain why the Model SEED GWMRs have the least overlap with the curated set.

METRONOME is the least restrictive tool and able to cope with unannotated genome sequences, whereas Pathologic and Model SEED both require prior genome annotations in specific formats (GenBank and RAST/SEED). Pathologic and Model SEED both build the GWMRs from a single source whereas METRONOME can be configured to use multiple sources. METRONOME requires the use of REST web services to generate a

draft GWMR. Pathologic is part of a wider software package, Pathway Tools, which requires a license and installation but does not require an internet connection to perform the reconstruction. Model SEED is entirely web based and the reconstruction is done entirely on external servers.

Each of the three tools operates differently from each other and is suitable for different use cases. It could therefore be argued that it is not appropriate to compare the performance of METRONOME to that of Pathologic and Model SEED. However by comparing in this way confidence can be gained into METRONOMEs ability to generate draft GWMRs.

It is clear that Pathologic and Model SEED are good at recovering reactions from their associated databases, however depending on what is considered to be the gold standard set of reactions, the performance of the draft GWMR process varies. METRONOMEs architecture allows for several sub-modules to be created to extract metabolic reactions from a variety of sources increasing its accuracy. However, with each extraction source added, the likelihood of extracting erroneous reactions increases, reducing the precision of METRONOMEs draft GWMR.

The performance of the automated draft GWMR tools are accessed by measuring the proportion of reactions that are recovered from a pre-defined curated set. This curated set of reactions is a merger of two well curated metabolic reconstruction resources. There is a possibility that good values for the number of curated reactions and the precision those reactions could be achieved by chance. To further test this, a random gene set could be used as input to the various tools and the precision and accuracy of the resulting draft GWMRs measured. Another way would be to add noise into the input data by inserting

random genes and inspect the differences in performance of the tools. If significantly better results are achieved with this random or noisy data, it could indicate inadequacies in the automated draft reconstruction processes.

3.5. Conclusion

METRONOME is a lightweight flexible platform for automated draft GWMR that through appropriate configuration of sub-modules can take an unannotated nucleotide sequence as an input and return a draft GWMR as its output in the SBML format. Currently, METRONOME implemented OrthoMCL enzyme assignment, KEGG data mining and MetaCyc data mining sub modules, which extract enzymatic reactions from these sources.

METRONOME then merges these extracted enzymatic reactions to form a single coherent network using the MetaNetX database. METRONOME does not require a genome annotation, meaning that it is suitable for use with newly sequenced genomes. By incorporating new modules, reactions can be extracted from multiple sources, including online databases and directly from SBML files. Through the use of the MetaNetX reconciliation project, the information from these sources can be coherently merged, and a large amount of useful database cross-links can be included. The end result of this is larger GWMRs that recover a high proportion of the known biochemical reactions for the two model organisms tested.

Although METRONOME recovers the highest proportion known reactions for two model organisms, when compared to two popular automated draft GWMR tools, the GWMRs have a slightly lower precision than Pathologic. The precision could be improved by including a pathway inference step or by designing a more accurate enzyme assignment

sub-module, which once implemented can be straightforwardly incorporated into METRONOMEs highly flexible platform.

4. Draft GWMR of *Daphnia magna* using METRONOME Platform

Species of the genus *Daphnia* are renowned models in ecotoxicology and are widely used as indicators of water quality and environmental health. They are also key models in evolutionary biology and the study of adaptive responses to environmental change. Genome-wide metabolic reconstructions (GWMRs) provide a platform for analysis, visualisation and contextualisation of high throughput omics data sets (Francke et al, 2005), can help in understanding the global properties of metabolic networks (Ma & Zeng, 2003), and can guide metabolic engineering and hypothesis driven discovery (Oberhardt et al, 2009).

A GWMR of a species of *Daphnia* is useful for constructing computational hypotheses about metabolic response to environmental insults or stressors. This chapter details the draft GWMR of *Daphnia magna* using the METRONOME platform (Chapter 3) which is then subsequently used to make predictions about metabolic response to two environmental stressors relevant to human-driven pollution (Chapter 5).

Version 2.4 of the draft genome sequence of the Xinb3 strain of *D. magna*[1] taken from wFleaBase (Colbourne et al, 2005) is used as the input to the METRONOME platform, which is configured to assign enzymes using OrthoMCL (Section 4.1) and to extract reactions and metabolites from the KEGG and MetaCyc databases (Section 4.2). The contents of the resulting draft GWMR are assessed by investigating the overlap between core and literature reported metabolic pathways (Section 4.3).

[1] April 2010 - https://wiki.cgb.indiana.edu/display/magna/Daphnia+magna+Genome

4.1. Enzyme assignment

The first step when using the METRONOME platform is to assign enzymes to the input genome sequence. Enzyme assignment is achieved using the OrthoMCL sub-module (Section 3.2.1), which consists of two key steps. The first uses the OrthoMCL algorithm to assign ortholog groups to the input sequence. The second uses the OrthoMCL web service to recover enzyme assignments for each of the assigned ortholog groups.

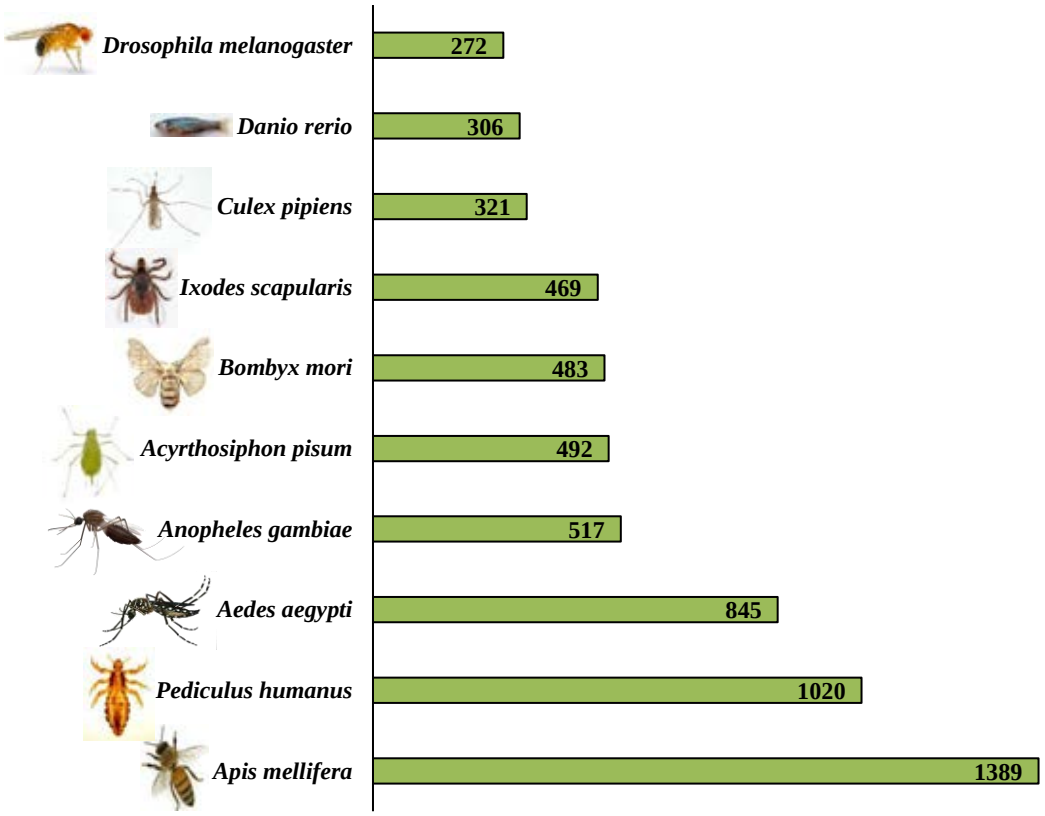


Figure 4.1: Ten species with the highest number of OrthoMCL group matches with the xinb3 V 2.4 *D. magna* genome sequence. The data is generated using the enzyme assignment module of the METRONOME platform.

The OrthoMCL algorithm found 17,468 ortholog matches of which 295 were not assigned to an OrthoMCL group. Of the remaining 17,173 matches, 8,255 unique OrthoMCL

groups across 109 species are found. Figure 4.1 shows the ten species with the most ortholog matches. The species with the most matches is *Apis mellifera*, or the western honey bee, and all but one of the top ten matches are phylogenetically similar species. Table 4.1 details all of the species for which orthologs are found. From the list of OrthoMCL groups, a total of 1,267 E.C. numbers, of which 1,142 were complete, are recovered and subsequently used in the data mining process (Section 4.2).

Table 4.1: Number of unique OrthoMCL group matches within the xinb3 V 2.4 *D. magna* sequence per species. The data is generated using the enzyme assignment module of the METRONOME platform.

Species	**OrthoMCL groups**	**Species**	**OrthoMCL groups**
Apis mellifera	1389	Micromonas	4
Pediculus humanus	1020	Mycobacterium tuberculosis	4
Aedes aegypti	845	Trypanosoma congolense	4
Anopheles gambiae	517	Agrobacterium tumefaciens	3
Acyrthosiphon pisum	492	Burkholderia mallei	3
Bombyx mori	483	Leishmania infantum	3
Ixodes scapularis	469	Listeria monocytogenes	3
Culex pipiens	321	Plasmodium falciparum	3
Danio rerio	306	Trichomonas vaginalis	3
Drosophila melanogaster	272	Yersinia enterocolitica	3
Nematostella vectensis	269	Emericella nidulans	2
Gallus gallus	201	Candida glabrata	2
Mus musculus	153	Coccidioides immitis	2
Takifugu rubripes	129	Cyanidioschyzon merolae	2
Tetraodon nigroviridis	121	Cryptococcus neoformans	2
Monodelphis domestica	112	Dehalococcoides ethenogenes	2
Equus caballus	111	Entamoeba invadens	2
Macaca mulatta	92	Leishmania braziliensis	2
Rattus norvegicus	91	Rhodopirellula baltica	2
Ciona intestinalis	90	Ricinus communis	2
Canis lupus	88	Staphylococcus aureus	2
Ornithorhynchus anatinus	73	Toxoplasma gondii	2
Homo sapiens	69	Wolbachia endosymbiont of Culex quinquefasciatus	2
Phytophthora ramorum	62	Aquifex aeolicus	1
Trichoplax adhaerens	56	Archaeoglobus fulgidus	1
Pan troglodytes	44	Babesia bovis	1

Species	OrthoMCL groups	Species	OrthoMCL groups
Chlorobium tepidum	28	Campylobacter jejuni	1
Brugia malayi	25	Cryptosporidium muris	1
Caenorhabditis elegans	20	Cryptococcus bacillisporus	1
Caenorhabditis briggsae	18	Coccidioides posadasii	1
Bacillus anthracis	15	Escherichia coli	1
Burkholderia pseudomallei	13	Entamoeba dispar	1
Schistosoma mansoni	13	Francisella tularensis	1
Ralstonia solanacearum	12	Gibberella zeae	1
Dictyostelium discoideum	11	Haloquadratum walsbyi	1
Physcomitrella patens	11	Leishmania mexicana	1
Phanerochaete chrysosporium	10	Methanococcus maripaludis	1
Volvox carteri	10	Neurospora crassa	1
Chlamydomonas reinhardtii	9	Ostreococcus tauri	1
Oryza sativa	8	Plasmodium chabaudi	1
Vibrio cholerae	8	Rickettsia prowazekii	1
Arabidopsis thaliana	7	Saccharomyces cerevisiae	1
Brucella suis	7	Salmonella enterica	1
Monosiga brevicollis	7	Shigella flexneri	1
Tetrahymena thermophila	7	Streptococcus pneumoniae	1
Clostridium botulinum	6	Schizosaccharomyces pombe	1
Geobacter sulfurreducens	6	Synechococcus	1
Thalassiosira pseudonana	6	Trypanosoma brucei	1
Archaeoglobus fulgidus	5	Trypanosoma brucei	1
Aspergillus oryzae	5	Trypanosoma cruzi	1
Coxiella burnetii	5	Treponema pallidum	1
Kluyveromyces lactis	5	Thermoplasma volcanium	1
Methanocaldococcus jannaschii	5	Wolinella succinogenes	1
Candida albicans	4	Yersinia pestis	1
Laccaria bicolor	4		

4.2. Data mining and network merging

Two METRONOME data mining sub-modules are used to extract biochemical reactions from the KEGG (Kanehisa et al, 2014) and MetaCyc (Caspi et al, 2014) databases (Section 3.2.2). METRONOME builds draft GWMRs from each data source before merging them. Table 4.2 shows the number of reactions and metabolites that are in each of the *D. magna* draft GWMRs. The KEGG network has 2,249 reactions and 2,256 metabolites and the MetaCyc network has 2,028 reactions and 2,562 metabolites.

Table 4.2: Total number of reactions and metabolites in the KEGG, MetaCyc and merged *D. magna* draft networks generated using the METRONOME platform.

Network	# Reactions	# Metabolites
KEGG	2,249	2,256
MetaCyc	2,028	2,562
Merged	3,273	3,473

The network merging procedure described in section 3.2.3 uses the MetaNetX (see section 3.2.3) database (Moretti et al, 2016) to merge the KEGG and MetaCyc networks. The merged network contains 3,273 reactions and 3,473 metabolites. Figure 4.3 shows a visualisation of the merged draft GWR of *D. magna* using the software package Cytoscape (Smoot et al, 2011).

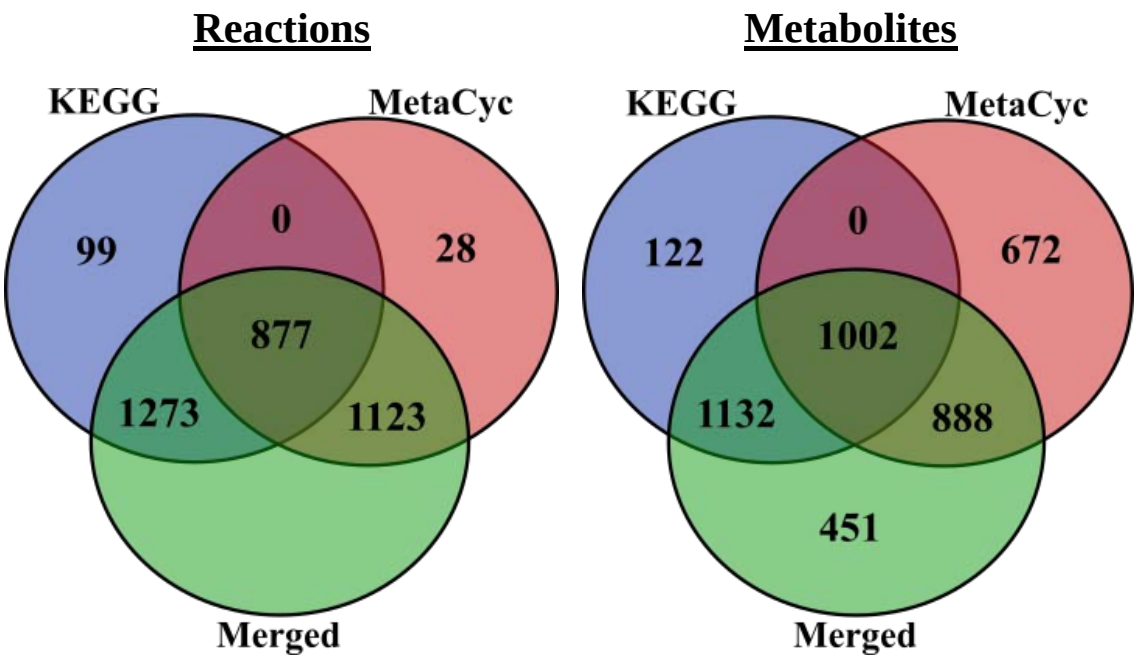


Figure 4.2: Overlap of reactions and metabolites between the KEGG, MetaCyc and merged draft *D. magna* GWMRs generated using the data mining and network merging sub modules in the METRONOME platform.

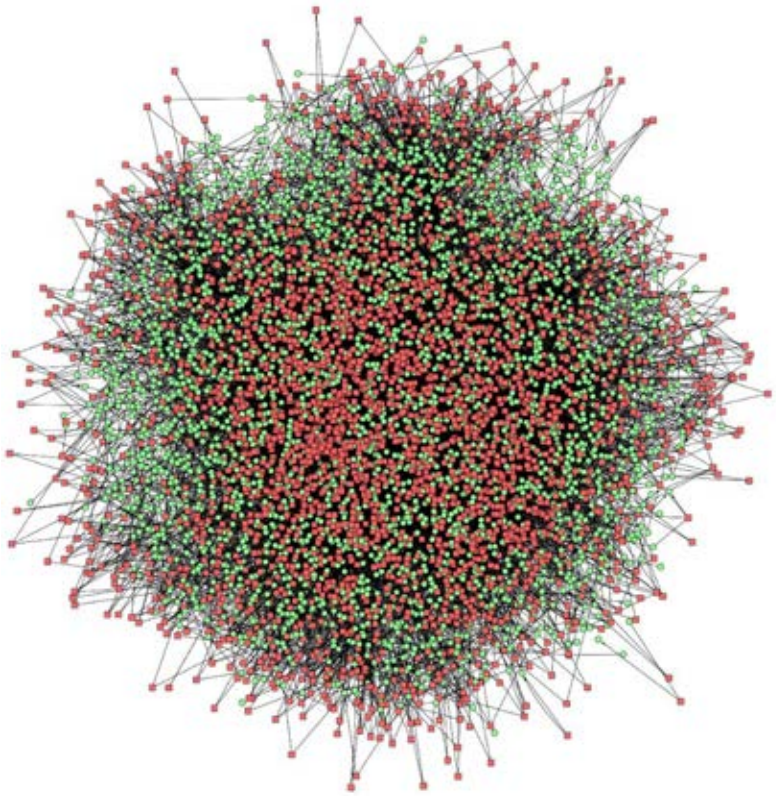


Figure 4.3: Cytoscape (Smoot et al, 2011) visualisation of the merged *D. magna* draft GWMR. Red and green nodes represent reactions and metabolites respectively. A metabolite node is linked to a reaction node if it is either a substrate or product of the reaction.

Figure 4.2 shows the overlap between the KEGG, MetaCyc and merged *D. magna* draft networks. There are 877 reactions and 1,002 metabolites that were directly comparable between the KEGG and MetaCyc networks. The KEGG network has 1,273 and 1,132 reactions/metabolites not in the MetaCyc network, and the MetaCyc network has 1,123 and 888 reactions/metabolites not in the KEGG network.

There are 99 reactions and 122 metabolites in the KEGG network, and 28 reactions and 672 metabolites in the MetaCyc network that do not have an entry in MetaNetX and are therefore not included in the merged network. In total, 127 reactions and 342 metabolites extracted during the data mining process are not included in the merged network.

There are 451 metabolites in the merged network that are not directly mapped to metabolites in either the KEGG or MetaCyc networks. This is likely due to how the MetaNetX database reconciles reactions from several sources (Table 3.1), this means that the reaction definitions can differ from the equivalent KEGG and MetaCyc definitions. Upon inspection, these 451 metabolites are all linked to MetaCyc entries that are labelled, "*compound class*". This suggests that the MetaNetX reconciliation procedure (described in section 3.2.3.1) has reconciled some compounds in reactions in their more generic class form. An illustration of this is that the MetaCyc compound class *an aliphatic N-acetyl-diamine* (MetaCyc id - Aliphatic-N-Acetyl-Diamines) is included in the 451 metabolites not found in the KEGG or MetaCyc network, and the metabolite *acetylcadaverine* (MetaCyc id - CPD-10194) is in the 672 metabolites unique to the MetaCyc GWMR. The metabolite *acetylcadaverine* sits below the metabolite class *an aliphatic N-acetyl-diamine* in the MetaCyc ontology.

4.3. Network interrogation

There is no published GWMR of *D. magna*, and there is little reported on the *D. magna* metabolome (Jones et al, in preparation). Subsequently, it is difficult to know how accurate the generated *D. magna* draft GWMR is. The KEGG collection of databases (Kanehisa et al, 2014) includes a database that contains a number of metabolic pathways which are made up of a number of KEGG modules. There is a KEGG pathway called the reference pathway (KEGG id: map01100), which contains all metabolic pathways represented in KEGG. The KEGG Mapper software (Kanehisa, 2013) allows for KEGG pathway maps to be coloured based on some user provided data. Figure 4.4 shows the KEGG reference pathway with all reactions and metabolites that are in the *D. magna* draft GWMR which have KEGG ids coloured in black. It can be seen that there is fairly wide coverage of the KEGG reference pathway in the reconstruction, with many complete pathways including core pathways corresponding to energy metabolism, carbohydrate metabolism, lipid metabolism and nucleotide metabolism.

4.3.1. Core KEGG Modules

Figure 4.5 shows three core KEGG modules (TCA cycle, Glycolysis pathway and Urea cycle) also coloured using KEGG Mapper, with reactions and metabolites present in the *D. magna* draft GWMR coloured in pink. Table 4.3 summarises the coverage of these modules. For all three modules there is 100% coverage of metabolites, however none have all of the reactions, with the lowest coverage being the TCA cycle with two thirds of the reactions covered. The best coverage in terms of reactions is the urea cycle which is only missing one reaction.

Table 4.3: Coverage of three core KEGG modules (Figure 4.5) in the *D. magna* draft GWMR.

KEGG Module Id	KEGG Module Name	Module Reactions	Module Metabolites	Reactions in Model	Metabolites in Model
M00009	Citrate cycle (TCA cycle, Krebs cycle)	18	12	66.66%	100.00%
M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	15	11	73.33%	100.00%
M00029	Urea cycle	5	9	80.00%	100.00%

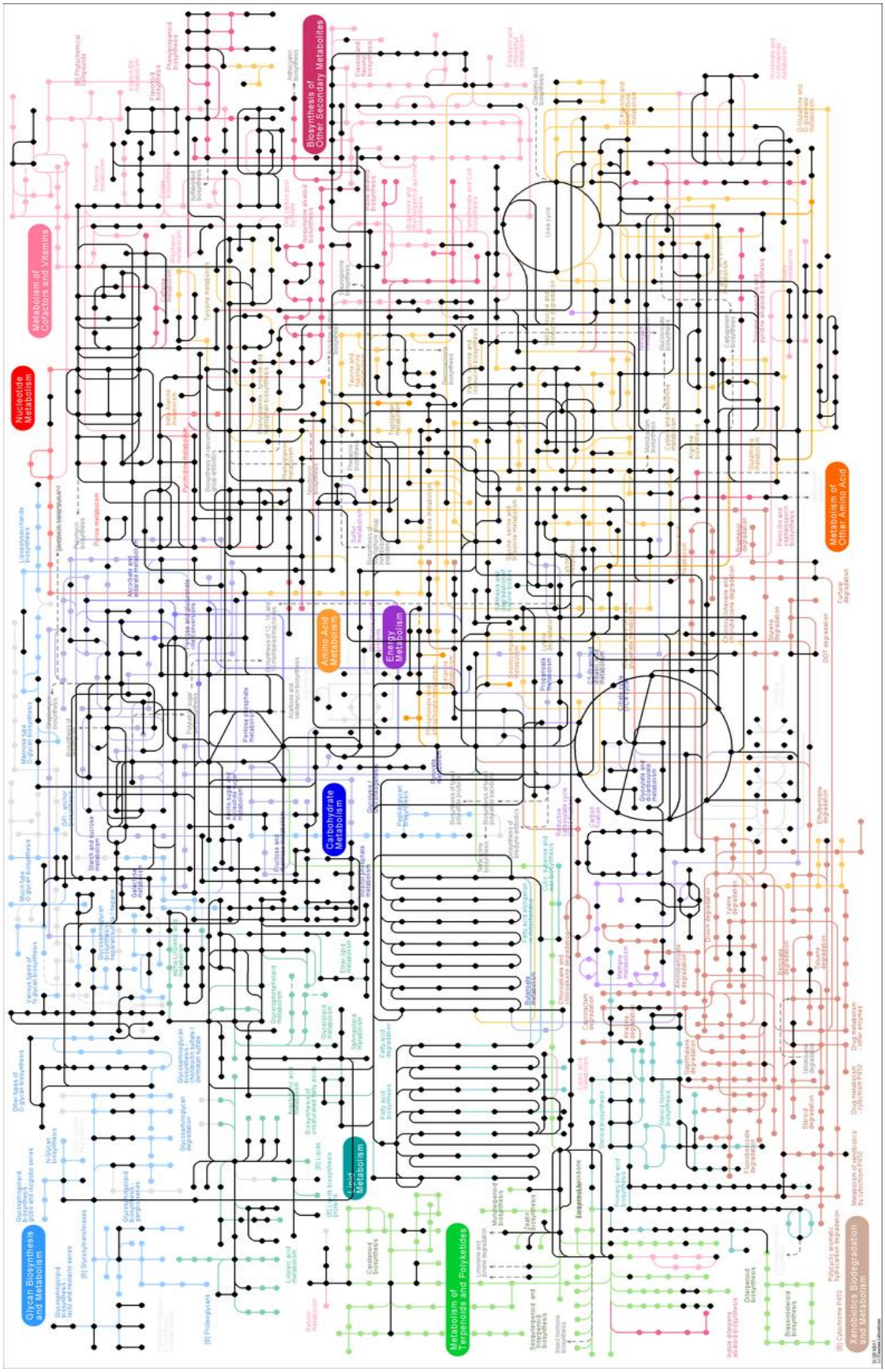


Figure 4.4: KEGG reference pathway with all reactions and metabolites in the *D. magna* draft GWMR coloured in black.

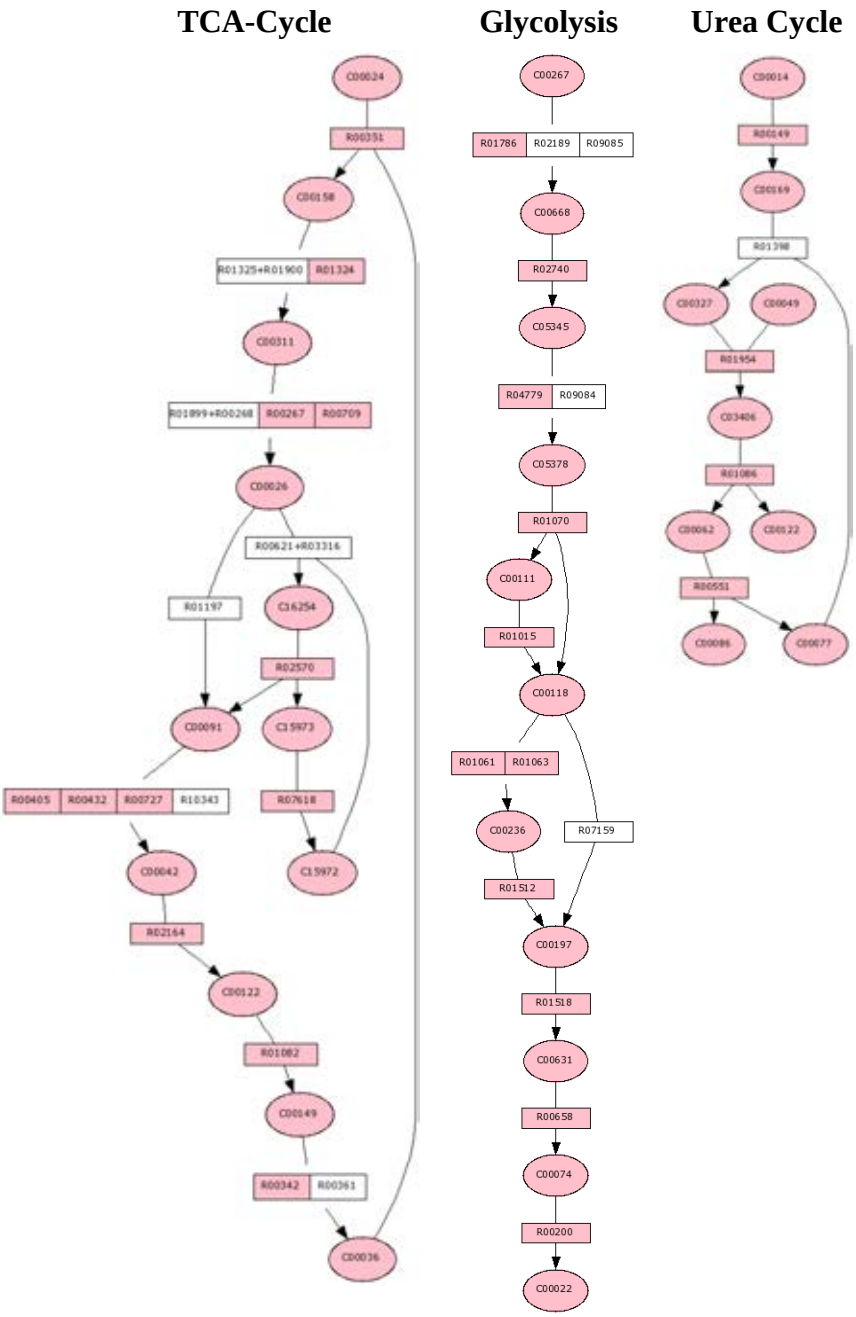


Figure 4.5: Three core pathways coloured using KEGG Mapper (Kanehisa, 2013), all reactions and metabolites that are present in the *D. magna* draft GWMR are coloured pink.

4.3.2. KEGG Pathways from literature

Table 4.4 lists 13 KEGG pathways that have been reported in toxicology studies of *D. magna* (Garreta-Lara et al, 2016; Poynton et al, 2011), along with the coverage of them in the *D. magna* draft GWMR. For each of the pathways, the percentage coverage of the

reactions and metabolites is calculated. As with the core pathways listed in section 4.3.1, the general trend is that the metabolite coverage is better than the reaction coverage. In total 84.30% of metabolites and 54.94% of reactions from the pathways are in the *D. magna* draft GWMR. Five of the pathways have 100% metabolite coverage, and ten have a coverage of 80% or higher.

Table 4.4: Coverage of 13 KEGG pathways reported in *D. magna* transcriptomics and metabolomics toxiocology studies (Garreta-Lara et al, 2016; Poynton et al, 2011) in the *D. magna* draft GWMR.

KEGG Pathway Id	KEGG Pathway Name	Pathway Reactions	Pathway Metabolites	Reactions in Model	Metabolites in Model
ko00020	TCA cycle	63	52	66.67%	90.38%
ko00052	Galactose metabolism	26	31	38.46%	70.97%
ko00061	Fatty acid biosynthesis	10	10	70.00%	100.00%
ko00220	Arginine biosynthesis	16	21	43.75%	80.95%
ko00250	Alanine, aspartate and glutamate metabolism	19	19	57.89%	73.68%
ko00260	Glycine, serine and threonine metabolism	21	19	66.67%	94.74%
ko00280	Valine, leucine and isoleucine degradation	11	14	81.82%	100.00%
ko00290	Valine, leucine and isoleucine biosynthesis	23	28	47.83%	82.14%
ko00310	Lysine degradation	12	10	83.33%	100.00%
ko00400	Phenylalanine, tyrosine and tryptophan biosynthesis	22	26	13.64%	69.23%
ko00480	Glutathione metabolism	2	3	100.00%	100.00%
ko00500	Starch and sucrose metabolism	6	7	16.67%	85.71%
ko00620	Pyruvate metabolism	2	2	50.00%	100.00%

4.4. Discussion and conclusion

The METRONOME platform (Chapter 3) has been used to generate a draft GWMR of *D. magna* using the xinb3 reference genome taken from wFleabase (Colbourne et al, 2005). 1,142 complete enzymes are assigned using the OrthoMCL algorithm and are used to extract biochemical reactions from the KEGG and MetaCyc databases and construct two networks which are subsequently merged into a single network containing 3,273 reactions and 3,473 metabolites, which forms the final *D. magna* draft GWMR.

The contents of the network are then investigated by looking at the coverage of the KEGG reference pathway, three core KEGG pathway modules and thirteen KEGG pathways that have been highlighted in *D. magna* toxicology studies. A large amount of the KEGG reference pathway is included in the network, with many complete pathways present. The three core KEGG modules; TCA cycle, glycolysis pathway and urea cycle are well represented in the model, with 100% metabolite coverage and 71% reaction coverage. The thirteen literature reported KEGG modules are well represented in terms of metabolites, with 84.30% coverage, but have only 54.94% reaction coverage. The (Garreta-Lara et al, 2016) study however, which exclusively contributed nine of the thirteen pathways, contains no evidence for all of the metabolites in a given pathway being present. Only the fact that at least two metabolites that are within the pathways have been measured to be effected by the treatments being considered is reported. There is therefore no evidence that all of the metabolites within the KEGG pathways have been observed in *D. magna*.

It is difficult to say how well the model represents the metabolome of *D. magna* as the complete list of expected reactions and metabolites is unknown. It is clear the *D. magna* draft GWMR is missing some reactions from some of the pathways that are expected to

be present as all of the participating metabolites and surrounding reactions are present. As discussed in section 3.4, the METRONOME platform would benefit from the inclusion of a pathway inference module such as the rule-based inference system used in Pathologic (Karp et al, 2011). That being said, the reaction, and especially the metabolite coverage of the highlighted pathways is satisfactory. The accuracy of the draft network could be improved by iterative time-consuming manual curation and performing Flux Balance analysis.

5. Computational Hypothesis Generation of *Daphnia magna* Metabolic Response Using Active Modules

Traditional approaches for analysing transcriptomics datasets to assess metabolic changes involve performing enrichment analysis using the KEGG database tools (Wrzodek et al, 2011) or Gene Ontology (GO) terms (Ashburner et al, 2000) using Gene Set Enrichment Analysis (GSEA) (Subramanian et al, 2005). These approaches are limited because they rely on inflexible pre-defined metabolic pathways or ontologies that do not necessarily represent the highly interconnected nature of the various metabolic networks.

Active module identification is an alternative approach that uses biological network reconstructions to identify hot spots within a network that are not constrained by traditional ontologies or pre-defined pathways. The active module identification approach involves first scoring a network using transcriptomics data and then heuristically searching for highly connected sub-networks, or hot-spots, within the network that have substantially different scores compared to the background network. These hot-spots reveal the coordinated response of a biological network to the observed changes in gene expression in an unbiased way that does not rely on pre-defined pathways. The approach can be applied to different types of biological networks such as protein interaction networks and metabolic networks.

The aim of this chapter is to take the draft GWMR of *D. magna* detailed in Chapter 4 and analyse it using the AMBINET algorithm (Bryant et al, 2013b), an extension of the original active modules algorithm (Ideker et al, 2002) for use with metabolic networks. Two transcriptomic RNA-Seq datasets from the STRESSFLEA (Orsini et al, 2016) project are used to score the network to be used by the AMBIENT algorithm. These

datasets measure the effect on the gene expression of *D. magna* when exposed to environmentally relevant concentrations of human-induced environmental stressors that are relevant in the context of human-driven pollution.

The result of this is a set of sub-networks, or hot-spots, which contain reactions and metabolites representing areas of the *D. magna* metabolome that are predicted to be highly affected by the two conditions tested. This type of analysis has previously not been done with *Daphnia* and uses the AMBIENT algorithm in a different way than previously reported. Formerly, AMBIENT is used to investigate known organism behaviour (Bryant et al, 2013a; Bryant et al, 2013b), whereas here, unknown organism response to environmental stressors is investigated. The extracted sub-networks are subsequently expanded to KEGG modules and KEGG pathways that are derived from the active modules. The sub-networks, KEGG modules and KEGG pathways form computationally generated hypothesis about metabolic response of *D. magna* to the effects of the induced environmental stressors at different layers of granularity.

5.1. Introduction

The publication of the *Daphnia pulex* (Colbourne et al, 2011) and *D. magna* (Colbourne et al, in preparation) genome sequences has enabled RNA-Seq transcriptomic measurements of *D. magna* to be made. These measurements allow for the detailed analysis of the links between genes and the environment by measuring the effect on gene expression of an environmental stressor. Capitalizing on this, the STRESSFLEA consortium, a research network funded by the ESF EUROCORES Programme EuroEEFG generated a comprehensive set of RNA-Seq datasets obtained from exposing three isolates of *D. magna* to twelve biotic and abiotic environmental perturbations (Orsini et al, 2016). This rich RNA-Seq dataset enabled the identification of early stress

responses to ecologically relevant biotic and abiotic environmental perturbations at a transcriptomic level. These responses will undoubtedly have a downstream, effect at the metabolomics level.

Several methods exist for interrogating GWMRs based on experimental data. Metabolomics data can result in a specific set of metabolites of interest which can then be mapped onto the relevant nodes in the network. A number of techniques exist for extracting possible paths through the network that link the mapped metabolites of interest. These include methods that extract the shortest (Holme, 2009) or lightest (Croes et al, 2005; Croes et al, 2006) path between metabolites, find paths based on atom mapping (Blum & Kohlbacher, 2008) and chemical similarity between side compounds, and a method that is based on the PageRank algorithm used by Google for web searching (Lemetre et al, 2013). All of these methods require metabolomics data that identifies metabolites of interest, which may not always be available.

Gene Set Enrichment Analysis (GESA) is a method that identifies genes that are over or under represented or expressed in a transcriptomics dataset (Subramanian et al, 2005) and can be used to extract Gene Ontology (Ashburner et al, 2000) (GO) terms (Lemetre et al, 2013). These GO terms can act as a functional annotation and then be mapped to metabolic pathways or processes, thus highlighting area of a metabolome that should be affected by the varying gene expression observed in the transcriptomics data. These highlighted areas can be mapped onto a GWMR and then analysed using the methods previously outlined.

Transcriptomics data can also be directly integrated with GWMRs. A number of methods exist for improving the prediction of flux-based analyses (Brandes et al, 2012; Colijn et

al, 2009; van Berlo et al, 2011), metabolic engineering of microbial cells (Kim & Reed, 2012; Lee et al, 2012) and to generate context (or tissue) specific metabolic models based on gene expression patterns (Agren et al, 2012; Wang et al, 2012; Zur et al, 2010).

The AMBIENT algorithm (Bryant et al, 2013b), an extension to the active modules approach (Ideker et al, 2002), uses a search heuristic to identify sub-networks within a metabolic reconstruction that are significantly affected by a change in gene expression. It works by linking the gene expression data to reactions in the GWMR based on the enzymes or proteins that specific reactions are linked to (see section 5.2.1).

Here the AMBIENT algorithm, is used to identify sub-modules within the *D. magna* draft GWMR that are effected by two of the conditions studied in the STRESSFLEA project. KEGG pathway analysis is performed using the metabolites and reactions in the AMBIENT active modules. The resulting set of KEGG modules are predicted to be effected by the environmental stressors, therefore forming a computationally generated hypothesis of the effect on the *D. magna* metabolome.

5.2. Active module identification

The active module identification approach, is a generalised method for searching a network to find connected sets of nodes that are deemed to be highly active under a certain condition (Ideker et al, 2002). The approach was originally used with protein interaction networks to find active modules representing connected sets of genes with higher levels of differential expression than the overall network.

One of the most popular approaches to identify active modules is the significant area search method, which consists of three generalised steps (Figure 5.1). The first step scores the nodes and/or edges of the network based on some biological data such as transcriptomic data. The second step is to formulate a scoring function, which scores sub

networks so that the overall activity of the contained nodes and their interactions is represented. Finally a search strategy is used to optimise the scoring function and identify sub networks which become the active modules (Mitra et al, 2013).

A number of significant area search based methods for active module identification exist (Cabusora et al, 2005; Chowdhury & Koyuturk, 2010; Dao et al, 2011; Dittrich et al, 2008; Fortney et al, 2010; Huang & Fraenkel, 2009; Nacu et al, 2007; Scott et al, 2006; Segal et al, 2003; Sohler et al, 2004) which are all based on the procedure described by (Ideker et al, 2002). Due to the NP-hard nature of searching for active modules, heuristic search techniques are usually employed although deterministic methods have also been utilised (Dittrich et al, 2008; Qiu et al, 2010) including linear programming (Backes et al, 2012; Zhao et al, 2008). The original algorithm uses simulated annealing as its search heuristic (Ideker et al, 2002), but greedy search (Chuang et al, 2007; Hwang & Park, 2009; Nacu et al, 2007; Rajagopalan & Agarwal, 2005) and genetic algorithms (Klammer et al, 2010) have also been used.

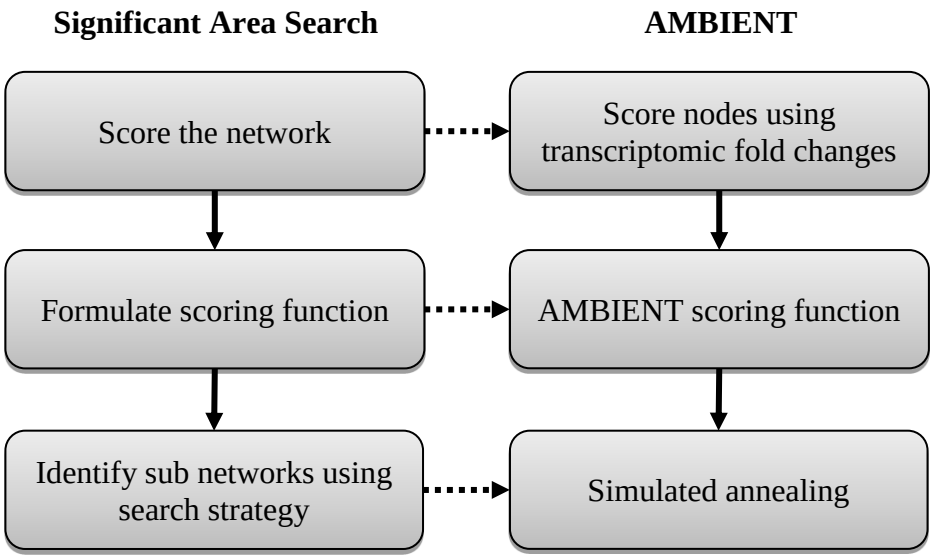


Figure 5.1: The key steps in the generalised significant area search based approach to active module identification and how AMBIENT maps to this framework.

5.2.1. AMBIENT

The AMBIENT (Active Modules for BIpartitE NeTworks.) algorithm (Bryant et al, 2013b) is an extension of the original active modules algorithm, that is specifically designed for metabolic networks represented as bipartite graphs. Bipartite graphs are where the network has two classes of nodes, in this case metabolite and reaction nodes. AMBIENT is a significant area search based algorithm and as such follows the three key steps (Figure 5.1).

Network scoring

In the current algorithm, only reaction nodes are scored based on some input data. Each reaction node in a metabolic reconstruction is linked to at least one enzymatic gene. Transcriptomic fold change data for these enzymatic genes can be used to score the reaction nodes according to the rules outlined in Table 5.1. If a reaction is linked to a single gene, the log fold change of the transcript corresponding to that gene is used. If a reaction is linked to a single enzyme and to multiple genes, the mean of the corresponding log fold changes is used. If multiple enzymes are linked to a reaction, the mean of each of the enzyme log fold changes is used

Metabolite nodes can also be scored using fold change data. However if this is not available, metabolite nodes are scored based on their degree. The degree of a metabolite node corresponds to how many reactions the metabolite is either a substrate or product of. Scoring metabolite nodes using the degree means that highly connected metabolites that participate in a number of metabolic processes are deemed more important. There is a danger to this approach however, as some metabolites like water or ATP are involved in an extremely large number of reactions, including them active modules can result in biologically meaningless active modules. These metabolites are known as currency

metabolites, and AMBIENT applies a penalty based on the degree of metabolites in active modules, thus favouring modules that do not contain very highly connected metabolites.

Table 5.1: Rules for scoring reaction nodes using transcriptomic data.

Reaction catalysed by	Score
Linked to a single gene	Log fold change of transcript
Single enzyme (multiple genes)	Mean of log fold changes
Multiple enzymes	Mean of enzyme log fold changes

Using this scoring mechanism will result in AMBIENT finding find up regulated modules. Down regulated modules can also be found, which is achieved by multiplying the reaction scores by -1. This has the effect giving the most down regulated reaction the highest score.

Scoring function

A module *m* represents a connected subgraph within the bipartite graph as a whole. Each module consists of r^m reactions and c^m metabolites. A module is assigned a score using equation (5-1).

$$S(m) = \ln(q)\left(\sum_{i}^{I} s(r_i^m) - \alpha \sum_{j}^{J} w\left(c_j^m\right)\right) \tag{5-1}$$

Where I is the number of reactions and J is the number of metabolites in module m, $q = r^m + c^m$ or the total nodes in the module m, $s(r_i^m)$ is the score of the reaction i and $w\left(c_j^m\right)$ is the degree of the metabolite j in the module m. The result of the second term being subtracted from the reaction scores is that modules that contain currency metabolites are penalised. α is a constant defined as

$$\alpha = \frac{|c| \sum_i^I s(r_i^+)}{|r^+| \sum_j^J w(c_j)} \tag{5-2}$$

Where $r^+$is all reactions with a positive score. The α term equates to the mean score of all positively scored reactions divided by the mean metabolite score across the complete network. This has the effect of ensuring that modules are not restricted by the weights of metabolite nodes and does not also generate large and biologically meaningless modules, essentially balancing the penalisation of currency metabolites but at the same time still favouring metabolites that have higher than average degree. The $\ln(q)$ term is added to favour large modules whilst at the same time not generating exceptionally large and therefore biologically meaningless modules, as the ln term saturates the overall module score as the size of the module increases.

Search strategy

As with the original active modules algorithm, simulated annealing is used as the search algorithm for AMBIENT. Simulated annealing is a search heuristic that mimics the atomic movement in a material that is heated and then cooled in a controlling manner. The algorithm is outlined in Algorithm 5.1. In the case of AMBIENT, each annealing step toggles a certain amount of neighbouring edges of the active modules rather than neighbouring nodes. This modification is made due to the bipartite representation of metabolic reconstructions.

1: **function** AMBIENT ALGORITHM

1. Set $t = t_{\text{init}}$, $T = T_{\text{init}}$
2. Select random set of edges, E of size t as the initial edge set
3. Calculate the score $\delta(E) = \sum_k S(m_k^E)$, where m^{E} is the set of connected components induced by E
4. Select t edges at random to form the toggle set, F
5. Propose a new edge set, $P = E \cup F - E \cap F$
6. Calculate $\delta(P) = \sum_k S(m_k^P)$
7. If $p < e^{\frac{\delta(P)-\delta(E)}{T}}$, where p is a random number drawn from [0,1], accept the proposed move and set $E = P$
8. Repeat steps 4 to 7 until the temperature reduction criterion is met
9. Set $t = 0.9t$ and $T = 0.9T$
10. If the maximum number of annealing steps is reached then END, otherwise go to step 3

2: **end function**

Algorithm 5.1: The AMBINET simulated annealing algorithm.

t represents the amount of edges that are toggled during a toggle step. *T* represents the effective temperature for annealing, which describes the acceptance of negatively scoring steps. The values of t_{init} and T_{init} are determined stochastically from the network (Bryant et al, 2013b). The temperature reduction criterion referenced in step 8 of the algorithm is set so that annealing steps occur until the sum of scores of the modules over does not change by at least 5%. If the temperature reduction criterion is met before the maximum number of annealing steps is reached, the values for *t* and *T* are multiplied by 0.9 and the annealing procedure resumes. This has the effect of reducing the amount of edges that are toggled during each annealing step, and relaxing the acceptance criteria in step 7.

5.3. *D. magna* active module identification using AMBIENT

The AMBIENT algorithm is used to find active modules in the draft GWMR of *D. magna* described in Chapter 4 based upon the transcriptomic data from the STRESSFLEA datasets (Orsini et al, 2016). The STRESSFLEA project uses *D. magna* as a model organism to investigate mechanisms of adaptation to environmental stressors using genomics tools. This involved exposing *D. magna* populations to twelve human induced

biotic and abiotic stressors. Transcriptomic RNA-Seq datasets were produced that measured the effect on gene expression that each of the twelve exposures had.

RNA-Seq fold change data from two of the twelve original conditions is used, Pb (278µg/L) and Carbaryl (8 µg/L). These conditions are chosen as the analysis of the gene expression data sets showed that they had some of the highest number of well annotated genes among the twelve data sets (Orsini et al, 2018), and that they are also ecologically relevant as they are common pollutants. Pb is a common by-product of heavy industry and Carbaryl is a once common highly toxic agricultural insecticide associated with a number of diseases including cancer and diabetes (Popovska-Gorevski et al, 2017). Both of these human-induced environmental stressors have had significant environmental impact and can be found in freshwater systems across the planet.

5.3.1. Results

For each stressor, the AMBIENT algorithm is run with the maximum annealing steps set to 100,000 and the number of empirical significance tests set to 10,000. The algorithm is executed twice for each condition, once for detecting up regulated active modules and once for detecting down regulated active modules. For the purpose of this study, the up and down regulated active modules are combined. Table 5.2 summarises the active modules found using AMBIENT. Figure 5.2and Figure 5.3 visualise the largest Carbaryl and Lead AMBIENT modules. Appendix A contains visualisations and details of the metabolites contained for all generated AMBIENT modules.

Table 5.2: Summary of AMBIENT active module identification on the *D. magna* GWMR and two STRESSFLEA transcriptomic data sets.

Condition	Module #	# Reactions	# Metabolites
Carbaryl	1	21	19
Carbaryl	2	6	6
Carbaryl	3	5	5
Carbaryl	4	3	3
Carbaryl	5	2	2
Carbaryl	6	1	2
Carbaryl	7	119	33
Carbaryl	8	11	12
Carbaryl	9	7	7
Carbaryl	10	4	5
Carbaryl	11	6	7
Carbaryl	12	4	5
Carbaryl	13	3	4
Carbaryl	14	2	2
Lead	1	129	47
Lead	2	5	11
Lead	3	12	3
Lead	4	12	6
Lead	5	6	3
Lead	6	1	4

Figure 5.2: Visualisation of the Carbaryl active module #7. Generated using MetExploreViz (Chazalviel et al, 2017).

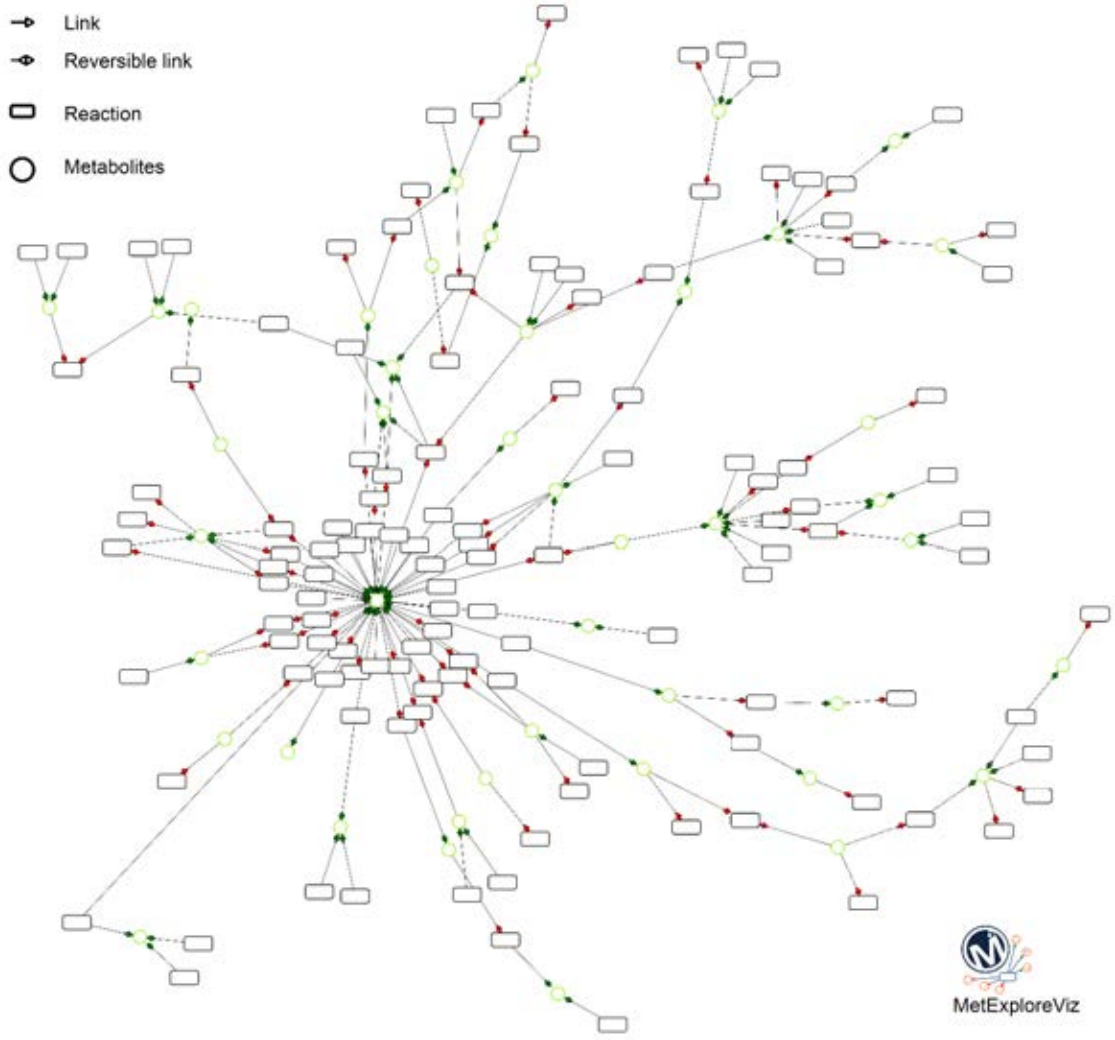


Figure 5.3: Visualisation of the Lead active module #1. Generated using MetExploreViz (Chazalviel et al, 2017).

5.3.2. KEGG analysis

The identified active modules represent previously unknown areas of the *D. magna* metabolome that are predicted to be affected by the Lead and Carbaryl exposures. Chapter 7 presents a metabolomics study that seeks to validate these predictions experimentally, as such, the predictions need to be placed into a context that can be validated experimentally.

A total of 178 metabolites are contained within the identified active modules. Metabolite annotation is a significant challenge for metabolomics studies with even the most

advanced techniques resulting in low percentages of the measured signals being annotated (Benton et al, 2015; Stanstrup et al, 2013; Tautenhahn et al, 2012). In anticipation of potential low annotation rates from the metabolomics dataset, the KEGG collection of databases and software tools (Kanehisa et al, 2004) is used to extrapolate the contents of the active modules to different levels.

The KEGG pathways database contain a large collection of metabolic pathways that contain metabolites and reactions. Each KEGG pathway consists of a number of KEGG modules, and also belongs to an area of metabolism. Figure 5.4 shows what this hierarchal structure looks like. The metabolites and reactions contained within the identified active modules are used along with the KEGG mapper software (Kanehisa, 2013) to identify KEGG modules, pathways and areas of metabolism have some representation in the identified active modules. The metabolomics datasets generated during the validation study will be analysed in the same way, to assess the effectiveness of the computationally generated hypotheses.

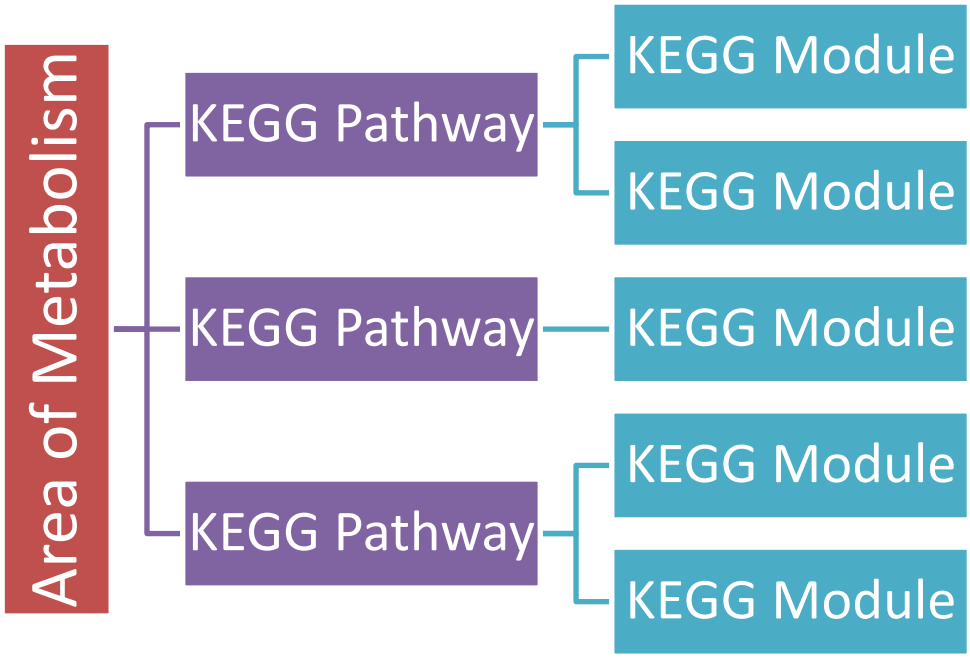


Figure 5.4: Relationship between areas of metabolism, KEGG Pathways and KEGG Modules.

For each of the stressors, the reactions and metabolites from the identified active modules are passed to KEGG mapper. KEGG modules that contain at least two metabolites or reactions represented in the detected AMBIENT modules are extracted. KEGG pathways that contain any of these modules and areas of metabolism to which these KEGG pathways belong are also extracted.

5.3.2.1. Carbaryl

A total of 14 AMBIENT active modules are identified using the Carbaryl STRESSFLEA transcriptomic data, containing 193 reactions and 113 metabolites. In total 18 KEGG modules were found to contain at least two metabolites or reactions in the AMBIENT active modules. The coverage of these KEGG modules in terms of reactions and metabolites for the *D. magna* draft GWMR and AMBIENT active modules is shown in Table 5.3.

Table 5.3: KEGG modules that contain at least two metabolite or reactions in the AMBIENT active modules for the *D. magna* draft GWMR scored with the Carbaryl STRESSFLEA dataset. The overall coverage of each KEGG module in the *D. magna* draft GWMR and the coverage of the KEGG module in the AMBIENT active modules is shown.

Module		GWMR Coverage	AMBIENT module coverage
M00001	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	100.00%	9.09%
M00004	Pentose phosphate pathway (Pentose phosphate cycle)	100.00%	18.18%
M00014	Glucuronate pathway (uronate pathway)	90.91%	27.27%
M00066	Lactosylceramide biosynthesis	100.00%	66.67%
M00068	Glycosphingolipid biosynthesis, globo-series, LacCer => Gb4Cer	100.00%	33.33%
M00073	N-glycan precursor trimming	100.00%	33.33%
M00078	Heparan sulfate degradation	70.00%	10.00%
M00104	Bile acid biosynthesis, cholesterol => cholate/chenodeoxycholate	100.00%	21.43%
M00114	Ascorbate biosynthesis, plants, glucose-6P => ascorbate	80.00%	10.00%
M00116	Menaquinone biosynthesis, chorismate => menaquinone	47.37%	26.32%
M00117	Ubiquinone biosynthesis, prokaryotes, chorismate => ubiquinone	36.84%	21.05%
M00128	Ubiquinone biosynthesis, eukaryotes, 4-hydroxybenzoate => ubiquinone	52.94%	23.53%
M00129	Ascorbate biosynthesis, animals, glucose-1P => ascorbate	88.89%	22.22%
M00131	Inositol phosphate metabolism, Ins(1,3,4,5)P4 => Ins(1,3,4)P3 => myo-inositol	100.00%	20.00%
M00372	Abscisic acid biosynthesis, beta-carotene => abscisic acid	18.18%	9.09%
M00549	Nucleotide sugar biosynthesis, glucose => UDP-glucose	100.00%	25.00%
M00563	Methanogenesis, methylamine/dimethylamine/trimethylamine => methane	28.57%	28.57%
M00565	Trehalose biosynthesis, D-glucose 1P => trehalose	85.71%	57.15%

Table 5.4 shows the KEGG pathways and areas of metabolism that are predicted to be affected by the Carbaryl treatment. 15 KEGG pathways that contain at least one of the 18 identified KEGG modules and at least two metabolites or reactions from the AMBIENT active modules are identified. 6 areas of metabolism that the 15 KEGG pathways belong

to are also identified. Visualisations of these 15 KEGG pathways are shown in Appendix B.

Table 5.4: KEGG pathways that contain at least contain at least one of the 18 identified KEGG modules and at least two metabolites or reactions from the AMBIENT active modules for the *D. magna* draft GWMR scored with the Carbaryl STRESSFLEA dataset. The area of metabolism that each pathway belongs to is also identified.

KEGG Pathway	**Area of Metabolism**
map00010: Glycolysis / Gluconeogenesis	Carbohydrate metabolism
map00030: Pentose phosphate pathway	Carbohydrate metabolism
map00040: Pentose and glucuronate interconversions	Carbohydrate metabolism
map00053: Ascorbate and aldarate metabolism	Carbohydrate metabolism
map00120: Primary bile acid biosynthesis	Lipid metabolism
map00130: Ubiquinone and other terpenoid-quinone biosynthesis	Metabolism of cofactors and vitamins
map00500: Starch and sucrose metabolism	Carbohydrate metabolism
map00510: N-Glycan biosynthesis	Glycan biosynthesis and metabolism
map00520: Amino sugar and nucleotide sugar metabolism	Carbohydrate metabolism
map00531: Glycosaminoglycan degradation	Glycan biosynthesis and metabolism
map00562: Inositol phosphate metabolism	Carbohydrate metabolism
map00600: Sphingolipid metabolism	Lipid metabolism
map00603: Glycosphingolipid biosynthesis - globo and isoglobo series	Glycan biosynthesis and metabolism
map00680: Methane metabolism	Energy metabolism
map01200: Carbon metabolism	Carbon metabolism

5.3.2.2. Lead

A total of 6 AMBIENT active modules are identified using the Lead STRESSFLEA transcriptomic data, containing 165 reactions and 65 metabolites. In total 10 KEGG modules were found to contain at least two metabolites or reactions in the AMBIENT active modules. The coverage of these KEGG modules in terms of reactions and metabolites for the *D. magna* draft GWMR and AMBIENT active modules is shown in Table 5.5.

Table 5.5: KEGG modules that contain at least two metabolite or reactions in the AMBIENT active modules for the *D. magna* draft GWMR scored with the Lead STRESSFLEA dataset. The overall coverage of each KEGG module in the *D. magna* draft GWMR and the coverage of the KEGG module in the AMBIENT active modules is shown.

Module		**GWMR Coverage**	**AMBIENT module coverage**
M00028	Ornithine biosynthesis, glutamate => ornithine	50.00%	33.33%
M00029	Urea cycle	92.86%	21.43%
M00057	Glycosaminoglycan biosynthesis, linkage tetrasaccharide	100.00%	44.44%
M00117	Ubiquinone biosynthesis, prokaryotes, chorismate => ubiquinone	36.84%	21.05%
M00128	Ubiquinone biosynthesis, eukaryotes, 4-hydroxybenzoate => ubiquinone	52.94%	23.53%
M00130	Inositol phosphate metabolism, PI=> PIP2 => Ins(1,4,5)P3 => Ins(1,3,4,5)P4	100.00%	33.33%
M00134	Polyamine biosynthesis, arginine => ornithine => putrescine	100.00%	60.00%
M00142	NADH:ubiquinone oxidoreductase, mitochondria	100.00%	33.33%
M00144	NADH:quinone oxidoreductase, prokaryotes	100.00%	33.33%
M00151	Cytochrome bc1 complex respiratory unit	100.00%	50.00%

Table 5.6 shows the KEGG pathways and areas of metabolism that are predicted to be affected by the Lead treatment. 8 KEGG pathways that contain at least one of the 10

identified KEGG modules and at least two metabolites or reactions from the AMBIENT active modules are identified. 5 areas of metabolism that the 8 KEGG pathways belong to are also identified. Visualisations of these 8 KEGG pathways are shown in Appendix B.

Table 5.6: KEGG pathways that contain at least contain at least one of the 10 identified KEGG modules and at least two metabolites or reactions from the AMBIENT active modules for the *D. magna* draft GWMR scored with the Lead STRESSFLEA dataset. The area of metabolism that each pathway belongs to is also identified.

KEGG Pathway	**Area of Metabolism**
map00130: Ubiquinone and other terpenoid-quinone biosynthesis	Metabolism of cofactors and vitamins
map00220: Arginine biosynthesis	Amino acid metabolism
map00330: Arginine and proline metabolism	Amino acid metabolism
map00480: Glutathione metabolism	Metabolism of other amino acids
map00532: Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	Glycan biosynthesis and metabolism
map00562: Inositol phosphate metabolism	Carbohydrate metabolism
map01210: 2-Oxocarboxylic acid metabolism	2-Oxocarboxylic acid metabolism
map01230: Biosynthesis of amino acids	Amino acid metabolism

5.4. Discussion and conclusion

The AMBIENT active modules algorithm (Bryant et al, 2013b) is used to identify sub-modules within the *D. magna* draft GWMR (presented in Chapter 4) using two of the transcriptomic data sets from the STRESSFLEA consortium project (Orsini et al, 2016). The two environmental stressors chosen, Carbaryl and Lead, are human induced through their use in industry and agriculture and have had significant effects on ecological systems. AMBIENT active modules identify several metabolites and reactions that are predicted to be affected by the selected environmental stressors. A total of 178 metabolites and 358 reactions make up these predictions. KEGG pathway analysis using these metabolites and reactions reveals a number of KEGG modules that are predicted to be affected. These KEGG modules are mapped to KEGG pathways which in turn belong to wider areas of metabolism (Figure 5.4).

Table 5.7: Summary of the computationally generated hypotheses. The contents of AMBIENT modules are used to derive KEGG modules, pathways and areas of metabolism. These form predictions of how the *D. magna* metabolome is affected at different levels of granularity.

	Carbaryl	**Lead**
AMBIENT Modules	14	6
KEGG Modules	18	10
KEGG Pathways	15	8
Areas of Metabolism	6	5

Table 5.7 summarises the number of KEGG modules, pathways and areas of metabolism that are predicted to be affected by the Carbaryl and Lead treatments. These predictions form computationally generated hypotheses about areas of the *D. magna* metabolome that are affected under the tested conditions. There are three levels that can be analysed; areas of metabolism, KEGG pathways and KEGG modules, each with increasing granularity.

It is worth noting that there are limitations to using transcriptomics data in the way in which it used by the AMBIENT algorithm. Transcriptomics data measures gene expression by looking at the abundance of mRNA, it does not measure protein abundance directly. The AMBIENT method assumes that if a gene is highly expressed, then the relevant enzymatic proteins are more active. This assumption may not always hold true (Horgan & Kenny, 2011). Ideally a proteomics study should also take place to validate these assumptions. This would however add significant cost and complication to any study (Petricoin et al, 2002a; Petricoin et al, 2002b).

An alternative approach would be to use GSEA based pathway enrichment. This however relies heavily on GO terms and on rigid pre-defined metabolic pathways or ontologies that do not necessarily represent the highly interconnected nature of the various metabolic networks. It does not leverage a key benefit of representing a metabolome using a GWMR, that pre-defined pathways are not imposed on the network. The technique used by AMBIENT can reveal the coordinated response of a biological network to the observed changes in gene expression in an unbiased way.

Chapter 7 details a metabolomics study that seeks to validate these predictions experimentally.

6. Closed-loop Optimisation of Liquid-Chromatography Mass Spectrometry

The preceding chapters in this thesis detail the draft GWMR of *D. magna* and it use for a computational toxicology study to predict the unknown metabolic response of *D. magna* to two environmental stressors. This resulted in computationally generated hypotheses that predict the effect of the stressors on the *D. magna* metabolome. To validate this approach a metabolomics study will be carried out (Chapter 7), and as one of the principal analytical technique for metabolomics, LC-MS will be used to make the metabolomics measurements.

Standard untargeted LC-MS methods exist that can be applied to this study, however in this case a LC-MS method that is optimised to detect as many of the metabolites that are predicted to be effected would be beneficial. LC-MS method development is not a trivial task and significant time and expertise. In this chapter, the MUSCLE (Multi-platform Unbiased optimisation of Spectrometry via Closed Loop Experimentation) software platform for automated closed-loop optimisation of LC-MS is presented. MUSCLE is designed in a modular way so that automated closed loop optimisation can be applied to targeted and untargeted LC-MS method optimisation across a range of LC-MS systems from any instrument manufacturer. MUSCLE can also be configured so that different optimisation algorithms can be applied. A modified version of the PESA-II algorithm is applied, and an extension PESA-II-FS is presented and also applied.

MUSCLE is used to optimise a LC-MS method for use in the *D. magna* computational toxicology study presented in this thesis using a semi-targeted optimisation approach.

MUSCLE is also used to optimise several LC-MS methods for both targeted and untargeted analyses.

6.1. Introduction

LC-MS is widely used in analytical laboratories for measuring a range of (bio)chemicals and as the principal technology for metabolomics and proteomics. There are two distinct types of mass spectrometry based analytical approaches, targeted and untargeted (Figure 6.1). Targeted analyses measure predefined number of chemically characterised and biochemically annotated molecules (Roberts et al, 2012), whereas untargeted analyses measure any molecule that can be ionized within a defined range of m/z values (Vinayavekhin & Saghatelian, 2010). Each approach has advantages and limitations. A targeted approach provides better quantification of known molecules at lower detection limits but does not allow for the discovery of unknown compounds. An untargeted approach provides a more comprehensive global measurement of the molecules within a sample but requires far more intricate informatics approaches to interpret the results (Menni et al, 2017).

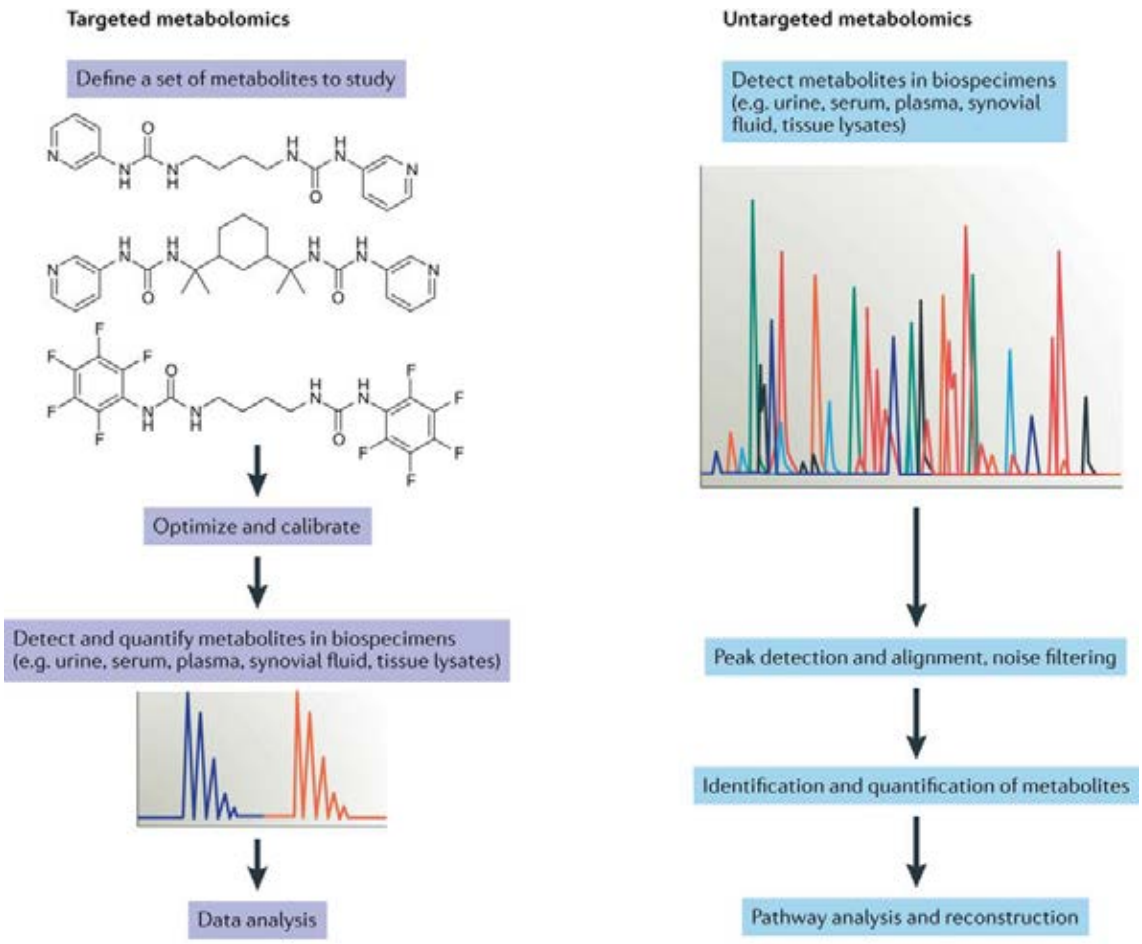


Figure 6.1: (Menni et al, 2017) Differences between targeted and untargeted metabolomics using MS. For targeted analysis, the metabolites of interest are predefined, and the MS is configured to measure these metabolites only. For untargeted analysis, all metabolites that are contained within the sample are measured. Untargeted analysis requires for the measured peaks to be assigned to metabolites.

Regardless of the LC-MS approach employed, development of new LC-MS methods or the transfer of existing methods between instruments and laboratories is time consuming and challenging. Simply put, this is because of the large number of LC and MS parameters that require optimisation. Varying all the possible parameters systematically to optimise the analysis of selected molecules is generally regarded as impossible because of the large search space created by the possible values for LC and MS parameters.

The traditional approach to achieving an LC-MS method involves a MS operator either fine tuning an existing method to suit the current set of needs, or developing a new method based on their expertise and experience with the analytical platform and the type of sample being used (Crowson & Beardah, 2001; Gassner & Weyermann, 2016; Koster et al, 2014; Vatansever et al, 2017; Widmer et al, 2002; Zelena et al, 2009; Zonaras et al, 2016). This inevitably introduces a human induced bias as MS operators draw from their past experiences when designing new LC-MS methods.

The problem of LC-MS method optimisation can be formulated as a multi-objective heuristic search problem that is suitable for optimisation using an evolutionary search based closed-loop optimisation approach (O'Hagan et al, 2005). Previous implementations of closed-loop optimisation of MS methods (both LC-MS and Gas Chromatography Mass Spectrometry (GC-MS)) were for a specific manufacturer's analytical platform (O'Hagan et al, 2005; O'Hagan et al, 2007; Zelena et al, 2009). Extending this to further instruments would require extensive reprogramming, therefore significantly limiting the deployability and suitability of this approach for use with the wide range of LC-MS analytical platforms.

Here the software platform MUSCLE (Multi-platform Unbiased optimisation of Spectrometry via Closed Loop Experimentation) (Bradbury et al, 2015) for fully automated closed-loop optimisation of LC-MS method development is presented. MUSCLE is completely instrument-manufacturer independent and can be used to optimise both targeted and untargeted analyses. MUSCLE is designed in a modular way so that different data processing pipelines, objective functions and algorithms can be used depending on the type of analysis, the analytical platform used and the requirements of the study. The application of MUSCLE is demonstrated across a range of LC-MS

platforms for both targeted and untargeted analyses. A method optimisation using a hybrid of these analysis types (semi-targeted) is also presented. In each instance MUSCLE optimised methods provide an improvement on manually optimised methods.

6.2. Methods

MUSCLE is a standalone desktop application written in the Java programming language. User-defined Visual Scripts imitate the keyboard and mouse commands that an analyst would use to manually change parameters and launch an LC-MS/MS analysis, enabling MUSCLE to control multiple LC and MS parameters on any instrument (Section 6.2.1). Once an automated optimisation is set up using the software, the closed-loop optimisation process begins (Figure 6.2). A multi-objective genetic algorithm (MOGA) optimises the values of the LC and MS parameters, based upon the values of user-defined objective functions that measure, e.g., analytical sensitivity and analysis time (Section 6.2.3). The data processing is handled in a modular way so that it can be customised depending on the LC-MS platform used and the type of analysis (Section 6.2.2).

Several components need to be configured to run a MUSCLE optimisation (Figure 6.3). An optimisation consists of an experiment and a configuration. An experiment describes the optimisation parameters and the mechanisms for changing the values using visual scripts (Section 6.2.1). Each experiment contains at least one visual script, which defines a set of keyboard and mouse commands used to enter instrument parameters. An experiment also contains parameter values for each optimisation parameter defined in a visual script. For each optimisation parameter, a minimum, maximum and step value is defined. These are used to define a possible set of discrete values that can be used for any of the optimisation parameters. For example, if for an optimisation parameter the

minimum value is 2.0, the maximum 3.0 and the step 0.5, the possible values are 2.0, 2.5 and 3.0.

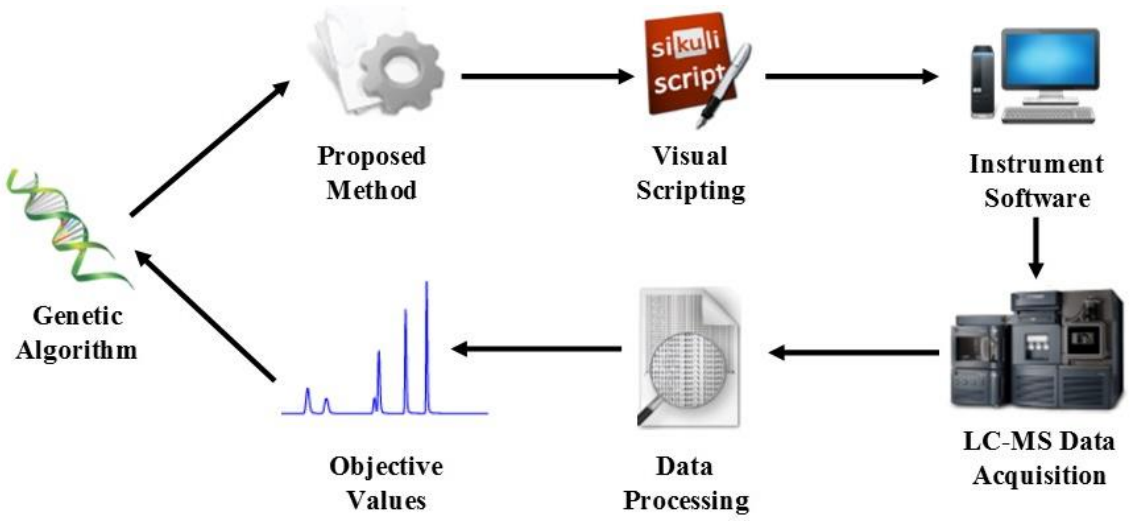


Figure 6.2: MUSCLE closed-loop optimisation process. The genetic algorithm decides on a value for each of the LC and MS parameters which forms the proposed method. The proposed method is then input into the instrument software and the LC-MS data acquisition is initiated using visual scripting (Section 6.2.1). The output data is then processed (Section 6.2.2) and assessed based on pre-defined objective functions (Section 6.2.3). The objective value information is passed back into the genetic algorithm and the process repeats until the desired number of LC-MS injections are complete.

A configuration defines the optimisation algorithm and objective functions along with how the output data for each LC-MS injection is processed. It includes an algorithm, a set of objective measures and a data processing configuration. The algorithm defines which optimisation algorithm is being used along with some algorithm specific parameters e.g. crossover and mutation rates. Up to three objective measures are selected for each optimisation (Section 6.2.3). The data processing configuration (Section 6.2.2) defines which type of analysis (e.g. targeted or untargeted) is being used along with some analysis type specific settings such as hyper-parameter values or paths to .bat files.

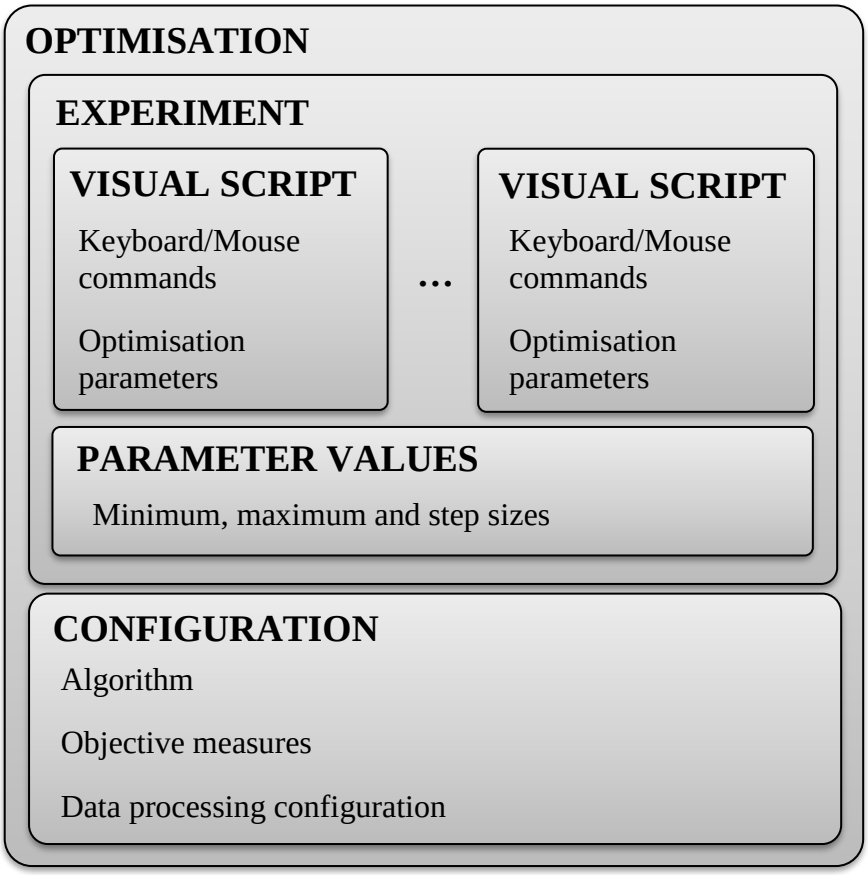


Figure 6.3: MUSCLE optimisation architecture. An optimisation consists of an experiment and a configuration. The experiment defines the optimisation parameters and the mechanisms for changing the corresponding values using the instrument software. A configuration defines the optimisation algorithm and objective functions along with how the output data for each LC-MS injection is processed.

6.2.1. Visual Scripting

To make MUSCLE LC-MS platform independent, a visual scripting approach is used. This is where there is no direct communications link between the application and the software that controls the LC-MS instrument. Instead visual scripting enables direct visual references to be made to objects displayed on the screen, e.g. a File menu item or button, and allows MUSCLE to mimic the keyboard and mouse actions that a user would make when operating an LC-MS instrument. The java library Sikuli (Yeh et al, 2009) is used, providing a powerful and flexible API to allow users to create visual scripts that

can: left/double/right click on selected objects on the screen, enter text into text fields, and press keyboard keys e.g. enter and backspace.

A wrapper around Sikuli allows for generation of visual scripts as a set of user defined commands (Table 6.1) that either enter LC or MS parameters or initiate LC-MS runs automatically. Click, double click and right click commands are defined by the user by selecting a region of the computer screen. The selected region is saved as an image and when the visual script is run, the region of the screen that matches the image is found, the mouse is moved to that region and then the click is performed. Figure 6.4 shows how the screen region for a click-based command is defined.

Figure 6.5 shows a visual script for changing the LC gradient parameters for an optimisation (taken from optimisation described in Section 6.3.3), and Figure 6.6 shows how the minimum, maximum and step values for optimisation parameters are defined.

Table 6.1: List of possible visual script commands.

Command Type	**Description**
Click	Click a region on the screen defined by a user selected image.
Double Click	Double click a region on the screen defined by a user selected image.
Right Click	Right click a region on the screen defined by a user selected image.
Enter Text	Enter some user defined text.
Press a Key	Press a keyboard key e.g. Enter, tab or delete.
Enter Value	Enter a value for an optimised parameter. The possible values that can be entered are defined in the experiment configuration.

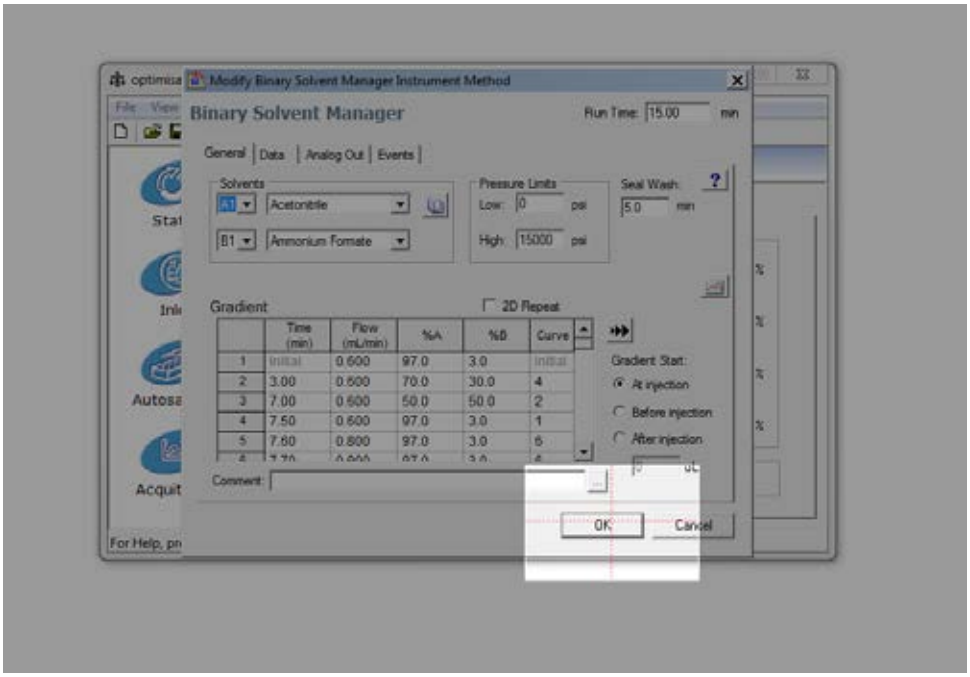


Figure 6.4: For click based commands, the user selects a region of the screen which is saved as an image. When the command is run, the centre of that image is found (indicated by the red cross) and then the click is performed.

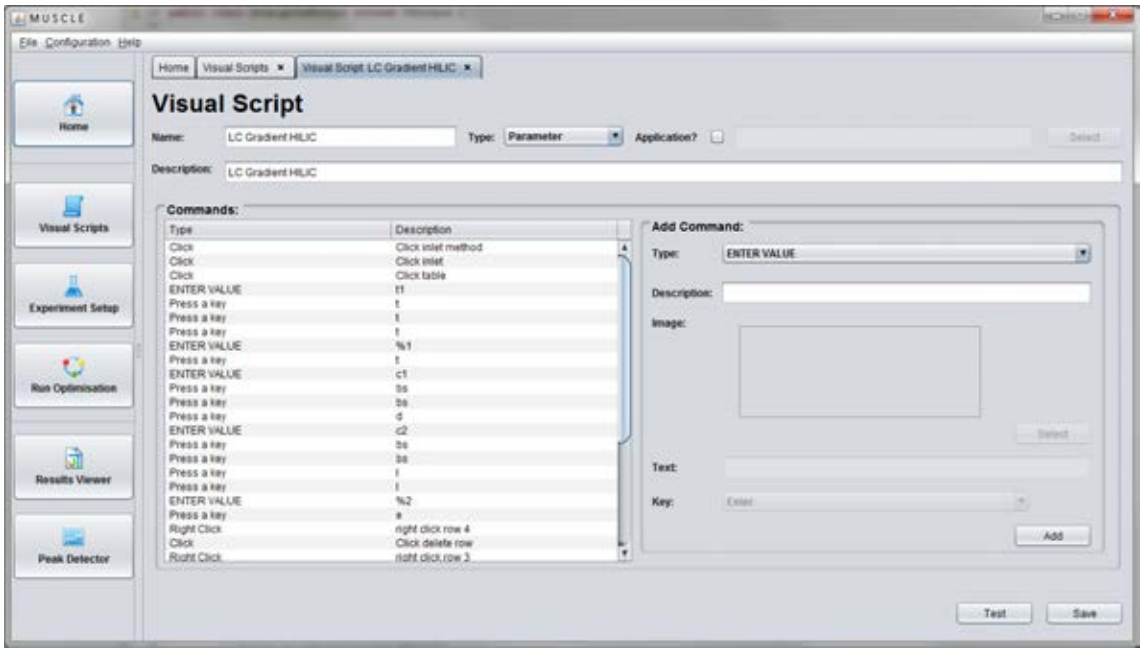


Figure 6.5: Screenshot from MUSCLE software showing a visual script for changing optimisation parameters associated with a LC gradient (taken from the experiment in Section 6.3.3). The commands table lists the commands in the visual script, e.g. click commands and press-a-key commands. Enter value commands relate to optimisation parameters and must have minimum, maximum and step values defined for them for any experiment using the visual script (Figure 6.6).

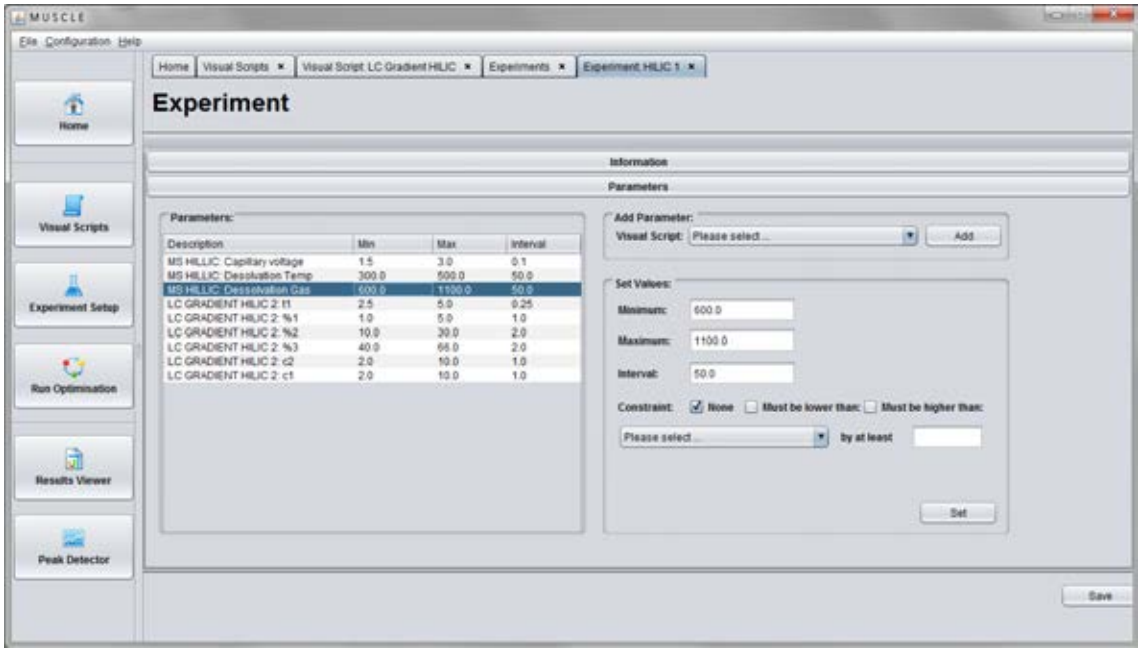


Figure 6.6: Screenshot from MUSCLE software showing how minimum, maximum and step values are defined for enter value commands in visual scripts.

This visual scripting approach is chosen as otherwise there would need to be programmatic access to the software working that controls the LC-MS instruments. This would require extensive customisation for each new instrument that is used. Visual scripting provides a solution that works in the same way for all instruments and is therefore future proof.

The main limitation of using this visual scripting approach is that it relies on the generation of visual scripts which look to identify pre-defined images that should be visible on a computer screen. If these images are not visible on the screen e.g. if a message box appears, then the visual script cannot continue. The visual scripts must be developed meticulously to avoid any such issues occurring. Despite this limitation, the visual scripting approach is preferred as the alternative would be to gain programmatic access to the instrument software. This would limit the ability of MUSCLE to be used on any

analytical platform, as different programmatic access to each instrument would be required and would likely need to be configured in a different way.

6.2.2. Data Processing

Processing of LC-MS data is handled in a modular fashion. This allows the MUSCLE platform to be used for any type of LC-MS analysis with little modification. The heuristic search algorithm used for optimisation requires a mechanism for assessing the quality of each physical closed loop experiment, in this case an LC-MS run. This is achieved using the objective functions outlined in section 6.2.3. The data processing module acts as an intermediary between the LC-MS output data (hereby referred to as raw data) and the objective functions module.

The data processing module has two components. The first component converts the raw data to an mzML file. The mzML format is an open-source file format for mass spectrometry data that is platform independent (Martens et al, 2011). File conversion is carried out using the msConvert application, which is part of the ProteoWizard toolkit (Chambers et al, 2012). msConvert has several program arguments that may be required depending on the type of analysis and the analytical platform being used. These can be set by modifying a .bat file, which runs the file conversion after each LC-MS run. Conversion to mzML is carried out for both targeted and untargeted analysis.

The second component of the data processing module is responsible for analysing the mzML file and assessing the quality of the corresponding chromatogram. The approach used is different depending on the type of analysis but the overall objective is the same, to detect the features present within the injected sample along with some information about them e.g. retention time and intensity.

6.2.2.1. Targeted Analysis

For targeted analysis, the MS instrument is configured to monitor selected transitions from a precursor ion to a product ion arising through fragmentation. This acquisition method is termed Selected Reaction Monitoring (SRM) and results in a signal being generated for each of the selected product ions in the mzML file, with a feature being represented as a peak within a signal. The detection of features thus becomes a two-dimensional peak detection problem for a series of signals, as opposed to untargeted analysis, whereby the signals contained within the mzML file represent a three-dimensional surface in which to detect peaks.

A peak detection algorithm (Algorithm 6.1), is implemented to detect features within a mzML file containing multiple SRM signals.

```
1: function FindPeaks(mzML, slopeThreshold, noiseThreshold, peakGroup)
2:     for each Signal s in mzML do
3:         noiseThreshold = calculateNoiseThreshold(s)
4:         f = firstDerivative(s)
5:         f = smoothSignal(f)
6:         z = zeroCrossingPoints(f)
7:         for each Point p in z do
8:             if f(p) − f(p + 1) > slopeTheshold then
9:                 if s(p) > noiseThreshold then
10:                    fit polynomial for points between s(p − peakGroup) and s(p + peakGroup)
11:                    calulate peak width and area
12:                    store peak height, retention time, width and area
13:                end if
14:            end if
15:        end for
16:    end for
17: end function
```

Algorithm 6.1: Targeted peak detection algorithm.

The first part of each signal is analysed to derive a value for a noise threshold. Each peak must be at least higher than this threshold to be considered valid. The noise threshold is simply the median intensity value for the first x values in the signal. x is a free parameter

to be set but is typically set so that approximately the first half of a minute is checked. What value constitutes half of a minute of signal depends on the instrument and its scan rate.

The first derivate of the signal (f) is then calculated, this is as the points at which the derived signal crosses the *x* axis, so called zero-crossing points, correspond to a local maxima and therefore the top of a peak (Nguyen et al, 2010). A moving average filter (Shumway & Stoffer, 2010) is applied to smooth f. The formula shown in (1) is applied three times to the signal, with each point in the signal being replaced by the average of M adjacent points. This has the effect of removing very small peaks in the signal that are a result of noise (Figure 6.7).

$$y(i) = \frac{1}{M} \sum_{j=-\frac{M}{2}}^{\frac{M}{2}} x(i+j) \tag{1}$$

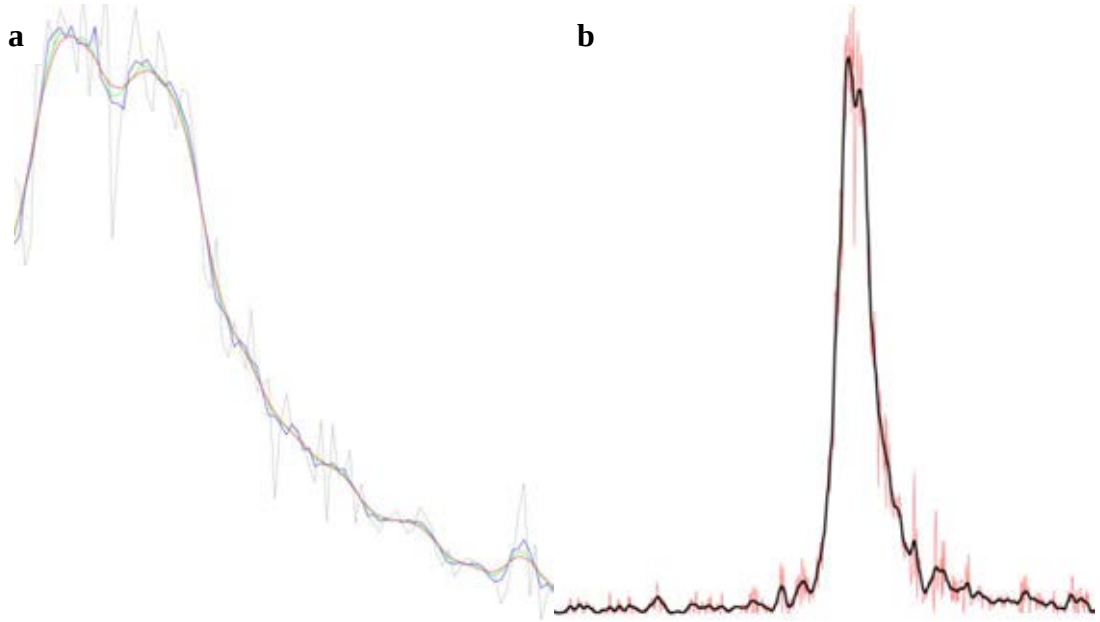


Figure 6.7: Effect of smoothing using moving average filter. Figure a demonstrates the effect of each application of the smoothing filter. The original signal is shown is black, the first pass of the filter is in blue, the second green and the third red. Figure b shows a targeted signal that has had the moving average filter applied three times. The red line is the original signal, and the black is the smoothed signal.

The zero crossing points are then found for the smoothed first derivative signal, a zero-crossing point corresponds to the top of a peak. The slope around the zero-crossing point is then checked to see if it is higher than a pre-defined slope threshold. This is as ideally an LC-MS peak should be sharp and not elongated (Zhang et al, 2009)**.**

If the value of the original signal at the corresponding zero-crossing point in the derivative signal is greater than the calculated noise threshold, a pre-defined number of points around the top point of each peak (defined by the peakGroup argument) are used to fit a polynomial. The polynomial coefficients are then used to calculate the peak width and area. For each signal, just the most intense peak is selected. For each peak, the height, width, chromatographic retention time, and area is stored.

6.2.2.2. Untargeted Analysis

For untargeted analysis, a global measurement of all molecules within a sample is taken. Thus, analysing the mzML file to detect peaks is a far greater challenge as there are potentially thousands of features to analyse as opposed to typically tens of features in a targeted analysis.

XCMS (Smith et al, 2006; Tautenhahn et al, 2008), part of the Bioconductor R package (Gentleman et al, 2004) is a software package for untargeted LC-MS feature detection and is one of the most widely used data processing tool for untargeted metabolomics (Benton et al, 2010; Kurczy et al, 2015). The MUSCLE platform can be configured to run any of the feature detection algorithms in XCMS by modifying a .bat file. The output of the XCMS feature detection algorithms is a .csv file with each row representing an LC-MS feature. The columns store information about each feature such as retention time, peak intensity and signal to noise ratio. The .csv file is parsed by MUSCLE with the information contained within being used to calculate the objective measures.

6.2.3. Objective Measures

The objective measures are also handled in a modular manner and can be configured based on the study and the type of analysis being conducted. The aim of each objective measure is to provide a measurement of quality of the output of the current LC-MS run. Each objective measure is either set to be maximised or minimised. For maximise objectives, a higher value represents a better quality output, whereas for minimise objectives, the inverse is true.

As the data processing procedures used for targeted and untargeted analysis (Section 6.2.2) give different outputs, different objective measurements need to be used depending

on whether a targeted or untargeted LC-MS method optimisation is being performed. Targeted objective measurements are set or calculated based on a list of peaks that is generated using the procedure outlined in section 6.2.2.1, whereas untargeted objective measurements are set or calculated based on the XCMS output that is parsed by MUSCLE (Section 6.2.2.2). Table 6.2 lists several objective measures for both targeted and untargeted analyses. Also listed is a semi-targeted objective, which is explained in section 6.3.2. An explanation of the objective functions and how they are used in each particular study is provided in section 6.2.3.

Table 6.2 Example objective measures. Objectives differ for each analysis type as they are calculated based on different objects. Each objective is set to be either maximised or minimised.

Analysis Type	Objective Type	Description
Targeted	Maximise	Total peak count
Targeted	Minimise	Run time
Targeted	Maximise	Total peak area
Targeted	Maximise	Average peak intensity
Untargeted	Maximise	Total feature count
Untargeted	Maximise	Total peak area
Untargeted	Maximise	Average peak area
Untargeted	Maximise	Average signal to noise
Untargeted	Maximise	Chromatographic separation
Semi-Targeted	Maximise	Number of pre-defined m/z values present in feature set

6.2.4. Closed-Loop Optimisation

Closed-loop evolutionary optimisation is a probabilistic search heuristic, whereby potential solutions are evaluated by conducting physical experiments (Knowles, 2009), which in the case of MUSCLE corresponds to LC-MS analyses. Each solution represents a set of control parameters for the LC-MS instrument and is generated using a MOEA. To evaluate each solution, a fitness value is calculated for each of the objective measures, with each one measuring some aspect of the quality of the LC-MS spectra obtained using

the GA selected instrument settings. Because the objectives are in conflict, a multi-objective GA must be used, which can efficiently find a set of Pareto optimal solutions.

Typically, GAs evaluate tens of thousands of solutions *in silico* during an optimisation process, with many optimisation experiments required to achieve a highly optimised search method. Because of the time and cost involved in evaluating each solution in an LC-MS study, conducting large numbers of optimisation experiments is not feasible. Therefore, in this instance, as is the case with many closed-loop optimisation approaches, a GA is required that can perform well when limited to just a few tens or hundreds of evaluations.

The principles of exploration and exploitation are the foundations of heuristic search. Exploration means visiting new and unexplored regions of the search space, whereas exploitation means visiting regions within the search space around previously visited points (Črepinšek et al, 2013). In any heuristic search, it is important to balance the exploration and exploitation of the search space, and this especially pertinent when conducting closed-loop optimisation due to the limited number of evaluations. When selecting an appropriate algorithm for this closed-loop problem it is important therefore to consider how the balance of exploration and exploitation can be manipulated to achieve the best result. The PESA-II (Corne et al, 2001), ParEgo (Knowles, 2006) and NSGA-II (Deb et al, 2002) MOEAs have all been applied to closed-loop optimisations (O'Hagan et al, 2005; O'Hagan et al, 2007; Small et al, 2011). The mechanisms of the PESA-II algorithm (Section 6.2.4.1) allow the balance of exploration and exploitation to be manipulated, and is therefore selected as the MOEA for use in MUSCLE. The open source implementation of the algorithm, along with the general MOEA framework provided by the jMetal library (Durillo & Nebro, 2011) is modified for use in MUSCLE.

6.2.4.1. PESA-II

The PESA-II algorithm (Corne et al, 2001) operates like a standard EA in the manner that two populations are maintained, an internal and an external population. The internal population is of fixed size and at each generation it stores new solutions generated from the external population. The external population (also known as and herby referred to as the archive set), only contains the non-dominated solutions that have so far been discovered from the heuristic search. A solution is deemed to be non-dominated if none of the values for the objective measures can be improved without the degradation of any of the other objective values (Kirlik & Sayın, 2014). For each generation of the algorithm, new solutions to be evaluated are generated by selecting solutions from the archive, which then have crossover and mutation operators applied and subsequently form the internal population.

PESA-II implements a bin-based selection procedure to increase diversity. The non-dominated solutions in the archive are kept in bins based on their fitness values so that solutions with similar fitness values are grouped together. Selection is then carried out uniformly across these bins as opposed to selecting uniformly across the whole of the archive. This has the effect of favouring isolated solutions in the objective space. Selecting parent solutions from the archive of non-dominated solutions results in greater exploitation of previously visited points.

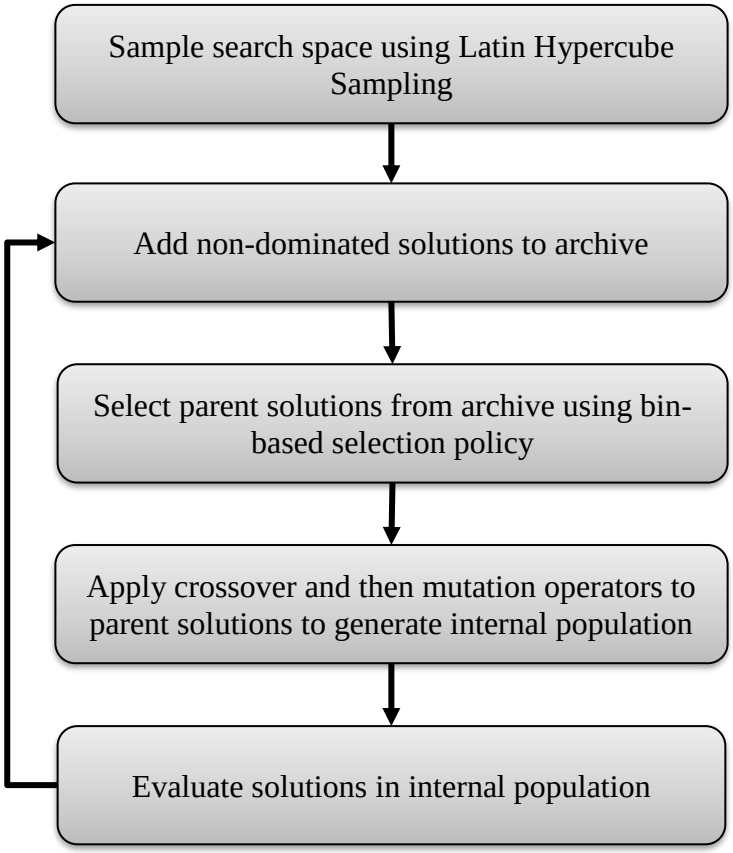


Figure 6.8: PESA-II algorithm with Latin-Hypercube sampling.

To initialise the archive, Latin Hypercube Sampling (LHS) (McKay et al, 2000) is used to generate the solutions for the first *n* runs, where *n* is defined for each optimisation in the optimisation configuration (Figure 6.3). The aim of LHS is to distribute the decision variable values evenly across the search space by using overlapping permutations of the possible values for the decision variables. LHS is used as an exploratory step before the PESA-II begins, with the hope of finding areas that contain local maxima that can then be exploited using the archive and bin-based selection. The number of solutions to be generated using LHS is optional and is typically between ten and twenty-five percent of the total solutions to be evaluated. Once the LHS evaluations are complete, the archive set is initialised by adding the solutions to it. If a solution is dominated by a solution that

is already in the archive set, it is not added. If a solution is non-dominated, it is added to the archive set. If adding a new non-dominated solution results in an existing solution in the archive set becoming dominated, the newly dominated solution(s) are removed from the archive set.

Subsequent values for the decision variables are generated using the PESA-II algorithm. An internal population is generated for each generation by selecting parent solutions from the archive set using the bin-based selection policy. As selection occurs on the non-dominated solutions in the archive set only, the parent solutions have high values for the objective functions. This means that areas of the search space that have shown to contain local maxima are exploited. The drawback of this method is that it increases the likelihood of getting stuck in the areas containing local optima, thus missing out on finding more global optima due to over exploitation and lack of exploration of the search space. To increase exploration of the search space, crossover and mutation operators are applied to the parent solutions to generate the new solutions to be evaluated.

The solutions are encoded for the GA using a binary representation, with a solution being represented by a single binary string containing a smaller substring for each parameter. To get a control parameter value the relevant binary substring is converted to a decimal number. Representing the solution using a binary string allows for efficient application of genetic operators. The crossover operator mimics breeding and takes two solutions (parent 1 and 2 which are represented as binary strings) of length y and picks a random point x such that $x < y$. The two binary strings are then cut at that point and a child solution is generated by taking the digits from before x from parent 1 and combining it with the digits after point x from parent 2, thus creating a new solution that is a combination of its two parent solutions. The mutation operator mimics genetic mutation

by choosing a random binary digit and flipping it, so if the digit is a 1, it is flipped to become a 0 and vice-versa.

6.2.4.2. PESA-II with Feature Selection

An extension to the PESA-II algorithm, PESA-II with Feature Selection (PESA-II-FS) is also proposed that uses feature selection to focus the heuristic search on decision variables that are deemed to be more influential to the overall fitness of the solutions. The intuition behind PESA-II-FS is that convergence of a heuristic search can be improved by focussing on the decision variables which have the strongest effect on the objective measures.

Feature selection is a procedure whereby a function for predicting the classification of a sample set is built using a training set (James et al, 2013). The principal behind feature selection is that several data features can be irrelevant or redundant to an observed classification and can therefore be removed from predictive models without the loss of too much information (Guyon & Elisseeff, 2003). In the context of LC-MS method development, feature selection can be used to identify instrument parameters that have relatively insignificant effects on the objective measures that determine the quality of each LC-MS run. Feature (or attribute) selection can be applied to select a subset of decision variables that have the greatest influence on the objective measures. The selected decision variables can then be the focus of the MOEA. As more evaluations are performed, the feature selection procedure can be re-run, thus re-selecting the subset of decision variables to focus the optimisation on.

Figure 6.9 describes the general procedure of the PESA-II-FS. The first step is the same as the modified PESA-II algorithm described in section 6.2.4.1. LHS is used to sample

the search space for the first n runs. The non-dominated solutions are then added to an archive. Feature selection is then performed to partition the decision variables. Variables that are deemed to be most influential to the objective measures are marked as selected variables, the others are marked as non-selected variables.

For each new evaluation, the values for the selected variables are generated using the PESA-II algorithm whereas the non-selected values have their values taken from a Latin Hypercube generated using only the values contained within the solutions in the archive set. This means that the data used for the next round of feature selection will contain differing values for all decision variables, but for the non-optimised decision variables, only values that have previously given favourable output (i.e. values that resulted in inclusion in the archive set) are used. If any of the solutions evaluated are non-dominated, they are added to the archive set using the same rules as PESA-II, i.e. the archive only contains non-dominated solutions.

The feature selection and decision variable partitioning is re-run after a predetermined number of evaluations, known as the round size. When the number of evaluations is equal to the round size, the decision variables are repartitioned into selected and non-selected variables once again. Figure 6.10 visualises how decision variables that are deemed to be influential on the objective measures may change after each round.

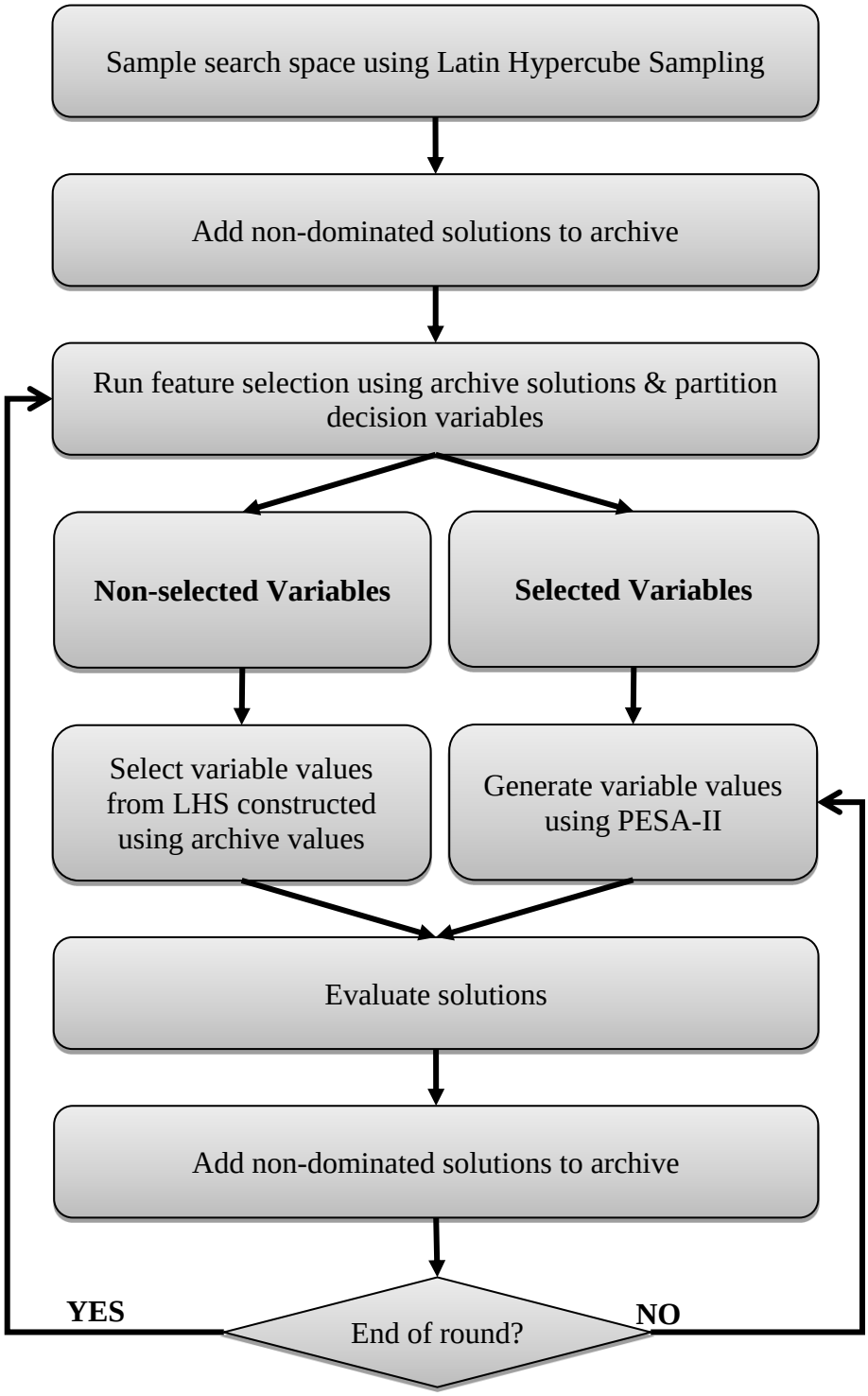


Figure 6.9: PESA-II-FS algorithm

For the MUSCLE implementation of PESA-II-FS, the WEKA java library (Hall et al, 2009; Smith & Frank, 2016) is used to perform feature selection. The wrapper subset

evaluation method (Kohavi & John, 1997) is used. This is an iterative technique which first selects features for classification, builds a multiple linear regression (MLR) model using the selected features and then assess the prediction capability of the model. Different combinations of features are tested, and the prediction rate of the model recorded. Due to the typically low number of data points involved with a closed-loop optimisation, an exhaustive search is used to find the set of features that have the best prediction rate, determined by the accuracy of the MLR model.

MLR is used to model the relationship between the exploratory variables and the response variable. In this case, the exploratory variables are the values used so far for the LC and MS parameters and the response variable is combined objective value outlined below. Once an MLR model has been built, the information can be used to create a prediction on the level of effect they each have on the outcome variable. The wrapper evaluation method selects the model that has the best cross-validated prediction accuracy, and the selected features used for the best-performing model (in this case LC and MS parameters) are selected for further optimisation using PESA-II (Figure 6.10).

Before feature selection is carried out, the values for the objective functions are normalised between 0 and 1 and then multiplied by a weighting factor. These values are then combined to form a single value for the objective functions. This sum of normalised and weighted objective measures is then used as the classification class value for the feature selection. The weights can be adjusted so that particular objective measures can have a lesser or greater impact on the assessment of the influence that the decision variables have on the combined objective measurements.

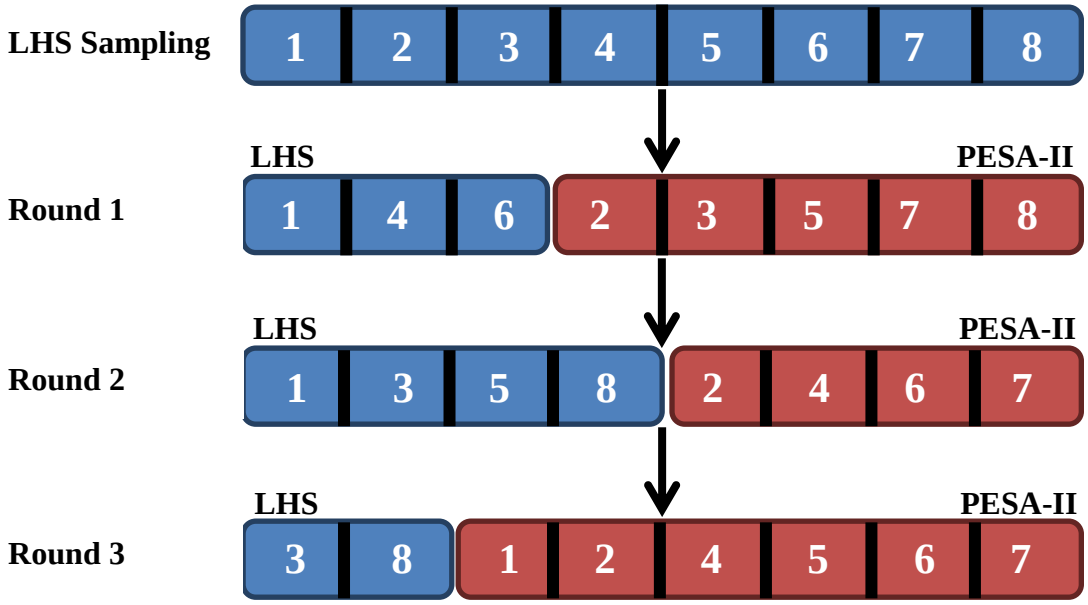


Figure 6.10: PESA-II with Feature Selection decision variable partitioning. After the initial LHS and after each round, feature selection selects variables that are deemed to have the most influence on the objective measures.

6.3. Results

MUSCLE closed-loop optimisations of LC-MS methods have been carried out for several different studies across a range of instruments and sample types (Table 6.3). Here the results of a targeted (section 6.3.1), semi-targeted (section 0) and an untargeted (section 6.3.3) closed-loop LC-MS method optimisation are presented. For each optimisation, a description is included that includes the aim of the optimisation, the optimisation parameters, the LC-MS variables and the objective measures. For each optimisation, the instrument parameters to be optimised are defined by giving a min max and step value.

Table 6.3: MUSCLE closed-loop LC-MS method optimisations.

Analysis Type	Description	Sample Type	Publication
Targeted	Detection of 6 steroids – 2 methods	Chemical standards	(Bradbury et al, 2015)
Targeted	Vitamin D analysis	Chemical standards	(Jenkinson et al, 2017)
Untargeted	HILIC method – 2 methods	Human urine	(Dunn et al, in preparation)
Targeted	Detection of 22 steroids	Chemical standards	(Taylor et al, 2015)
Targeted	Oestrogen analysis	Chemical standards	(Gilligan et al, 2014)
Targeted	Amino acids analysis	Chemical standards	N/A
Semi-Targeted	Daphnia metabolomics	Whole organism	Section 6.3.1

6.3.1. Targeted

Detailed descriptions of targeted MUSCLE optimisations are presented in two papers, which are summarised below:

BRADBURY J, GENTA-JOUVE G, ALLWOOD JW, DUNN WB, GOODACRE R, KNOWLES JD, HE S, VIANT MR (2015) MUSCLE: AUTOMATED MULTI-OBJECTIVE EVOLUTIONARY OPTIMIZATION OF TARGETED LC-MS/MS ANALYSIS. BIOINFORMATICS 31: 975-977

This paper (Appendix C) introduces the MUSCLE software package, including the concepts and implementations of visual scripting and closed-loop evolutionary optimisation. MUSCLE is demonstrated by optimising LC-MS/MS methods for the targeted analysis of a laboratory-prepared mixture of six difficult to chromatographically separate steroids (Figure 6.11) using two different manufacturers LC-MS/MS instruments and their associated software.

For both optimisations, the peak detection procedure outlined in section 6.2.2.1 is used. The objective measures for the optimisation are; maximise the number of separated peaks, maximise the total peak area and minimise run time. Run time is here defined as the

elution time of the last detected peak. Both optimisations included 200 injections and took approximately 48 hours. In both instances, the chromatograms in the final archive set are inspected manually and assessed by an analyst who then selected a preferred method. The preferred methods were then compared to methods that had previously been optimised by an experienced analytical chemist.

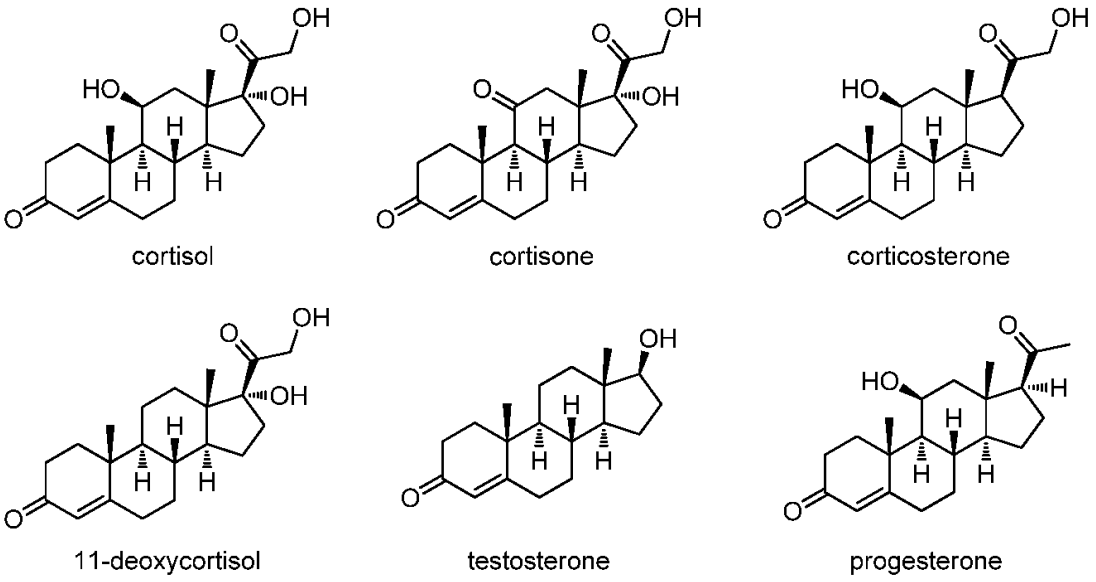


Figure 6.11: Selected steroids for LC-MS/MS optimisation.

In the first study, a Thermo Scientific UHPLC Ultimate 3000 coupled TSQ Vantage is used. The optimised method provided a faster (34.5%) and more sensitive (10%) analysis when compared to the manually optimised method. For the second study, a Waters ACQUITY UPLC Xevo TQ LC-MS/MS system is used. Again, the MUSCLE optimised method provided a faster (18.5%) and more sensitive (104%) analysis when compared to the manually optimised method. For more detailed information about the optimisation and instrument parameters, see Appendix C.

JENKINSON C, BRADBURY J, TAYLOR A, ADAMS JS, HE S, VIANT MR, HEWISON M (2017) AUTOMATED DEVELOPMENT OF AN LC-MS/MS METHOD FOR MEASURING MULTIPLE VITAMIN D METABOLITES USING MUSCLE SOFTWARE. ANALYTICAL METHODS 9: 2723-2731

In this paper (Appendix D), MUSCLE is used to optimise an LC-MS/MS method for the measurement of multiple vitamin D metabolites. The aim of this optimisation was to reduce the run time of a previously manually optimised method, to increase instrument throughput. The optimisation was performed on a Waters ACQUITY UPLC Xevo TQ LC-MS/MS system. A 200 injection optimisation was performed, taking approximately 30 hours.

The user-selected optimal run reduced the run time from 8.2 to 6.2 minutes, a 24% improvement, with the MUSCLE optimised method not compromising separation of any of the compounds with equal m/z. Furthermore, during analysis of the output files, an additional three minute method for the accurate quantification of the compound 25OHD3 was highlighted (Jenkinson et al, 2016). This additional method separated 25OHD3 from 3-epi-25OHD3, which has the same molecular weight.

Figure 6.12 shows the LC gradients and Figure 6.13 shows the chromatograms for the manually and MUSCLE optimised methods.

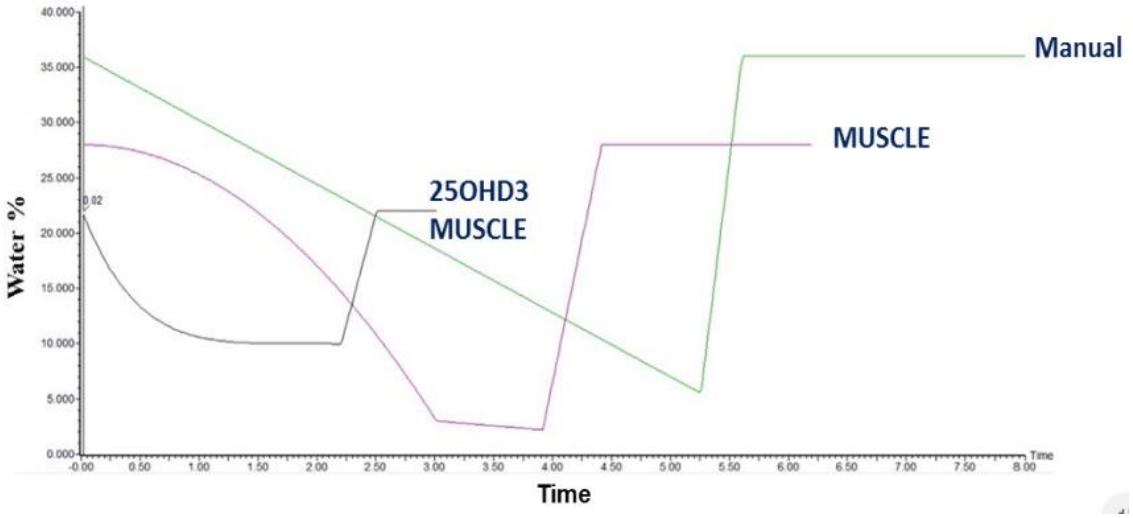


Figure 6.12: Overlaid LC gradients showing the manually optimised method (green line), the MUSCLE optimised method (purple line) and the additional 24OHD3 quantification method (black line).

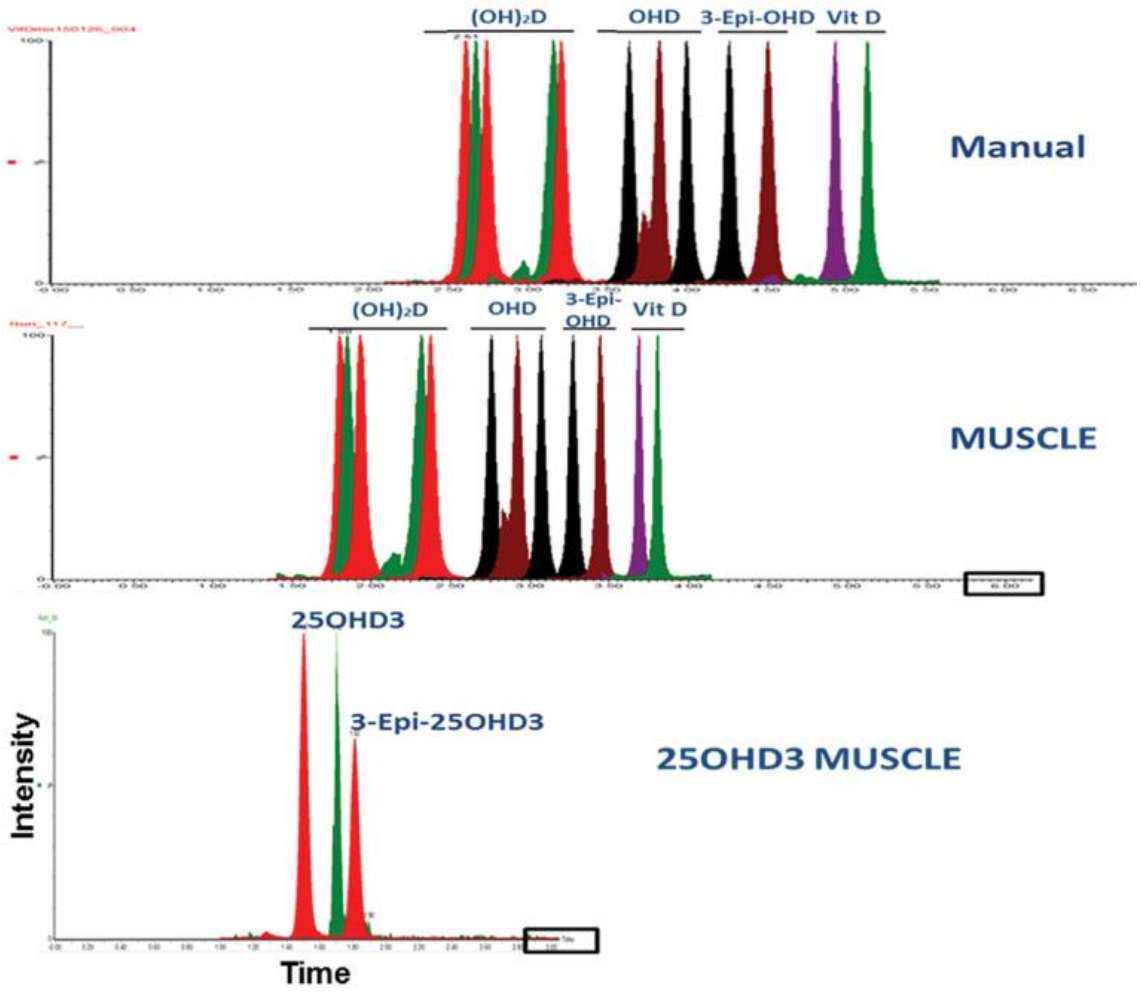


Figure 6.13: Chromatograms for the manually optimised method, the MUSCLE optimised method and the additional 25OHD3 quantification method.

For more detailed information about the optimisation and instrument parameters, and method validation see Appendix D.

6.3.2. Semi-targeted

Chapter 5 presents a set of computationally generated hypotheses that predict how the *D. magna* metabolome is effected by two environmental perturbations using a draft GWMR built using the METRONOME pipeline (Chapter 4). These predictions are previously unknown, so a metabolomics study is designed to validate these predictions (Chapter 7),

in which LC-MS is used as the analytical platform. The aim of this optimisation is develop a non-targeted method for use in this study.

The predictions from chapter 5 include a number of compounds that are predicted to be effected. The LC-MS method used for the study would ideally be able to separate as many of these compounds as possible. In order to achieve this, a list possible m/z values are calculated based on the predicted compounds and their possible adducts. The method is optimised to detect as many of these m/z values as possible. This kind of optimisation is termed semi-targeted, as it is looking for a targeted set of peaks within an untargeted analysis.

System Information

Separations are performed using a Thermo Scientific Ultimate 3000 ultra-high performance liquid chromatograph with a Thermo Hypersil Gold aQ column (130Å, 1.9µm, 1 mm X 100 mm, 3µm guard cartridge). Mobile phase A is comprised of 0.1% formic acid in water. Mobile phase B is comprised of 0.1% formic acid in methanol. All solvents used are Fisher Optima LC-MS grade. The flow rate is set to 0.08 mL/min and the column temperature is set at 40 degrees Celsius.

Mass spectral detection is performed with a Thermo Scientific Q Exactive tuned to 70,000 mass resolution. Data is collected in profile mode in positive ESI mode with a mass range of 100-1,000 Daltons with a scan time of 0.2 seconds.

Each injection consists of 2 µl sample and each run lasts 28 minutes.

Instrument Parameters

The instrument parameters that are optimised are shown in Table 6.4. The total number of combinations of these parameters is 1.167×10^9.

Table 6.4: Optimisation parameters and the minimum, maximum and step sizes used for the closed-loop optimisation.

Parameter	Type	Min Value	Max Value	Step Size
t_1 (start of second step)	LC	6.0	10.5	0.5
t_2 (end of second step)	LC	11.5	15.5	0.5
$\%_1$ (initial conditions)	LC	40	80	2
c_1 (1^{st} step curve)	LC	1	9	1
c_2 (2^{nd} step curve)	LC	1	9	1
Spray voltage	MS	3.0	4.0	0.1
Auxiliary gas flow rate	MS	10	20	1
Sheath gas flow rate	MS	24	40	2
S-lens RF	MS	40	100	10

Figure 6.14 is a visualisation of the LC gradient optimisation parameters.

The method is required have a maximum time of 30 minutes. This particular method also requires an equilibration time of 10 minutes. This means that the gradient must return to the starting conditions at the 20 minute mark, a hold time of 4.5 minutes at 100% B is also included meaning that the gradient ramp must be complete at 15.5 minutes. A two-step gradient is chosen, with the start point of the second step being optimised by the parameters t_1 and $\%_1$. t_1 defines the time that the second step starts, and $\%_1$ defines the mobile phase composition at the start of the second step. The time at which the gradient reaches 100% B and starts the hold is optimised with the t_2 variable. If an optimised

method uses a value for t_2 that is less than 15.5, the method can be shortened. The curve of each step is optimised using the c_1 and c_2 parameters.

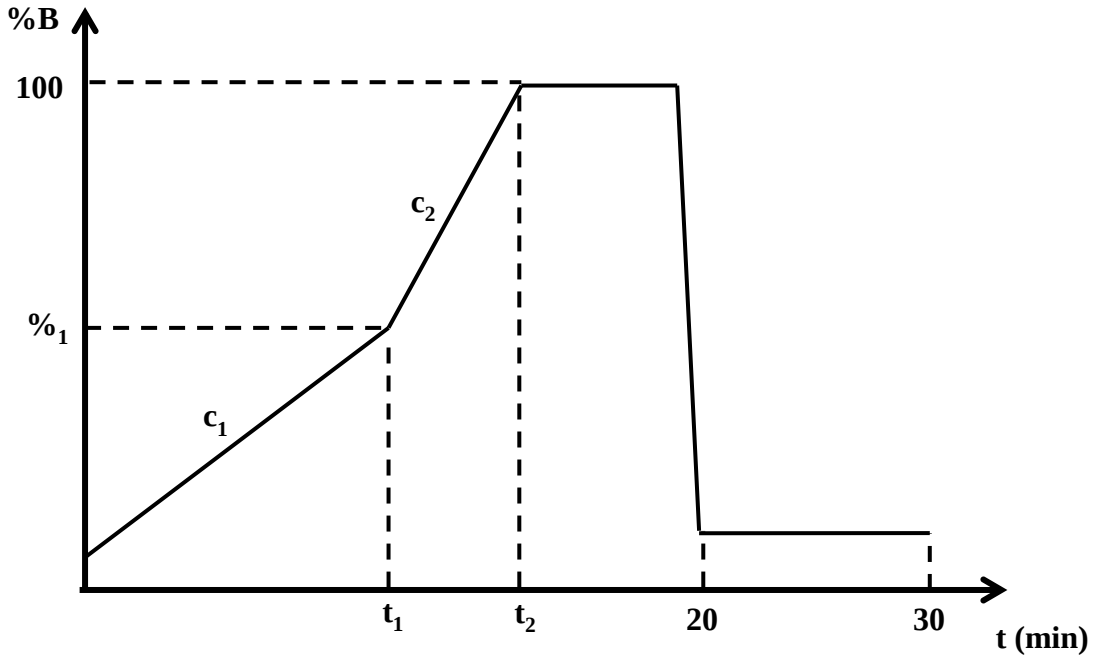


Figure 6.14: Visualisation of the LC gradient optimisation. The parameters t_1, t_2, $\%_1$, c_1 and c_2 are optimised during the optimisation. Table 6.4 shows the possible values that can be entered for the parameters.

Algorithm Configuration

The standard closed-loop PESA-II algorithm described in section 6.2.4.1 is used. The maximum number of injections is set to 250, with 50 initial LHS injections. The crossover rate is set to 0.7 and the mutation rate is set to 0.2. The objective measures used are; maximise the number of features, maximise the separation and maximise the number of m/z values in the XCMS output from a predetermined *hit-list*.

The number of features is maximised as a good method for this study would include as many features as possible. Only peaks with a signal to noise value of above the XCMS signal to noise threshold of 60 will be present in the XCMS output. This means that only peaks with good intensity values will be counted.

Chromatographic separation is important as a poorly separated method can suffer from ion suppression, where co-eluting compounds interact to supress or enhance or MS signal response resulting in inaccurate measurement (Furey et al, 2013). To calculate separation, the following procedure is used:

1. Cluster all detected peaks based on their retention times using the DBScan clustering algorithm (Ester et al, 1996).
2. Calculate a distance matrix for all peaks within each cluster.
3. Calculate the average distance for each clusters distance matrix.
4. Separation objective is calculated by the sum of the average distances divided by the total number of clusters.

The intuition behind using this method for measuring separation is that in an untargeted analysis, peaks with similar chemical structures will elute from the LC column at similar times. Therefore, clusters of chemically similar compounds will naturally occur and by basing the value on separation on these naturally occurring clusters, methods which have greater intra-cluster separation will be favoured. DBScan is used as the clustering algorithm as it does not require the number of clusters to be pre-defined.

The third objective is the semi-targeted objective measure. As this method is to be used for in a study that is looking for a pre-defined list of compounds in an untargeted spectrum, an ideal method would be able to detect as many of these peaks as possible. Performing accurate peak annotation during an optimisation is not feasible, so instead the semi-targeted objective measure calculates the percentage of m/z values are detected in the XCMS output based on a predefined list. For this study, there are 178 predicted

compounds. This resulted in 252 possible m/z values that are inside the 100-1,000 Dalton mass range used for the MS.

Data Processing

As this is an untargeted analysis, XCMS (Smith et al, 2006; Tautenhahn et al, 2008) is used for feature detection. Table 6.5 contains the XCMS parameters used during the optimisation. These parameters are taken from a database of instrument specific recommended parameters at the XCMS online web site (Institute, 2017). The signal to noise threshold parameter (snthresh) is set to a higher value than is recommend (60 instead of 10) to decrease the time taken for each run to be analysed during the optimisation.

Table 6.5: XCMS parameters used during optimisation.

Parameter	Value
method	centWave
ppm	5
peakwidth	5, 20
snthresh	60
prefilter	3, 100
mzdiff	0.01

Optimisation Results

Table 6.6 shows the final archive set of the optimisation. Figure 6.15 visualises the archive, with an axis for each objective measure. All of the corresponding chromatograms from the archive set are inspected manually and assessed by an analyst and the three most preferred chromatograms are selected for further inspection. These runs are highlighted in Table 6.6 and Figure 6.15, with the chromatograms shown in B-D sub figures.

Table 6.6: Final archive for the optimisation. Each row in the table is a non-dominated solution contained in the final archive set. Three manually selected methods are highlighted in red.

Run #	# Peaks Objective	Separation Objective	% Target Peaks
113	17435	2409.2	16
114	13637	1135.9	17
121	18947	4870.4	14
139	15228	1374.7	16
148	19630	17352.8	18
151	19506	9082.1	19
153	17552	2764.1	14
154	18742	2954.0	19
169	16071	1772.2	18
170	13652	718.9	15
173	15917	1459.3	21
183 (Figure 6.15 B)	**17732**	**5574.7**	**26**
185	17809	4658.2	23
189	18210	4878.1	21
194	16839	4501.2	20
197	16474	1984.7	20
199	19279	17207.3	24
206	16554	3615.8	22
207 (Figure 6.15 C)	**19118**	**21748.5**	**25**
215	15113	1237.6	21
219	15550	2861.6	22
227	19544	10985.0	22
237 (Figure 6.15 D)	**19105**	**5599.6**	**25**
247	14714	1970.0	22

After manual inspection, run 237 is chosen as the preferred method as it has a good value for the semi-targeted objective, and a high number of detected features. Although run 183 has a better value for separation (albeit with fewer detected features), upon inspection, it is deemed that the overall features across the entire run is more preferable. The last three minutes of the run in particular has many more detected features than the other selected runs.

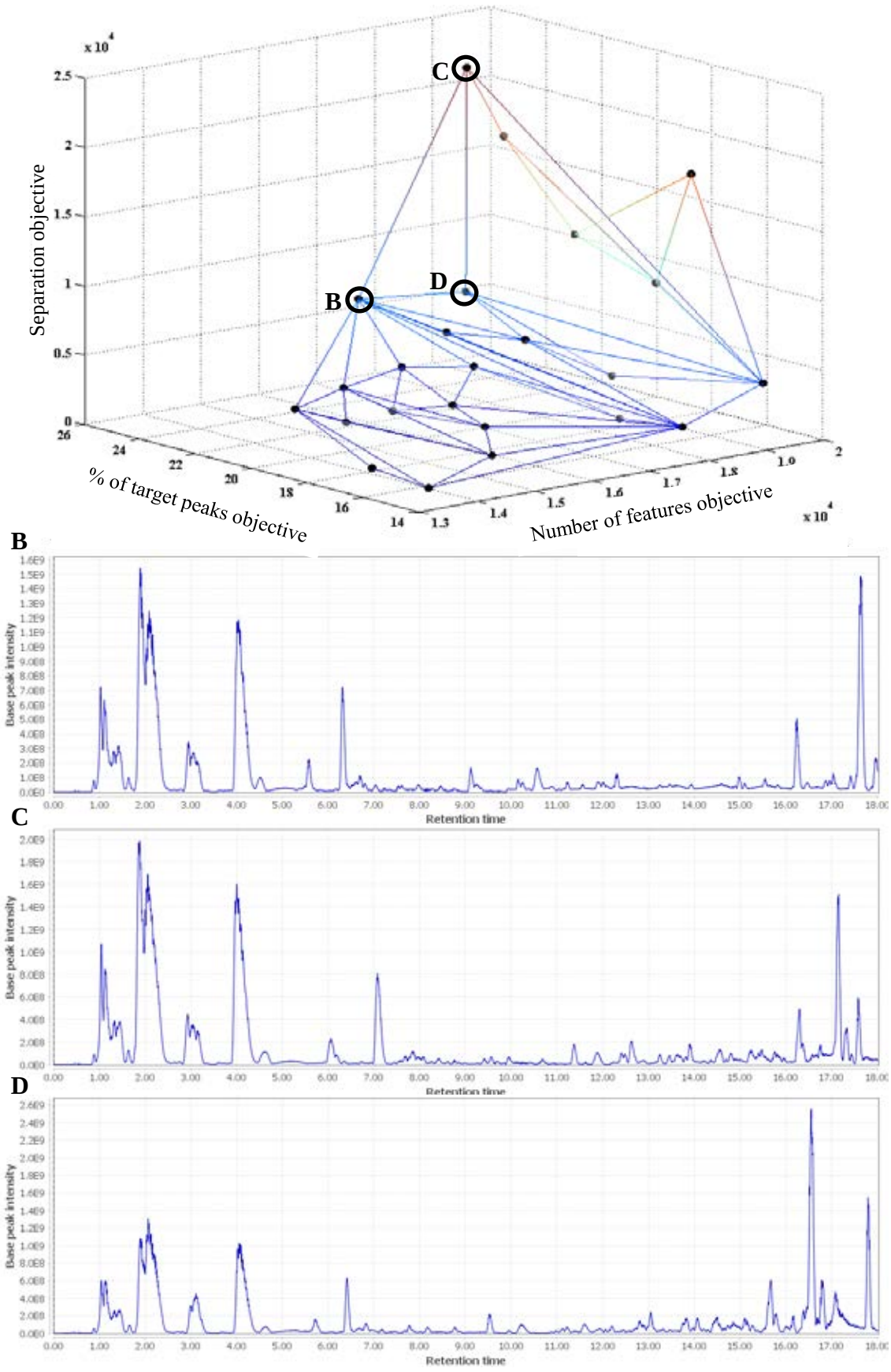


Figure 6.15: Optimisation Results PESA-II. The lines in A are for visualisation purposes only.

Method Validation

The preferred method is run ten times to validate the method and to check its stability. For the validation, the XCMS processing is carried out with the same parameter values as during the closed-loop optimisation (Table 6.5Table 6.12) except for the signal to noise threshold (snthresh) which is set to 10 (the recommended value). Table 6.7 shows the mean, standard deviation and relative standard deviation for the three objectives, for the preferred method across the ten replicates. To access the performance of the optimisation the optimised method is compared with a previously developed manually optimised method, with the same statistics calculated across 10 replicates.

Table 6.7: Validation run statistics. The mean, standard deviation and relative standard deviation of each of the objective measures is calculated for the selected optimised method as well as for a previously developed manually optimised method.

Run		# Peaks	Separation	% Target Peaks
Manual	Mean	31149.05	14752.25	52.15
	Standard Deviation	1394.21	5031.78	4.59
	Relative Standard Deviation	4.48	34.11	8.80
237	Mean	44867.76	85847.24	64.12
	Standard Deviation	1830.65	25877.06	3.81
	Relative Standard Deviation	4.08	30.14	5.94

Table 6.8 shows the improvements achieved in each of the three objectives from the manually optimised method.

Table 6.8: The percentage improvements in the objective measures for the selected optimised method compared to the previously developed manually optimised method. The values are calculated based on the mean values for the objective measurements across 10 replicates.

Objective	Run 237
# Peaks	44.04%
Separation	481.93%
Semi-Targetd	22.95%

Method Parameters

Table 6.9 shows the values for the instrument parameters used in the final optimised method.

Table 6.9: Parameter values for the selected optimised method.

Parameter	Type	Run 237
t_1 (start of second step)	LC	6.5
t_2 (end of second step)	LC	15
$\%_1$ (initial conditions)	LC	58
c_1 (1^{st} step curve)	LC	2
c_2 (2^{nd} step curve)	LC	1
Spray voltage	MS	3.1
Auxiliary gas flow rate	MS	13
Sheath gas flow rate	MS	36
S-lens RF	MS	90

6.3.3. Untargeted

The aim of this optimisation is to develop an untargeted UPLC-MS assay for HILIC negative ion analysis of urine. The method must have an injection-to-injection time of approximately 15 minutes and must be robust enough to allow large-scale studies to be performed.

The optimisation is carried out twice, once using the standard PESA-II algorithm described in section 6.2.4.1, and once with the PESA-II with feature selection algorithm described in section 6.2.4.2.

System Information

Separations are performed using a Waters Acquity liquid chromatograph with a Waters ACQUITY UPLC BEH Amide Column (130Å, 1.7 µm, 2.1 mm X 150 mm). Mobile phase A is comprised of 5 mMol ammonium acetate in 95% acetonitrile and 5 % water. Mobile phase B is comprised of 5 mMol ammonium acetate in 50% acetonitrile and 50%

water. All solvents used are Fisher Optima LC-MS grade. The flow rate is set to 0.6 mL/min and the column temperature is set at 40 degrees celcius.

Mass spectral detection is performed using a Waters Xevo G2 tuned to 35000 mass resolution using leucine enkephalin (554 peak) resolution. Data is collected in centroid mode in negative ESI mode with a mass range of 50-700 Daltons with a scan time of 0.1 seconds.

Each injection consists of 2 µl sample and each run lasts 15 minutes.

Instrument Parameters

The instrument parameters that are optimised are shown in Table 6.10. The total number of combinations of these parameters is 6.04×10^8. The same parameters are used for both optimisations.

Table 6.10: Optimisation parameters and the minimum, maximum and step sizes used for the PESA-II and PESA-II-FS closed-loop optimisations.

Parameter	Type	Min Value	Max Value	Step Size
t_1 (start of second step)	LC	2.5	5.0	0.25
$\%_1$ (initial conditions)	LC	1	5	1
$\%_2$ (step condition)	LC	10	30	2
$\%_3$ (final condition)	LC	40	66	2
c_1 (1^{st} step curve)	LC	2	10	1
c_2 (2^{nd} step curve)	LC	2	10	1
Spray voltage	MS	1.5	3.0	0.1
Desolvation Gas Flow	MS	600	1100	50
Desolvation Temperature	MS	300	500	50

Figure 6.16 is a visualisation of the LC gradient optimisation parameters. The method is required to be 15 minutes in total as this is a standard length for an untargeted method as a 15 minute method allows for a 96 well plate to be processed in 24 hours. This particular method also requires an equilibration time of 7.5 minutes. This means that the gradient

must return to the starting conditions at the 7.5 minute mark, a hold time of 0.5 minutes is also included meaning that the gradient ramp must be complete at 7 minutes. A two-step gradient is chosen, with the start point of the second step being optimised by the parameters t_1 and $\%_2$. t_1 defines the time that the second step starts, whereas $\%_2$ defines the mobile phase composition at the start of the second step. The initial and final mobile phase composition is also optimised using the $\%_1$ and $\%_2$ parameters respectively. The curve of each step is optimised using the c_1 and c_2 parameters.

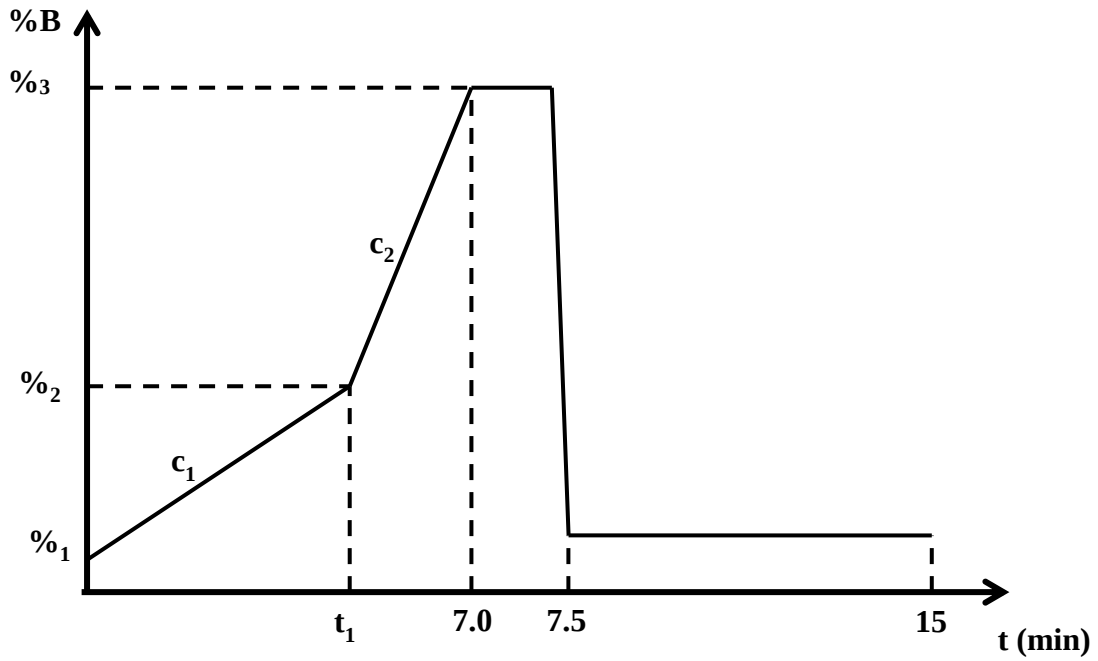


Figure 6.16: Visualisation of the LC gradient optimisation. The parameters t_1, $\%_1$, $\%_2$, $\%_3$, c_1 and c_2 are optimised during the optimisation. Table 6.10 shows the possible values that can be entered for the parameters.

Algorithm Configuration

For both optimisations, the same algorithm configuration is used. The maximum number of injections is set to 250, with 50 initial LHS injections. The crossover rate is set to 0.7 and the mutation rate is set to 0.2. The objective measures used are; maximise the number

of features, maximise the separation of features and maximise the mean signal to noise threshold of the detected features.

The number of features is maximised as a good method for this study would include as many features as possible. Only peaks with a signal to noise value of above the XCMS signal to noise threshold of 60 will be present in the XCMS output. This means that only peaks with good intensity values will be counted.

Chromatographic separation is important as a poorly separated method can suffer from ion suppression, where co-eluting compounds interact to supress or enhance or MS signal response resulting in inaccurate measurement (Furey et al, 2013). To calculate separation, first a distance matrix is constructed for all detected features, with the values in the distance matrix corresponding to retention time differences between features. The separation value is then set as the average value in the distance matrix. Chromatograms with good separation are expected to have a greater average distance between features. This objective conflicts with the total number of detected features, if more features are present in the chromatogram, there is less available space in the retention time axis for the features to inhabit.

The average signal to noise is optimised as it is a sound metric for feature quality. This objective is also in conflict with the total number of features. If a method results in more features being detected, there is a high chance that the extra features have signal to noise thresholds that are just above the signal to noise threshold, therefore dragging the average down.

PESA-II with Feature Selection configuration

The only additional parameter that needs to be set for the PESA-II with Feature Selection algorithm is round size. Round size determines how often the feature selection and subsequent variable partitioning (see Figure 6.9) is performed, with the first round beginning after the initial LHS phase. For this optimisation, a round size of 20 is selected. Table 6.11 shows the run numbers associated with each round for this optimisation.

Table 6.11: PESA-II with Feature Selection rounds.

Run Number	Round
1 – 50	LHS Sampling
51 – 70	Round 1
71 – 90	Round 2
91 – 110	Round 3
111 – 130	Round 4
131 – 150	Round 5
151 – 170	Round 6
171 – 190	Round 7
191 – 210	Round 8
211 – 230	Round 9
231 - 250	Round 10

Data Processing

As this is an untargeted analysis, XCMS (Smith et al, 2006; Tautenhahn et al, 2008) is used for feature detection. Table 6.12 contains the XCMS parameters used during the optimisation. These parameters are taken from a database of instrument specific recommended parameters at the XCMS online web site (Institute, 2017). The signal to noise threshold parameter (snthresh) is set to a higher value than is recommend (60 instead of 10) to decrease the time taken for each run to be analysed during the optimisation.

Table 6.12: XCMS parameters used during optimisation.

Parameter	Value
method	centWave
ppm	15
peakwidth	2, 25
snthresh	60
prefilter	3, 100
mzCenterFun	wMean
mzdiff	0.01

Optimisation Results – Standard PESA-II

Table 6.13 shows the final archive set of the PESA-II optimisation. Figure 2A visualises the archive, with an axis for each objective measure. All of the corresponding chromatograms from the archive set are inspected manually and assessed by an analyst and the three most preferred chromatograms are selected for further inspection. These runs are highlighted in Table 6.13 and Figure 6.17, with the chromatograms shown in B-D sub figures.

Table 6.13: Final archive for the PESA-II optimisation. Each row in the table is a non-dominated solution contained in the final archive set. Three manually selected methods are highlighted in red.

Run #	# Peaks Objective	Separation Objective	S/N Objective
142	1053	7.89E-04	784.3
155	1054	7.92E-04	598.2
165	1325	4.89E-04	584.9
169	924	0.001	523.2
173 (Figure 6.17 B)	**1224**	**6.05E-04**	**671.3**
176	1296	5.25E-04	650.9
194 (Figure 6.17 C)	**1022**	**8.25E-04**	**657.7**
195	909	0.00101	646.6
202	846	0.00108	620.3
206	882	0.00122	600.9
211	792	0.00128	456.2
212	820	0.00124	685.8
213 (Figure 6.17 D)	**886**	**0.00122**	**525.9**
223	444	0.00341	609.5
225	522	0.00306	633.7
231	724	0.00173	705.2
232	538	0.00286	1127.3
233	674	0.00173	629.3

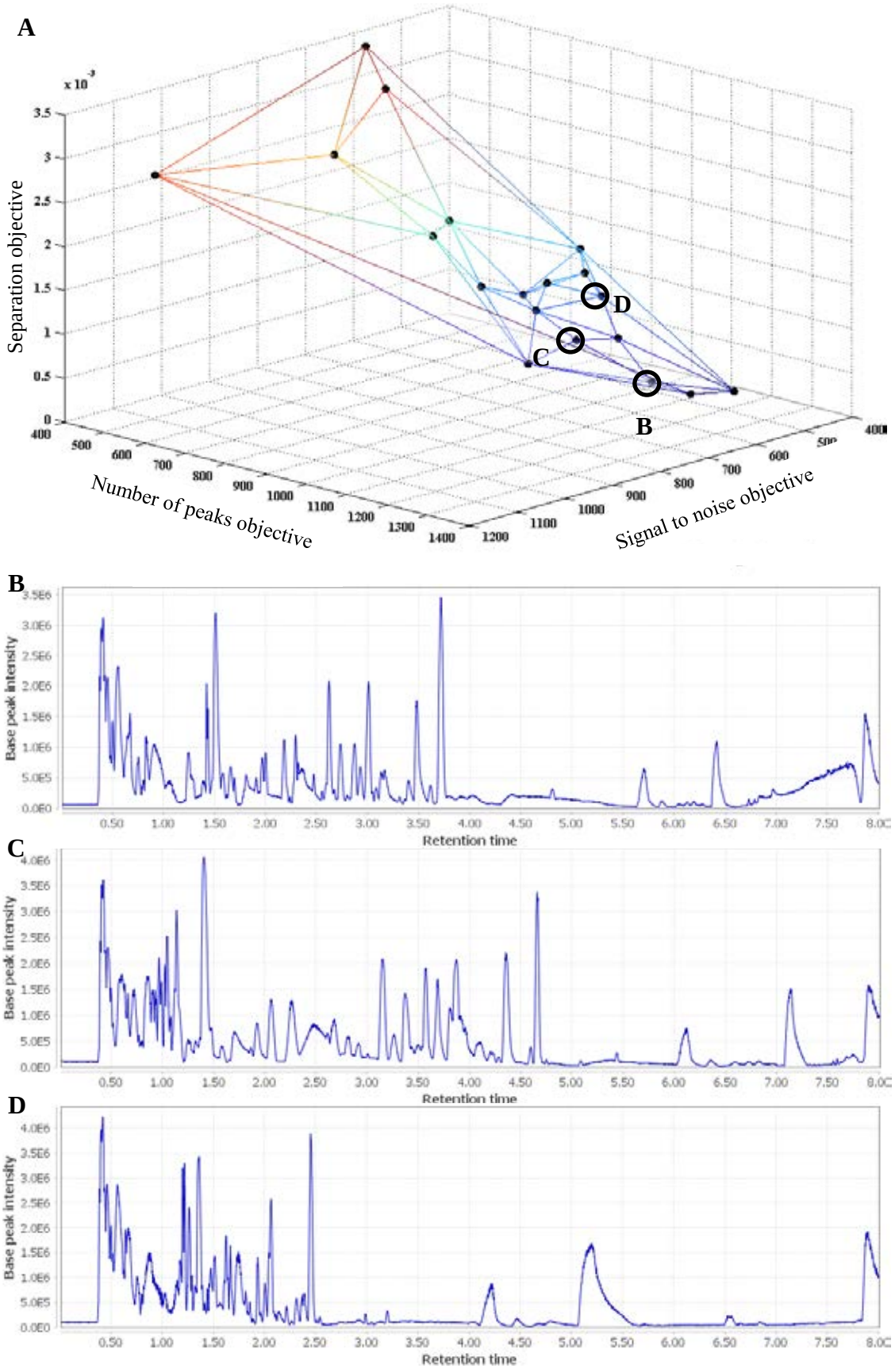


Figure 6.17: Optimisation Results PESA-II. The lines in A are for visualisation purposes only.

Method Validation – Standard PESA-II

Each of the three preferred runs is run ten times to validate the method and to check its stability. For the validation, the XCMS processing is carried out with the same parameter values as during the closed-loop optimisation (Table 6.12). Table 6.14 shows the mean, standard deviation and relative standard deviation for the three objectives, for each of the preferred runs across the ten replicates. To access the performance of the optimisation the optimised method is compared with a previously developed manually optimised method, with the same statistics calculated across 10 replicates.

Table 6.14: Validation run statistics. The mean, standard deviation and relative standard deviation of each of the objective measures is calculated for the three selected optimised methods as well as for a previously developed manually optimised method.

Run		# Peaks	Separation	S/N
Manual	Mean	1903.7	0.00027	1863.41
	Standard Deviation	81.36366	2.46E-05	326.0159
	Relative Standard Deviation	4.2742	8.99901	17.4957
173	Mean	1945.6	0.00029	1963.52
	Standard Deviation	44.50518	1.22E-05	614.3667
	Relative Standard Deviation	2.28748	4.26709	31.289
194	Mean	2004.78	0.00027	2883.86
	Standard Deviation	61.21637	1.84E-05	1604.251
	Relative Standard Deviation	3.05352	6.85055	55.6285
213	Mean	2016.71	0.00029	3860.03
	Standard Deviation	61.83234	2.84E-05	2048.498
	Relative Standard Deviation	3.06599	9.68148	53.0694

Table 6.15 shows the improvements achieved in each of the three objectives for the three preferred methods optimised using the modified PESA-II algorithm when compared to the manually optimised method.

Table 6.15: The percentage improvements in the objective measures for the three selected optimised methods compared to the previously developed manually optimised method. The values are calculated based on the mean values for the objective measurements across 10 replicates.

Objective	Run 173	Run 194	Run 213
# Peaks	2.21%	5.32%	5.94%
Separation	7.41%	0%	7.41%
S/N	5.37%	54.76%	107.15%

Method Parameters - Standard PESA-II

Table 6.16 shows the values for the instrument parameters used in the three preferred optimised methods.

Table 6.16: Parameter values for the three selected optimised methods.

Parameter	Type	Run 173	Run 194	Run 213
t_1 (start of second step)	LC	4.25	4.0	3.0
$\%_1$ (initial conditions)	LC	2	5	4
$\%_2$ (step condition)	LC	18	22	22
$\%_3$ (final condition)	LC	54	58	44
c_1 (1^{st} step curve)	LC	4	7	3
c_2 (2^{nd} step curve)	LC	6	9	8
Spray voltage	MS	2.3	2.4	3.0
Desolvation Gas Flow	MS	1050	750	600
Desolvation Temperature	MS	450	500	500

Optimisation Results - PESA-II with Feature Selection

Table 6.17 shows the final archive set of the PESA-II-FS optimisation. Figure 3A visualises the archive, with an axis for each objective measure. All of the corresponding chromatograms from the archive set are inspected manually and assessed by an analyst and the three most preferred chromatograms are selected for further inspection. These runs are highlighted in Table 6.17 and Figure 6.18 with the chromatograms shown in B-D sub-plots.

Table 6.17: Final archive for the PESA-II-FS optimisation. Each row in the table is a non-dominated solution contained in the final archive set. Three manually selected methods are highlighted in red.

Run #	# Peaks Objective	Separation Objective	S/N Objective
159	749	0.00156	714.6
169 (Figure 6.18 B)	**906**	**0.001063**	**464.4**
174	979	0.000983	745.8
175 (Figure 6.18 C)	**880**	**0.001131**	**638.7**
181	651	0.001862	564.5
188	898	0.001092	585.5
189	703	0.001741	554.6
208	813	0.001265	622.2
210	678	0.001759	807.8
220	783	0.001377	778
226	871	0.001155	538.1
233	684	0.001772	593.7
236	804	0.001305	787.5
239 (Figure 6.18 D)	**728**	**0.001733**	**662.1**
245	743	0.001433	917.8
249	613	0.001977	1003.9

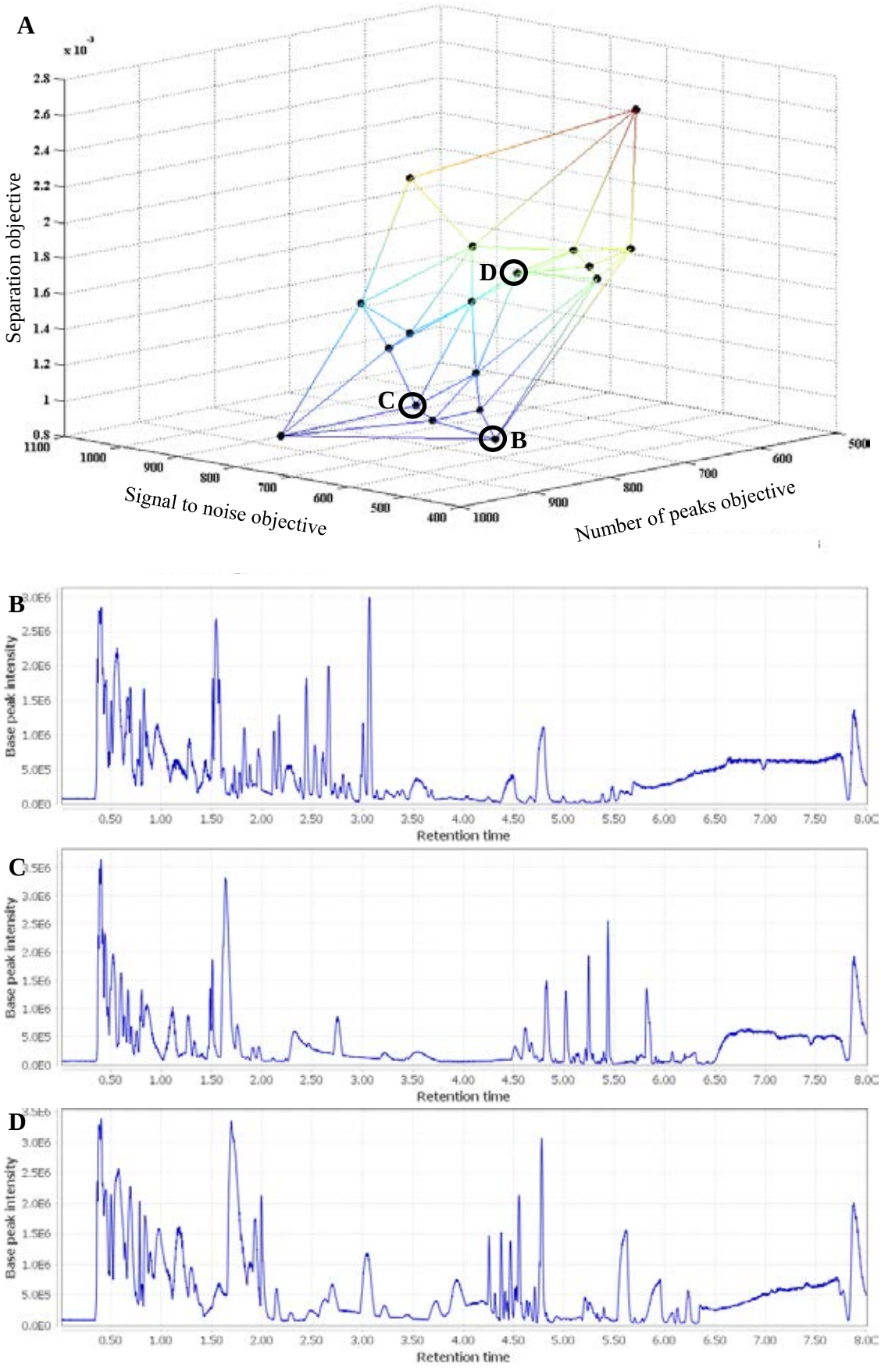


Figure 6.18: Optimisation Results PESA-II-FS. The lines in A are for visualisation purposes only.

Table 6.18 shows which LC-MS parameters are deemed to be the most influential to the overall fitness of the solutions for each round of the closed-loop optimisation (see Table 6.11). A blue cell means that a LC-MS parameter is considered to have a significant influence for a given round, whereas a red cell means that the given parameter is not significant. As the round number increases, there are more data points (evaluated LC-MS runs) to be used for feature selection. Therefore, the accuracy of the feature selection should increase after each round.

Table 6.18: Visualisation of feature selected parameters for each optimisation round. Blue cells show which LC-MS parameters are selected for optimisation with PESA-II. Red cells show the LC-MS parameters that are not selected and whose values are selected based on a Latin Hypercube constructed from parameter values taken from the archive set (see section 6.2.4.2).

Parameter	**Round Number**									
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
t_1 (start of second step)										
$\%_1$ (initial conditions)										
$\%_2$ (step condition)										
$\%_3$ (final condition)										
c_1 (1^{st} step curve)										
c_2 (2^{nd} step curve)										
Spray voltage										
Desolvation Gas Flow										
Desolvation Temperature										

Table 6.18 shows that throughout the optimisation, the LC parameters t_1, $\%_1$ and $\%_2$ (see Figure 6.16) had a significant influence on the overall fitness of solutions. These parameters play a big role in defining the characteristics of the LC gradient and are expected to have a large influence on the total number of peaks observed as well as their chromatographic separation (Pitt, 2009). The $\%_3$ parameter was deemed to have a significant effect in six of the 10 rounds, the gradient curve parameter c_1 is significant for all but the third and fourth rounds, while the c_2 parameter was deemed less significant as more data points for MLR became available.

The impact of the LC parameters that control the latter part of the LC gradient, $\%_3$ and c_2 are the LC parameters that are deemed to be less influential most often. Their impact potentially varies depending on the values that have been selected for the variables controlling the first part of the LC gradient. This can perhaps explain why their influence is not consistently deemed significant.

The influence of the MS parameters desolvation gas flow and desolvation temperature were consistent throughout the optimisation. The gas flow was deemed insignificant throughout the optimisation, and the temperature was significant in rounds two to ten. The insignificance of the gas flow parameter could be explained because the possible values selected are with in fairly narrow window (see Table 6.10), and also that the impact of the value of this parameter on a run is coupled to the value used for the flow rate (Banerjee & Mazumdar, 2012), which is kept constant throughout the optimisation. The spray voltage MS parameter was the most variable parameter, deemed significant in five of ten rounds.

Method Validation - PESA-II with Feature Selection

Each of the three preferred runs is run ten times to validate the method and to check its stability. For the validation, the XCMS processing is carried out with the same parameter values as during the closed-loop optimisation (Table 6.12). Table 6.19 shows the mean, standard deviation and relative standard deviation for the three objectives, for each of the preferred runs across the ten replicates. To access the performance of the optimisation the optimised method is compared with a previously developed manually optimised method, with the same statistics calculated across 10 replicates.

Table 6.19: Validation run statistics. The mean, standard deviation and relative standard deviation of each of the objective measures is calculated for the three selected optimised methods as well as for a previously developed manually optimised method.

Run		# Peaks	Separation	S/N
Manual	Mean	1903.7	0.00027	1863.41
	Standard Deviation	81.36366	2.46E-05	326.0159
	Relative Standard Deviation	4.2742	8.99901	17.4957
169	Mean	2041.75	0.00027	1705.95
	Standard Deviation	84.55028	2.63E-05	369.0105
	Relative Standard Deviation	4.14107	9.91819	21.6308
175	Mean	2134.2	0.00024	2938.19
	Standard Deviation	89.20986	2.43E-05	921.1826
	Relative Standard Deviation	4.18001	10.1599	31.352
239	Mean	2305.8	0.0003	3321.04
	Standard Deviation	51.80905	4.43E-05	1080.9
	Relative Standard Deviation	2.2469	14.7597	32.547

Table 6.20 shows the improvements achieved in each of the three objectives for the three preferred methods optimised using the modified PESA-II-FS algorithm when compared to the manually optimised method.

Table 6.20: The percentage improvements in the objective measures for the three selected optimised methods compared to the previously developed manually optimised method. The values are calculated based on the mean values for the objective measurements across 10 replicates.

Objective	Run 169	Run 175	Run 239
# Peaks	7.26%	12.11%	21.13%
Separation	0%	-11.11%	11.11%
S/N	-8.45%	57.68%	62.82%

Method Parameters - PESA-II with Feature Selection

Table 6.21 shows the values for the instrument parameters used in the three preferred optimised methods.

Table 6.21: Parameter values for the three selected optimised methods.

Parameter	**Type**	**Run 169**	**Run 175**	**Run 239**
t_1 (start of second step)	LC	3.5	4.25	3.5
$\%_1$ (initial conditions)	LC	1	2	1
$\%_2$ (step condition)	LC	26	24	26
$\%_3$ (final condition)	LC	42	46	44
c_1 (1^{st} step curve)	LC	5	10	10
c_2 (2^{nd} step curve)	LC	5	3	6
Spray voltage	MS	1.5	2.5	2.3
Desolvation Gas Flow	MS	600	700	950
Desolvation Temperature	MS	500	500	500

6.4. Discussion & Conclusion

The flexible modular MUSCLE software platform for automated closed-loop optimisation of LC-MS method development is presented. MUSCLE can be used to optimise targeted and untargeted analyses across a number of LC-MS analytical platforms. Visual scripting is used to change instrument parameter values (section 6.2.1), which allows MUSCLE to be used on any analytical platform with no modification to the source code. Users simply generate visual scripts which are then used as part of a wider user defined optimisation architecture (Figure 6.3).

The modular design allows for data processing and objective measures to be adjusted for each optimisation. Currently, the data processing module can be configured to use the in-built targeted peak detection function (Section 6.2.2.1) or to use XCMS to detect peaks in untargeted spectra. The modular design also allows for different evolutionary algorithms to be used for the optimisation. A modified PESA-II algorithm (section 6.2.4.1) and the PESA-II-FS algorithm (section 6.2.4.2) are presented. The PESA-II-FS algorithm iteratively uses feature selection to build a predictive model that assesses which instrument parameters are having the greatest effect on the quality of the LC-MS output. These parameters then become the focus of the optimisation, increasing the exploitative nature of the search. Both algorithms are specifically designed for use in closed-loop optimisations, where typically the number of solution evaluations are severely limited due to the expenses and time needed to conduct each physical, real-world evaluation.

MUSCLE has been used for several applications (Table 6.3) and is here demonstrated through its use in studies to optimise targeted (section 6.3.1), untargeted (section 6.3.3) and semi-targeted (section 6.3.2) analyses. Each optimisation is performed on a different analytical platform and is optimised using different optimisation criteria. In all cases

MUSCLE optimised methods provided improvements from manually optimised methods (Table 6.8, Table 6.15 & Table 6.20).

The untargeted optimisation of a HILIC method using urine samples (section 6.3.3) compared the modified PESA-II and the PESA-II-FS algorithms. The objective measure improvements of the closed-loop optimised methods compared to a manually optimised method is shown in Table 6.22. The best PESA-II-FS algorithm (PESA-II-FS run 239) achieves a far higher grater of peaks, whilst at the same time achieving a better separation than any of the PESA-II optimised methods. It also achieves a decent average signal to noise ratio that is better than the second best PESA-II optimised value (PESA-II run 194), but not as good as the first (PESA-II run 213). An explanation for this could be that as higher number of peaks that are detected in PESA-II-FS run 239 are a result of MS instrument parameters that increase the sensitivity of the method. This will have the effect of increasing the number of peaks that just cross the signal to noise boundary defined in the XCMS parameters. The big increase in detected peaks will also influence the separation objective value, as can be seen in the PESA-II-FS run 175 objective measure, where an increase of peaks is observed, but with a lower value of separation.

Table 6.22: Improvements in objective functions for the three selected methods for optimisations using the PESA-II and PESA-II-FS algorithms. The percentage improvements are based on a comparison with the same manually optimised method.

Algorithm	**PESA-II**			**PESA-II-FS**		
Objective	Run 173	Run 194	Run 213	Run 169	Run 175	Run 239
# Peaks	2.21%	5.32%	5.94%	7.26%	12.11%	21.13%
Separation	7.41%	0%	7.41%	0%	-11.11%	11.11%
S/N	5.37%	54.76%	107.15%	-8.45%	57.68%	62.82%

The PESA-II-FS algorithm performed better than the original PESA-II MUSCLE algorithm, especially in terms of method sensitivity but further testing is required. The

PESA-II and the PESA-II-FS algorithms share the first two steps (Figure 6.8 & Figure 6.9), both have an archive set initialised with non-dominated solutions obtained using Latin Hypercube Sampling. The initialisation on the Latin Hypercube is stochastic in nature, and this case, each optimisation used a different Latin Hypercube. In future studies, the same Latin Hypercube can be used for both optimisations to give a fairer comparison between the two algorithms.

In conclusion, MUSCLE is used to develop LC-MS methods in a fully automated way using a closed-loop optimisation approach that increased analytical sensitivity and/or shortened the analysis times for a number of different analyses across a range of different analytical platforms and analysis types. This closed-loop approach has the potential to benefit many scientific fields that make use of LC-MS including metabolomics, proteomics and pharmacology.

MUSCLE is used to optimise an untargeted LC-MS method for measuring the Daphnia metabolome. The fully automated closed-loop optimisation is carried out in a semi-targeted way, with the aim of being able to measure as many of the metabolites that are predicted to be effected in the computational toxicology study detailed in this thesis. The optimised method can detect 64% of m/z values that are associated with the predicted metabolites identified using the active module identification approach detailed in Chapter 5. The optimised method will be used for the collection of all metabolomics data obtained in study outlined in the next chapter.

7. Validation of Computationally Generated Hypotheses Using Metabolomics Study

Previously, computationally generated hypotheses which predict the unknown effects on the *D. magna* metabolome when exposed to two environmentally relevant stressors are generated (Chapter 5) using a draft *D. magna* GWMR (Chapter 4) and two transcriptomics datasets taken from the STRESSFLEA project (Orsini et al, 2016). As the predicted metabolic effects are previously unknown, and this is an untested computational toxicology methodology for assessing the metabolic impact of environmental perturbations (Figure 2.1), a metabolomics toxicology study is conducted to assess the validity of the computationally generated predictions.

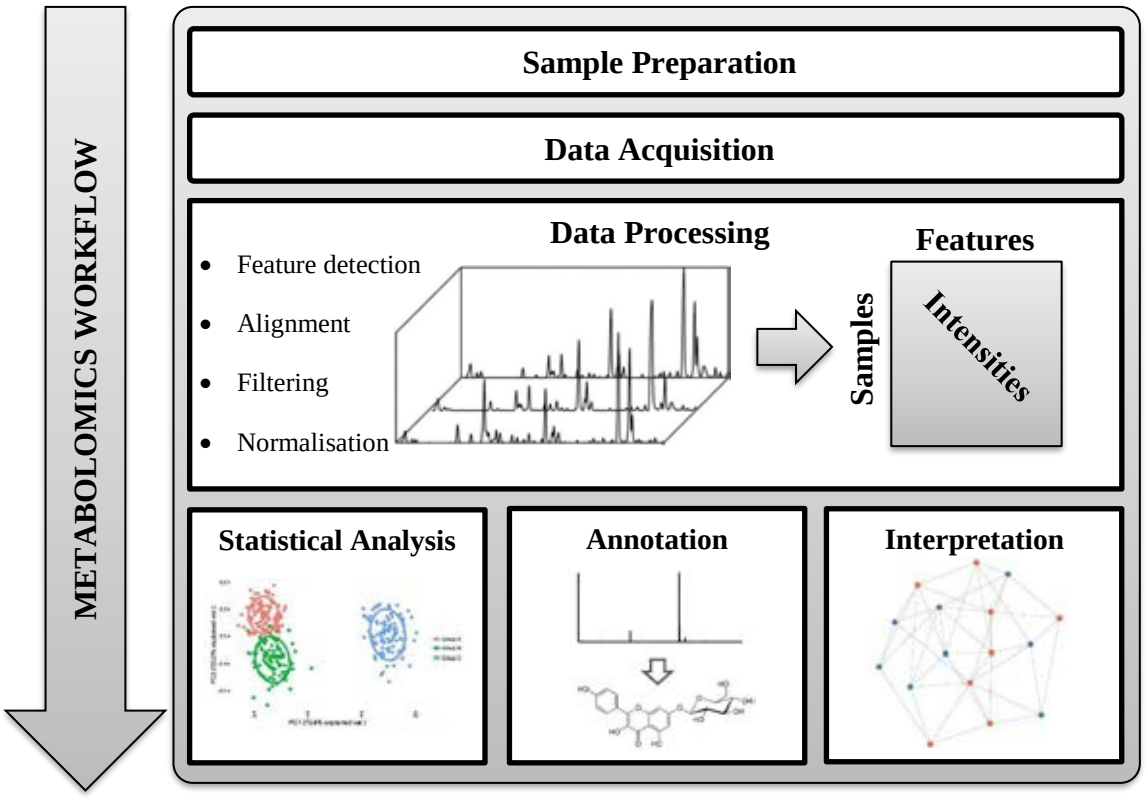


Figure 7.1: The typical metabolomics mass spectrometry based workflow. Adapted from (de Souza et al, 2017)

The procedure for doing this follows the typical metabolomics mass spectrometry based workflow (Figure 7.1). First samples are prepared based on an experimental design (section 7.1), which is followed by data acquisition (section 7.1.4). Acquired data is then processed to generate a data matrix (section 7.2) which is subsequently used for statistical analysis (section 7.3). This analysis identifies features of interest which are then annotated before being interpreted biologically (section 7.4).

7.1. Sample preparation

The computationally generated hypotheses in Chapter 5 are created using a draft *D. magna* GWMR (Chapter 4) and two transcriptomics datasets taken from the STRESSFLEA project (Orsini et al, 2016). These datasets measured the transcriptional responses of *D. magna* when exposed to environmentally relevant, sub-lethal, doses of the insecticide Carbaryl (8 µg/L) and Lead (278 µg/L). These doses are chosen as they reflect realistic human-induced pollution in inland waters. An experiment that exposes the same *D. magna* sub-species to the same dosages of Carbaryl and Lead as the transcriptomics study (Orsini et al, 2016) is designed. These exposures were observed to induce significant transcriptional response (Orsini et al, 2018).

7.1.1. Experimental design

D. magna has a parthenogenetic life cycle that allows the rearing of populations of genetically identical individuals (clones) from a single genotype. For this study a genotype collected from a system of ephemeral rock pools (Xinb3, South west Finland 59.833183, 23.260387) is used, which is the same genotype used in the transcriptomics experiments as well as the generation of a reference genome (NCBI accession number: LRGB00000000), which was used to generate the draft GWMR (Chapter 4).

The *D. magna* exposures experiments from the STRESSFLEA transcriptomics study are repeated for the purpose of making metabolomics measurements. These measurements will allow for the computationally generated hypotheses to be tested. Measurements are made at different time points: 4h, 8h, 12, and 24h, with first time point coinciding with the previous exposures to detect changes in transcriptional response to environmental perturbations. The reason for collecting samples at these time points is there is an expected and unknown temporal delay in metabolic response compared to transcriptional response. The aim of sampling through time at regular intervals is to ensure that metabolic changes following exposure to Carbaryl and Lead are captured.

For each time point a control (no stress imposed) is run in parallel to the environmental perturbations. Each treatment, including controls, is performed on eight biological replicates. Each biological replicate is made up of 15 *D. magna* individuals taken from individual aquaria where the exposures are performed (Figure 7.2). For each replicate metabolomics measurements are made using LC-MS (section 7.1.4).

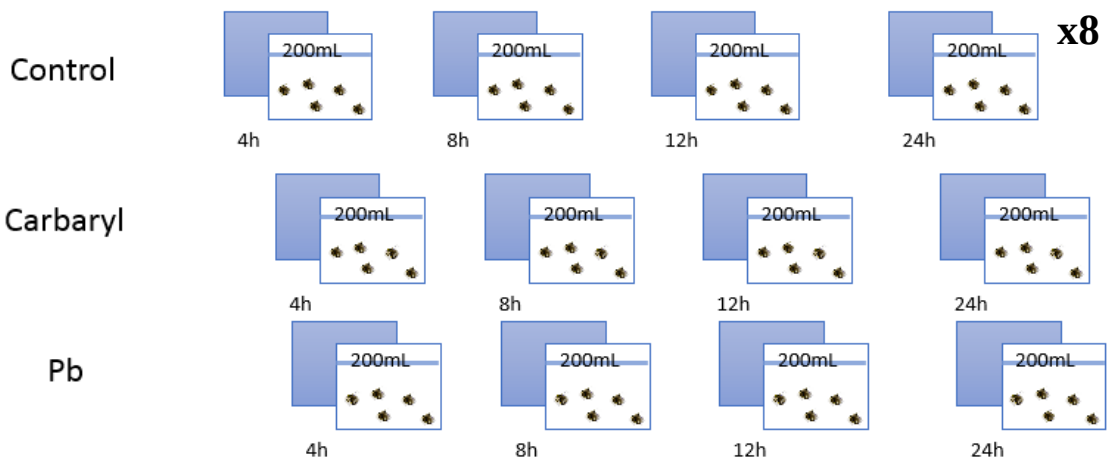


Figure 7.2: Biological replicates. For each replicate, a separate aquarium containing 15 individuals is used.

7.1.2. *D. magna* exposure details

Prior to the exposures, clonal populations of the genotype are synchronized in common garden conditions – controlled climate chambers with a fixed long day photoperiod (16h light/8h dark) at 20°C for at least two generations to reduce interference from maternal effect. Filtrated borehole water was used as growth medium and for exposure to the environmental perturbations. The animals were fed daily 0.8mg C/L of *Chlorella*.

The first generation is cultured at a density of 10 individuals/L, and increased to 50 individuals/L in large aquaria in the second generation to enable the harvesting of enough animals for the environmental perturbation exposures. The second clutch of the second generation are used for exposures to environmental perturbations. Batches of five-day old female juveniles randomly chosen from the offspring of the second generation of the synchronized animals at a density of 15 juveniles/L are exposed to the two environmental treatments for different lengths of time: 4h, 8h, 12, and 24h. The animal density for the exposures is determined from prior literature studies on *Daphnia* exposures (Jansen et al, 2011) and the previous study on transcriptional responses to these same stressors (Orsini et al, 2016).

All material is flash frozen in liquid nitrogen after collection and stored at -80°C in separate Precellys tubes per sample to quench metabolism prior to metabolite extraction.

7.1.3. Metabolite extraction

Each sample is taken up in 20µl of an equal mix of MeOH and water and then vortexed. Samples ae then spun at 15,000 rpm for 10 min at 4°C in a Biofuge rotor. 12µl is then pipetted into Thermo AB-0800 96-well plates in a controlled randomised order.

7.1.4. Data acquisition

All experimental measurement is performed using untargeted UHPLC-MS. Separations are performed using a Dionex Ultimate 3000 liquid chromatograph with a Thermo Hypersil Gold aQ, 100 x 1 mm column, with a 1.9 µm particle diatemer, and a a 10 x 1 mm, 3 µm guard cartridge.

Mass spectral detection is performed using a Q Exactive Orbitrap (Thermo Scientific) mass spectrometer equipped with a H-ESI II source, operated at 70,000 mass resolution (FWHM: 200m/z) in the positive ionisation mode. Data is collected in profile mode with a mass range of 100-1000 m/z. For each run, a 2µl injection is used.

The MUSCLE software (Chapter 6) package for automated closed loop LC-MS method optimisation (Bradbury et al, 2015) is used to develop the LC-MS method. In this instance, a semi-targeted optimisation approach is used. The LC-MS method optimisation is outlined in Section 6.3.2, which includes details the LC-MS parameters and the mobile and stationary phase compositions used.

7.2. Data processing

Several tools and workflows exist for processing and analysing untargeted LC-MS metabolomics data (Davidson et al, 2016; Giacomoni et al, 2015; Goecks et al, 2010; Institute, 2017; Smith et al, 2006; Tautenhahn et al, 2008; Xia et al, 2015). (Di Guida et al, 2016) present an investigation of the use of several approaches and provides a recommended workflow for analysing untargeted UHPLC-MS data sets. The workflow is applied here to process the experimental data (Figure 7.3).

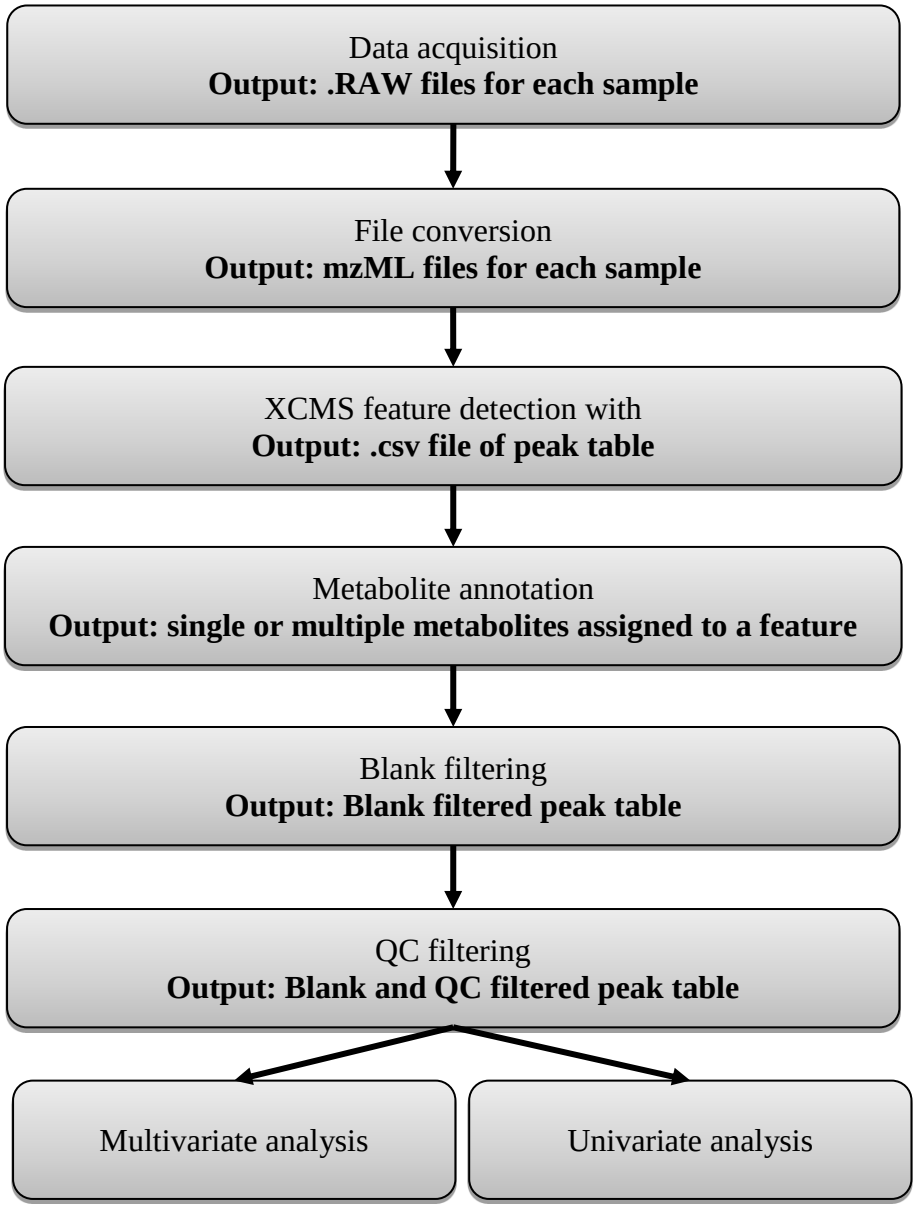


Figure 7.3: Data processing workflow.

7.2.1. File conversion

Each LC-MS run results in an output file whose format is dependent on the manufacturer of the instrument being used. These files are collectively known as .RAW files. The mzML file format (Martens et al, 2011) is an open source format for a number of different MS outputs including LC-MS. The XCMS feature detection software package (Smith et al, 2006; Tautenhahn et al, 2008) requires LC-MS files in the .mzML format. The .raw files for each LC-MS run are converted to the mzML format using the msConvert

application, which is part of the ProteoWizard toolkit (Chambers et al, 2012). In total 130 .RAW files across 14 groups (Table 7.1) are converted to .mzML files.

Table 7.1: Sample class mapping. There are 8 samples for each time point exposure combination. 30 QC samples and 4 blank samples are also injected.

CLASS	# SAMPLES
QC	30
Blank	4
Control 4h	8
Control 8h	8
Control 12h	8
Control 24h	8
Carbaryl 4h	8
Carbaryl 8h	8
Carbaryl 12h	8
Carbaryl 24h	8
Pb 4h	8
Pb 8h	8
Pb 12h	8
Pb 24h	8

7.2.2. XCMS feature detection

XCMS (Smith et al, 2006; Tautenhahn et al, 2008), part of the Bioconductor R package (Gentleman et al, 2004) is a software package for untargeted LC-MS feature detection and is one of the most widely used data processing tool for untargeted metabolomics (Benton et al, 2010; Kurczy et al, 2015). XCMS analyses the mzML files for peaks. The peaks detected in each file are then grouped across all the samples to produce a single peak table. LC-MS instruments are prone to drift over a period of time. This can result in the same peak in different samples being detected at different retention times in different samples. XCMS provides a retention time correction algorithm to correct for instrument drift.

The XCMS feature detection pipeline is carried out using a high-performance server-based Galaxy instance (Figure 7.4). Galaxy is a leading open source workflow platform

originally designed for next generation sequencing data analysis (Afgan et al, 2016). In recent years there has been a large community effort to introduce metabolomics tools into the Galaxy environment (Giacomoni et al, 2015; Goecks et al, 2010; Weber et al, 2017). The galaxy instance used in this case has been set up for metabolomics data processing and has many useful tools set up for use within it. XCMS Peak picking is performed using the centwave algorithm (Tautenhahn et al, 2008) using the parameters in Table 7.2. The values for the parameters are optimised taken from an optimisation (Nash et al, in preparation) using the IPO tool for automated optimisation of XCMS parameters (Libiseller et al, 2015).

Table 7.2: Centwave XCMS parameters used for data processing.

PARAMETER	VALUE
ppm	11
PeakWidth	3, 30
snthresh	10
prefilter	3, 100
mzdiff	0.001

The output after the XCMS workflow is a .csv file containing the peak table. The peak table contains peak intensities, with rows representing peaks, and columns representing samples. The resulting peak table contains 24,185 features across 130 samples.

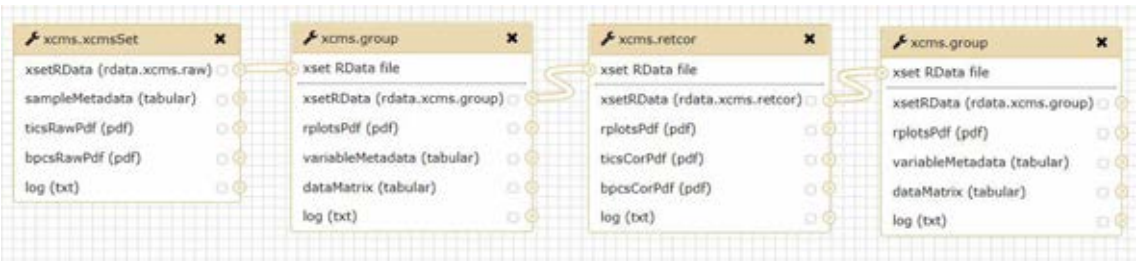


Figure 7.4: Galaxy XCMS workflow. .mzML files for each LC-MS run are used for feature detection using the xcms.xSet function. The features from each file are then grouped using the xcms.group function. The xcms.retcor function performs retention time correction, which accounts for instrument drift over the course of the whole analytical run. The xcms.group function must be called again after the retention time correction.

7.2.3. Metabolite annotation

Metabolite annotation is the process of assigning chemical formulas and thus chemical identities to MS spectra (section 1.6.4). Several software packages exist for automated annotation of metabolites in LC-MS data sets. In this instance the MI-Pack software (Weber & Viant, 2010) is used for metabolite annotation, and provides annotations matched to the KEGG Compound database (Kanehisa et al, 2014).

The MI-Pack software uses a mass-based annotation process to annotate metabolites using full scan MS spectrum. It assigns molecular formula to MS peaks and subsequently searches chemical databases using these derived formulas within a predefined ppm error. This has the potential to introduce false positive annotations, as many compounds can share the same molecular formula, and changes in the ppm tolerance used (in this case 2ppm was used) can introduce more candidate annotations (Lai et al, 2018). Ideally, once MS peaks have been identified to be of interest using statistical tests, MS/MS fragmentation spectra data should be collected to further increase confidence in the metabolite annotations (Dunn et al, 2013).

For the purpose of this study, MI-Pack is used to annotate the entire peak matrix, and the particular compounds of interest that are predicted to be effected are searched for in the list of Mi-Pack annotations. Of the 24,185 peaks, 6,450 (30%) have candidate KEGG Compound annotations.

7.2.4. Data filtering and missing value imputation

Blank subtraction and QC filtering is applied to the peak table. Blank subtraction involves subtracting the mean of any signals detected in the injected blank samples (see Table 7.1). Blank samples just contain the solvents used in the LC-MS method and the sample

preparation, so any signal that is measured when just the blank samples are analysed should not be included in the analysis and should be removed from the peak table.

QC samples are pooled from all other sample classes and are therefore chemically identical. QCs are used to assess the stability of the instrument over the course of the analysis. Any feature that has a relative standard deviation (RSD) value of greater than 25% in the QC samples are removed from the peak table as it is not a stably measured feature.

Missing value imputation is also performed on the peak table. The XCMS parameters will remove features that have excessively high missing values across the peak table based on the minfrac parameter, but some features will still have missing values for some of the samples. Missing value imputation looks at each class individually and finds features that do not have any value for at least one sample. Missing values can occur in peak tables because the metabolite is not detected for a sample, or the metabolite has a concentration lower than the detection limit (either the instruments detection limit or a limit set by the snthresh parameter in XCMS).

The k-nearest neighbour (KNN) missing value imputation method (Steuer et al, 2007; Troyanskaya et al, 2001) is recommended for use with UHPLC-MS data sets (Di Guida et al, 2016) and is used for this study. KNN imputation replaces missing values by taking the average of the 10 nearest non-missing values for the given feature within the same sample class. Euclidean distance is used as the measure of nearness. The KNN missing value imputation has the advantage of providing each missing value with a unique number, therefore maintaining some of the natural variance that is expected in a UHPLC-MS dataset (Di Guida et al, 2016).

7.2.5. Normalisation

The peak table undergoes normalisation and scaling before statistical analysis. Different approaches are applied depending on whether multivariate or univariate analyses are being performed. The recommendations from the extensive investigation carried out in (Di Guida et al, 2016) are followed (see Table 7.3).

Table 7.3: Summary of normalisation, missing value imputation and transformation/scaling techniques applied to the dataset for each statistical analysis.

	Normalisation	**Missing value imputation**	**Data transformation**	**Data scaling**
PCA	PQN	KNN	Generalised logarithm	None
PLS-DA	PQN	KNN	Generalised logarithm	None
Univariate	PQN	None	None	None

7.3. Statistical analysis

Several established statistical methodologies and workflows exist for interrogating untargeted metabolomics data sets. In a toxicology study, the goal of the statistical analysis is to assess if there are statistically significant differences in metabolomics responses between control and treatment experimental classes. If these differences are established, the features or peaks that contribute to these significant differences can be extracted. These features can be used to form a biological interpretation which in this context, can be used to validate the computationally generated hypotheses outlined in Chapter 5.

Multivariate and univariate statistical analysis is performed for each exposure separately and used to identify features that contribute to observed changes between treatment and control groups. The identified features are then used to make biological interpretations

and to assess the validity of the computationally generated hypotheses for each exposure separately in section 7.4.

7.3.1. Carbaryl treatment

7.3.1.1. Multivariate statistics

Principal Component Analysis (PCA) is performed on the Control and Carbaryl experimental classes. Plots of the pairwise comparisons of the first three principal components are generated (Figure 7.5, see Appendix E) These plots visualise the underlying multivariate variance between the experimental classes. Ideally samples from the same experimental classes should be tightly clustered on the plots, with inter-class separations between control and treatment sample groups.

No substantial inter-class separation is observed when looking at all 8 groups with no single principal component (PC) revealing substantial differences in the underlying variance between experimental classes. The Control 24h and Carbaryl 24h groups are the most grouped classes on the plots, with all other six classes displaying a wide spread across the plot.

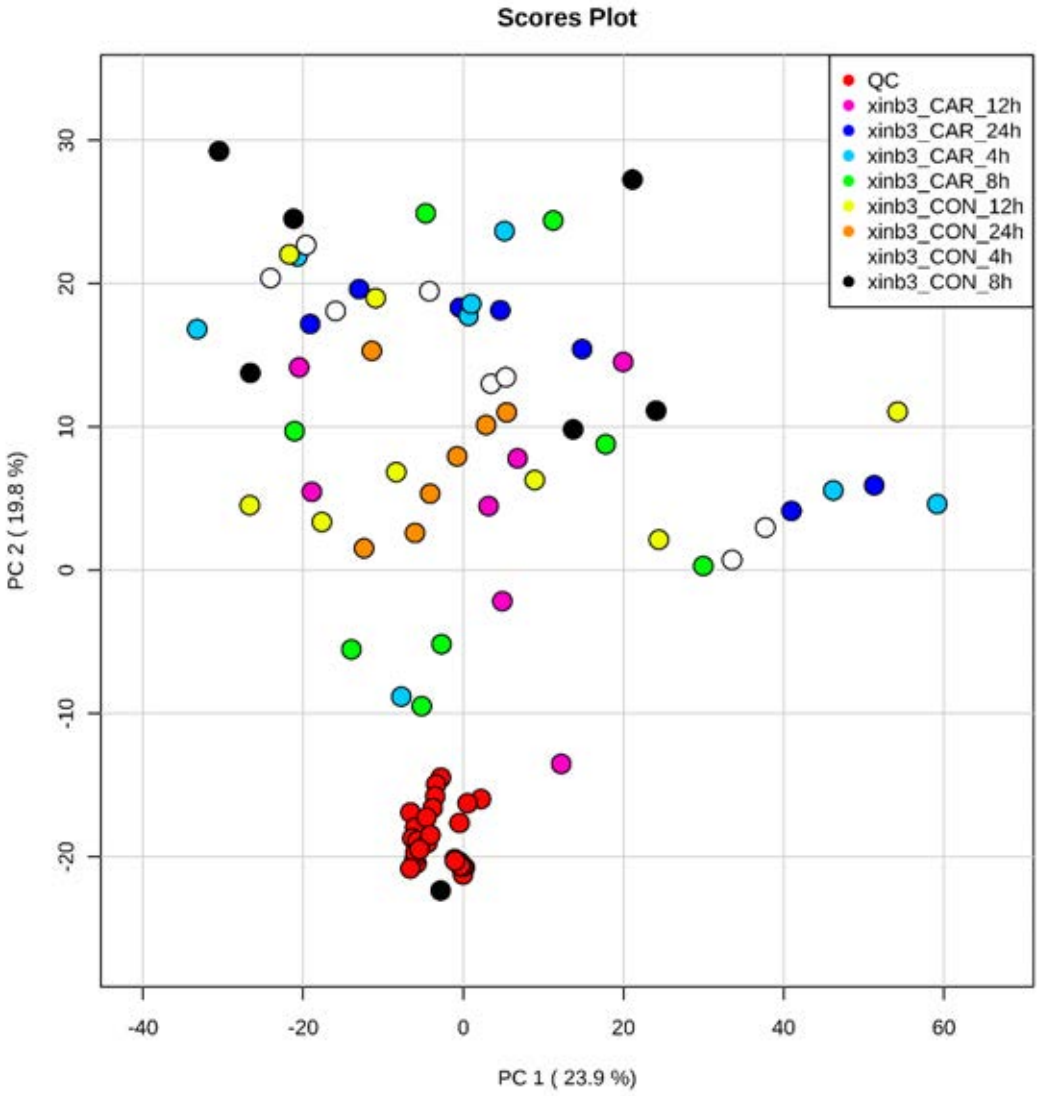


Figure 7.5: PC1 vs PC2 PCA plot of all Carbaryl and Control groups.

2 class PCA summary plots are subsequently generated that compare the Control classes with the Carbaryl treatment classes at each time point (see Appendix E). The 24h time point plot is the only plot where some separation between the classes is observed. Figure 7.6 shows the PC1 vs PC2 plot of the 24h time point. There is some multivariate variance between the Control and Carbaryl exposure 24h time point experimental classes.

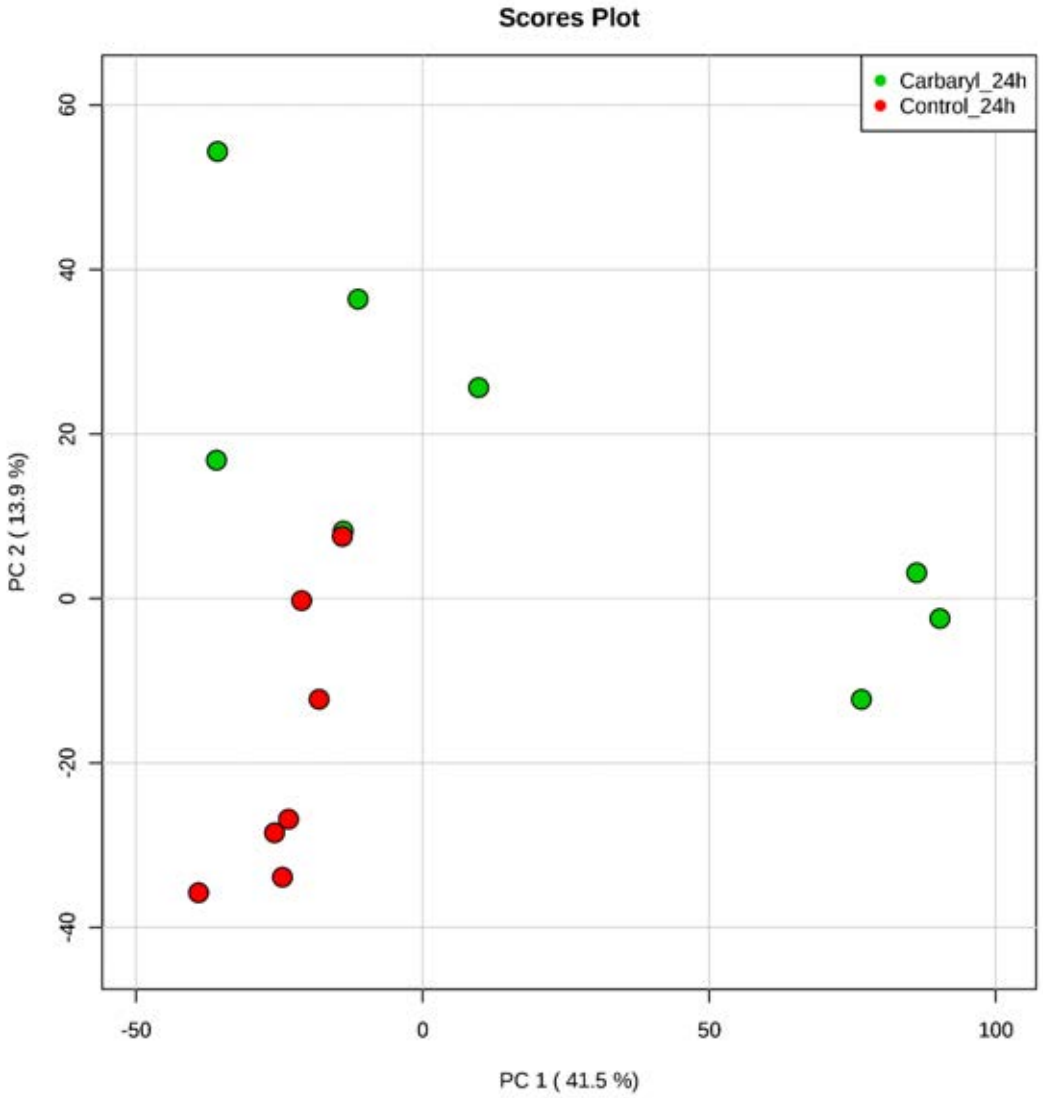


Figure 7.6: PC1 vs PC2 plot of Control 24h and Carbaryl 24h.

Partial Least Squares – Discriminant Analysis (PLS-DA) is performed on the Control and Carbaryl 24h classes to see if they can be separated. PLS-DA is a supervised classification technique that uses labelled data, whereas PCA is an unsupervised clustering technique that uses no prior knowledge. PLS-DA enhances the separation between groups of observations by rotating PCA components such that a maximum separation among these classes is obtained (Shaffer, 2002). The first and second PLS-DA components are plotted in Figure 7.7, and show that the PLS-DA model successfully separated the classes.

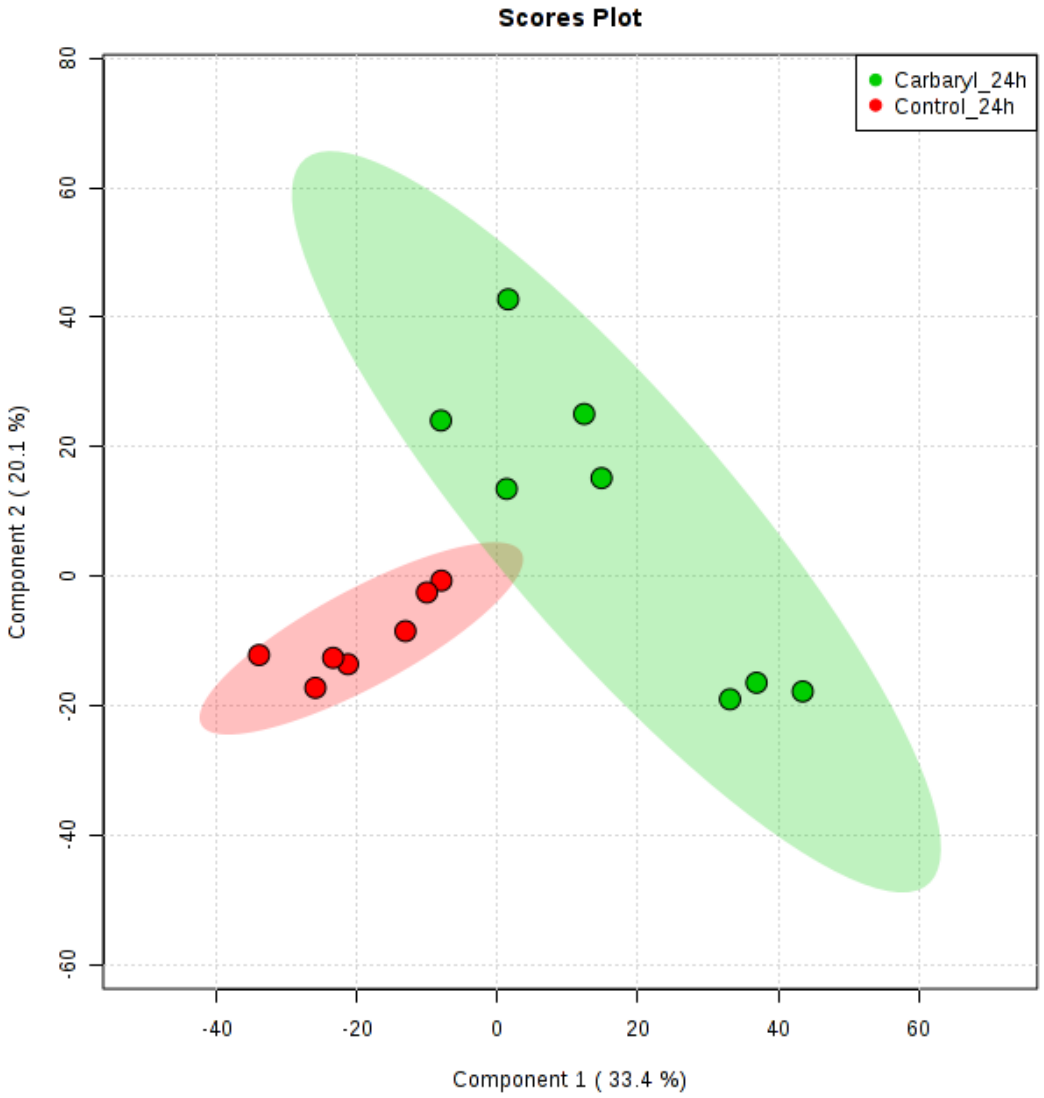


Figure 7.7: PLS-DA plot of the Control and Carbaryl 24h time point experimental classes.

PLS-DA models require validation and assessment to make sure that they aren't over-trained or over fitted. 10-fold cross validation is performed on the PLS-DA model to validate the classification model. This results in two important measures for quantitively assessing the performance of the model, R^2 and Q^2.

R^2 is a correlation index and is a quantitative measure between 0 and 1 that indicates how well the PLS-DA model can mathematically reproduce the data in the data set. Q^2 is also

a measure between 0 and 1 that measures to the quality of prediction and can be used as an indicator of whether the PLS-DA model is overfitted. A well fit PLS-DA model will have a R^2 value of 0.7 or 0.8, and a Q^2 value of greater 0.5 is acceptable, with a value of 0.9 deemed outstanding (Szymanska et al, 2012).

Figure 7.8 and Table 7.4 show the performance metrics for the 10-fold cross validation of the Control and Carbaryl 24h experimental classes PLS-DA model. The metrics are calculated for increasing number of components used in the model. 2 model components give the best R^2 and Q^2 values (0.85 & 0.57)

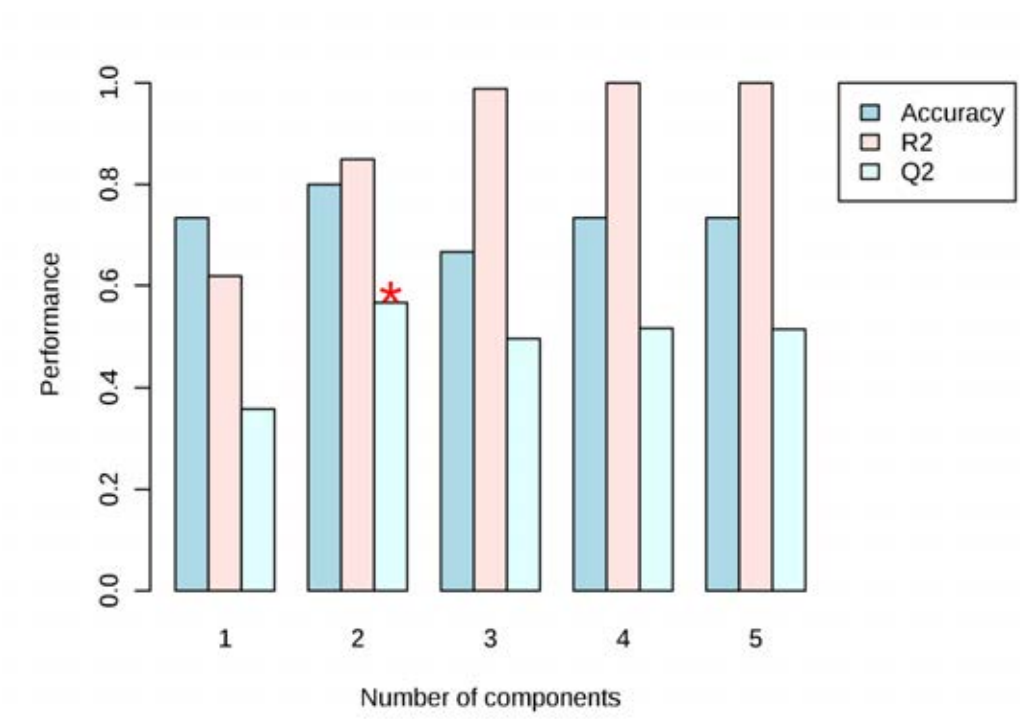


Figure 7.8: 10-fold cross validation of the PLS-DA model for the Control and Carbaryl 24h time point experimental classes.

Table 7.4: 10-fold cross validation performance metrics

Measure	1 comps	2 comps	3 comps	4 comps	5 comps
Accuracy	0.73333	0.8	0.66667	0.73333	0.73333
R2	0.6193	0.84964	0.98809	0.99956	0.99995
Q2	0.35854	0.56595	0.49492	0.51533	0.51389

Permutation testing is also performed on the PLS-DA model. Permutation testing gives a level of significance to the PLS-DA model in the form a p-value. The test creates several random PLS-DA models and checks their accuracy at classifying the samples into correct groups. The tests assume that there is no difference between the two randomly formed groups, with the sample labels being randomly permuted before a new classification model is produced. For each permuted model, the R^2 and Q^2 values are calculated and the number of times that the randomly generated models outperform the previously constructed model are recorded (Szymanska et al, 2012). 2,000 permutations are performed on the PLS-DA model, and the resulting p-value is 0.114. This value is higher than the ideal value of < 0.05.

A Variable Importance Projection (VIP) score can be calculated for each feature in the PLS-DA model. The VIP value measures how much each feature contributes to the separation in the PLS-DA model, it is calculated as a weighted sum of square of the PLS loadings for each component of the PLS-DA model. Figure 7.9 plots the VIP scores for the features that have the highest 25 values for the first component of the PLS-DA model. Variables with VIP scores of greater than 1 are considered to have substantial influence on the separation observed in the PLS-DA model. Since the separation on the PLS-DA plot (Figure 7.7) is achieved through a combination of the first and second components, features that have a VIP score of greater than 1 for both the first and second components are selected for biological interpretation (section 7.4.1). This amounts to a total of 480 features.

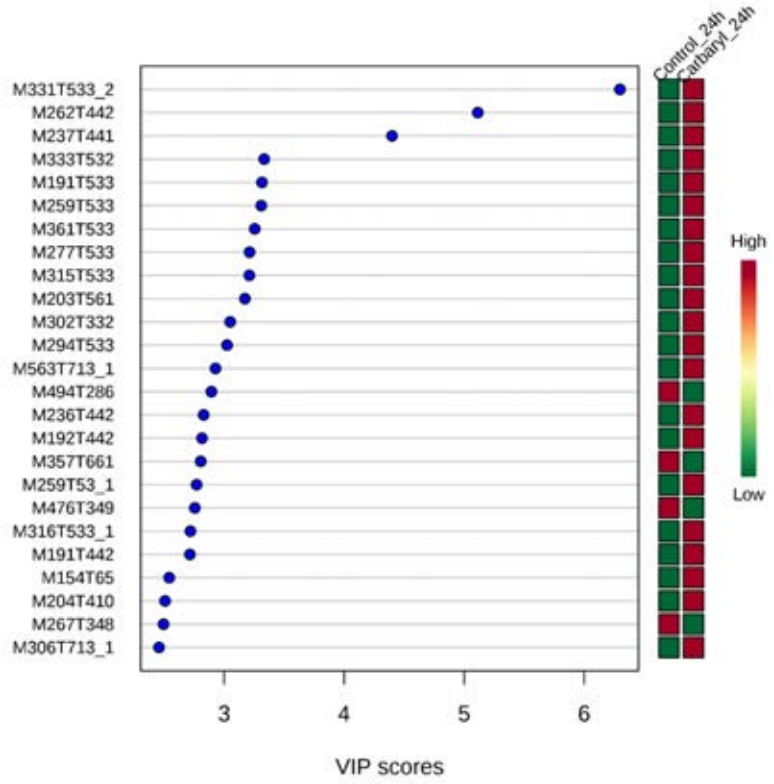


Figure 7.9: VIP scores plot of the 25 features that have the greatest VIP scores for the first component of the PLS-DA model.

7.3.1.2. Univariate statistics

For each time point, three analyses are performed to identify peaks that contribute to any differences between the control and treatment groups at that time point. A student's t-test with Benjamini-Hochberg false discovery rate (FDR) correction is applied to assess the significance of peak intensity changes between groups. Fold change analysis is used to identify peaks that have substantially different intensities between the control and treatment groups. A volcano plot is also generated which combines the t-test and fold change analysis to identify peaks that are both significantly different between groups and have a significant fold change. The t-test requires peaks to have an adjusted p-value less than 0.05 to be significant. Peaks with fold change values that are at least ±2.0 are selected during fold change analysis. The volcano plots highlight peaks that have both an adjusted

p-value of at least 0.1 and a fold change of at least ±2.0. Table 7.5 contains the number of significant or substantial peaks selected from each analysis.

Figure 7.10 is the t-test plot, Figure 7.11 is the fold change plot and Figure 7.12 is the volcano plot for the 24hr time point. Figures for the 4h, 8h and 12h time points are found in Appendix E.

Table 7.5: Number of significant peaks selected from the t-test, the number of peaks with a substantial fold change and the number of significant peaks identified from volcano plots for the Control and Carbaryl groups at each time point.

Time Point	T-Test: q-value < 0.05	Fold Change > ± 2.0	Volcano: q-value < 0.1 & Fold change > ±2.0
4h	0	199	2
8h	0	681	0
12h	0	416	0
24h	633	1,298	643

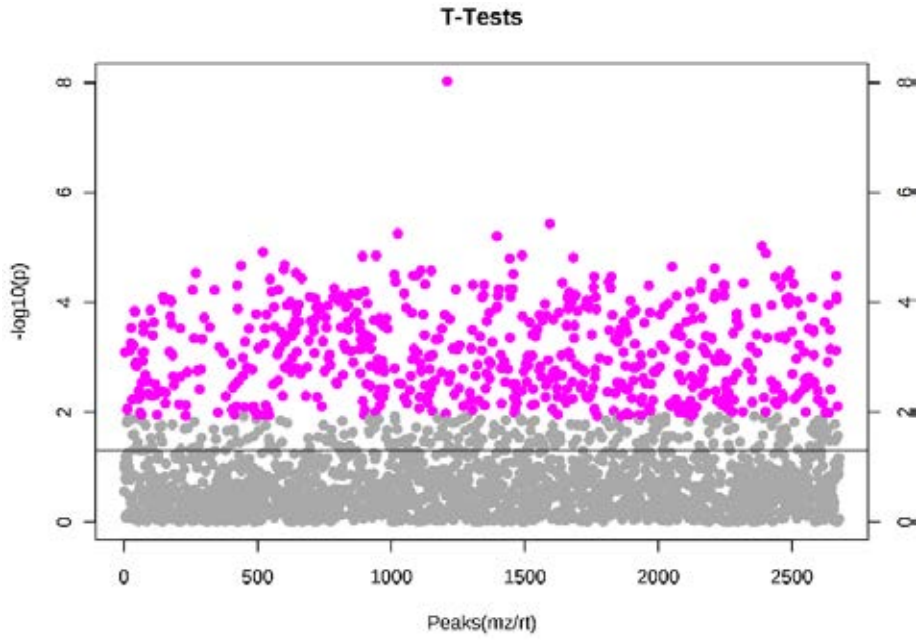


Figure 7.10: T-test plot for the Control vs Carbaryl 24h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 663 peaks have an FDR corrected p-value of less than 0.05

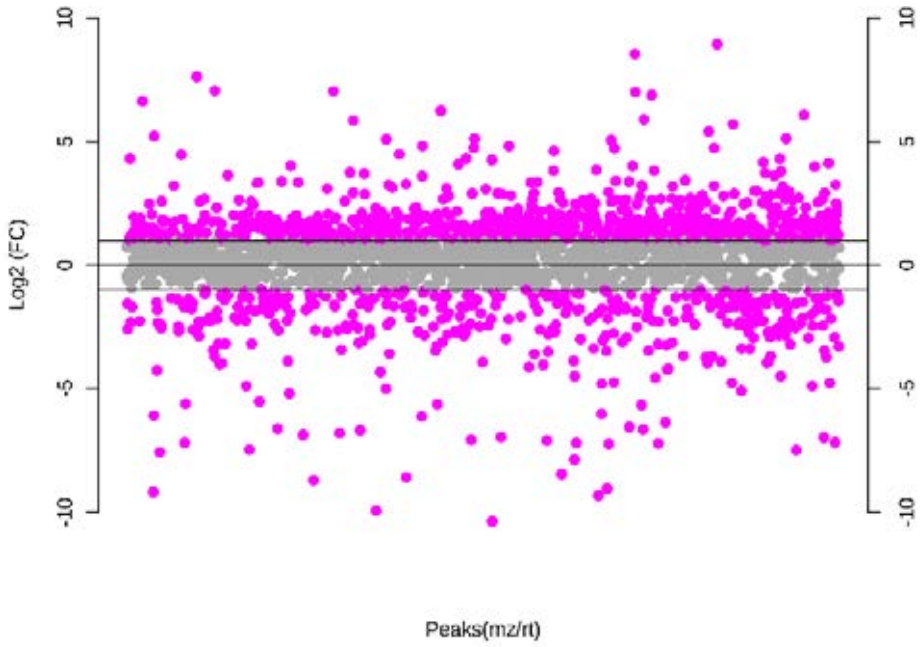


Figure 7.11: Fold change plot for the Control vs Carbaryl 24h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 1,298 peaks.

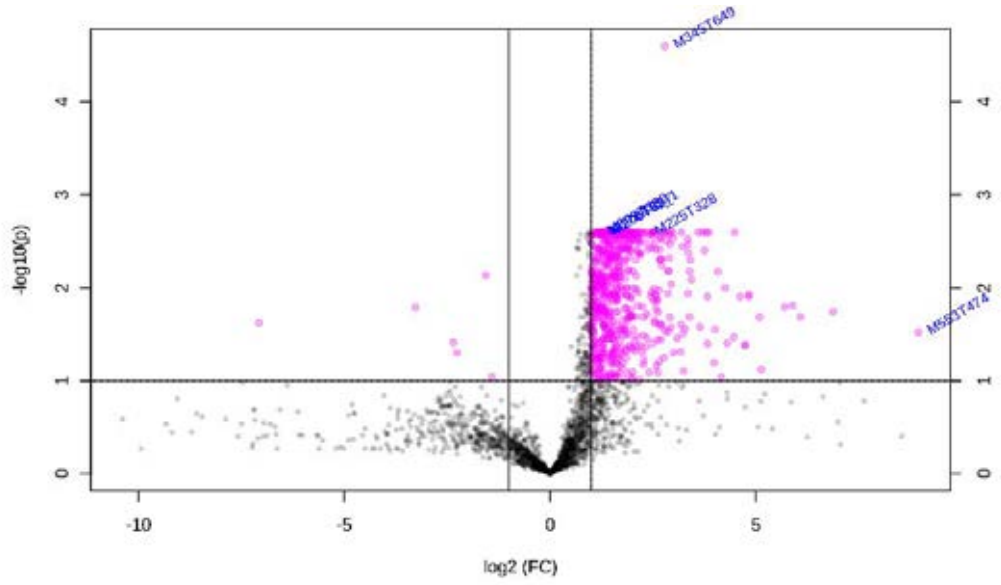


Figure 7.12: Volcano plot for the Control vs Carbaryl 24h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 643 peaks.

All time points have peaks with substantial fold changes, however these do not consider any statistical significance and thus have no associated p-values. The t-tests and volcano plots do assign statistical significance, and for these tests, just the 24h time point yields

significant peaks, 633 from the t-tests and 643 from the volcano plot. The 24h time point also has by far the most peaks with fold changes > ±2.0. The only exception to this is that 2 peaks are identified in the volcano plot for the 4h time point.

7.3.2. Lead treatment

7.3.2.1. Multivariate statistics

Principal Component Analysis (PCA) is performed on the Control and Lead experimental classes. Plots of the pairwise comparisons of the first three principal components are generated (Figure 7.13, Appendix E). These plots visualise the underlying multivariate variance between the experimental classes. Ideally samples from the same experimental classes should be tightly clustered on the plots, with inter-class separations between control and treatment sample groups.

No substantial inter-class separation is observed when looking at all 8 groups with no single principal component (PC) revealing substantial differences in the underlying variance between experimental classes. The Control 24h and Lead 24h groups are the most grouped classes on the plots, with all other six classes displaying a wide spread across the plot.

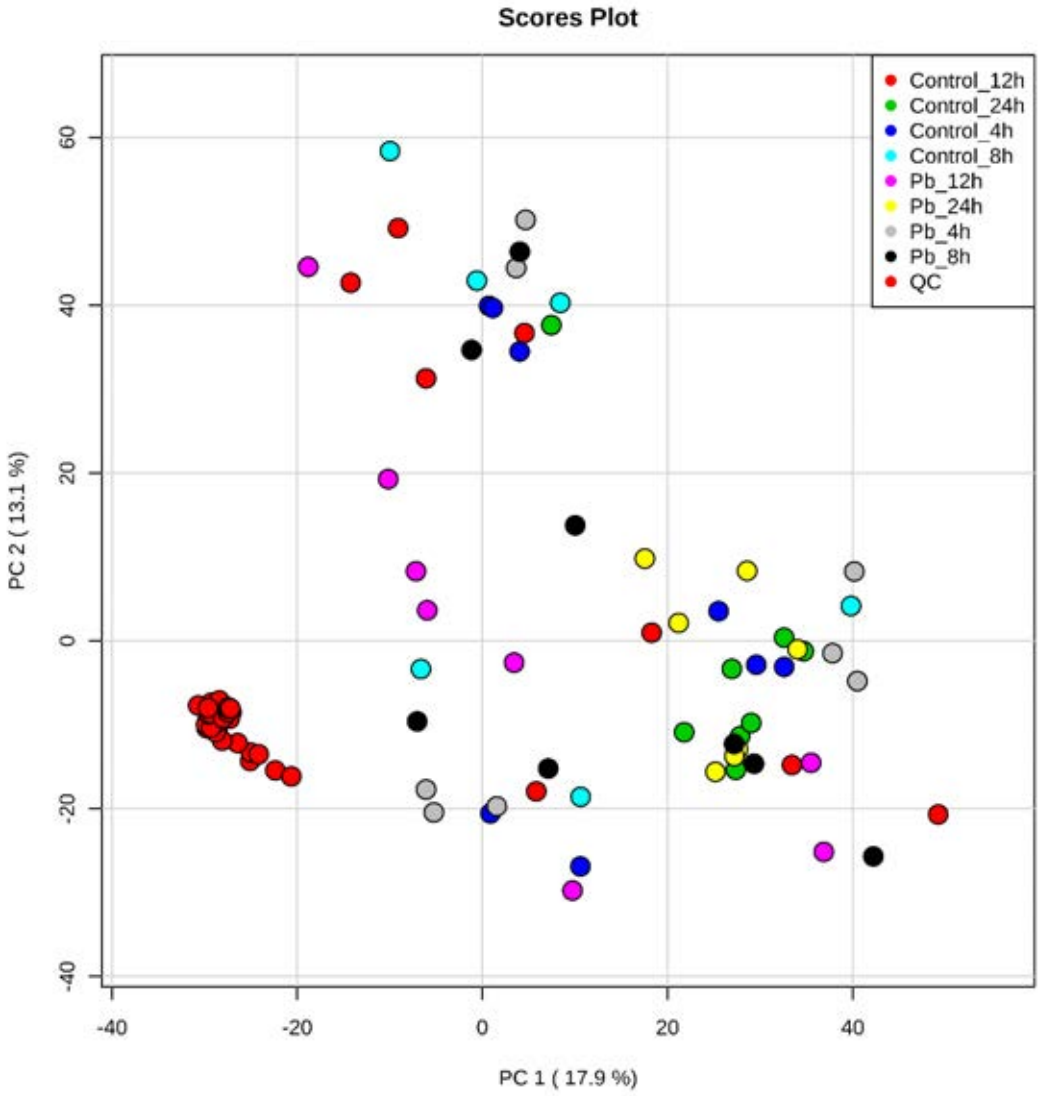


Figure 7.13: PC1 Vs PV2 plot of all Lead and Control groups.

2 class PCA summary plots are subsequently generated that compare the Control with the Lead treatment at each time point. PCA plots can be found in Appendix E. No substantial separation is observed between treatment and control groups at any of the time points.

PLS-DA is performed for the control and lead treatment classes at each time point to see if the classes can be separated. None of the PLD-DA models had acceptable R^2 or Q^2 values.

7.3.2.2. Univariate statistics

For each time point, three analyses are performed to identify peaks that contribute to any differences between the control and treatment groups at that time point. A student's t-test with Benjamini-Hochberg false discovery rate (FDR) correction is applied to assess the significance of peak intensity changes between groups. Fold change analysis is used to identify peaks that have substantially different intensities between the control and treatment groups. A volcano plot is also generated which combines the t-test and fold change analysis to identify peaks that are significantly different between groups and have a significant fold change. The t-test requires peaks to have an adjusted p-value less than 0.05 to be significant. Peaks with fold change values that are at least ±2.0 are selected during fold change analysis. The volcano plots highlight peaks that have both an adjusted p-value of at least 0.1 and a fold change of at least ±2.0. Table 7.6 contains the number of significant or substantial peaks selected from each analysis.

Figure 7.14 is the t-test plot, Figure 7.15 is the fold change plot and Figure 7.16 is the volcano plot for the 8hr time point. Figures for the 4h, 12h and 24h time points are found in Appendix E.

Table 7.6: Number of significant peaks selected from the t-test, the number of peaks with a substantial fold change and the number of significant peaks identified using volcano analysis for the Control and Lead groups at each time point.

Time Point	T-Test: q-value < 0.05	Fold Change > ± 2.0	Volcano: q-value < 0.1 & Fold change > ±2.0
4h	0	448	0
8h	0	806	0
12h	0	336	0
24h	0	446	0

All time points have peaks with substantial fold changes, however these do not consider any statistical significance and thus have no associated p-values. The t-tests and volcano

plots do assign statistical significance, and for these tests, none of the time points yield significant peaks.

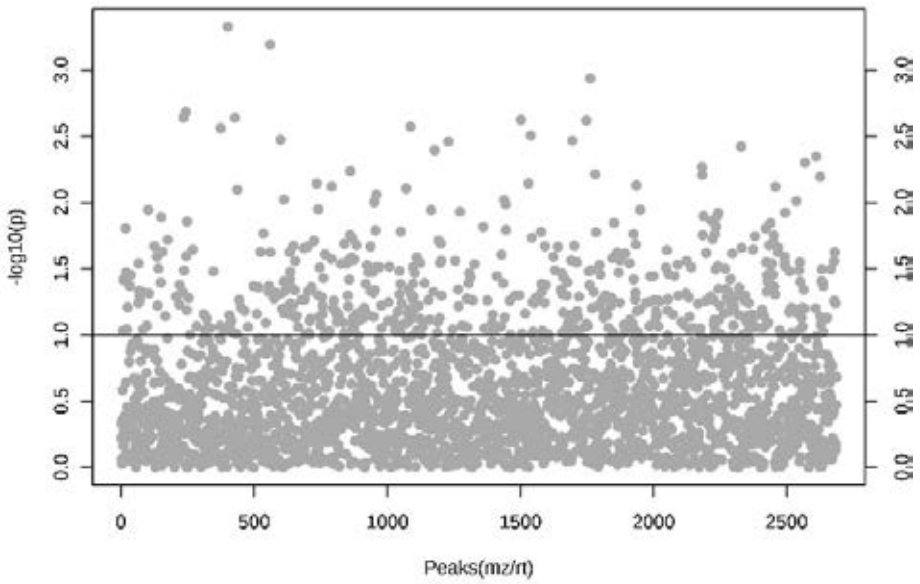


Figure 7.14: T-test plot for the Control vs Lead 8h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

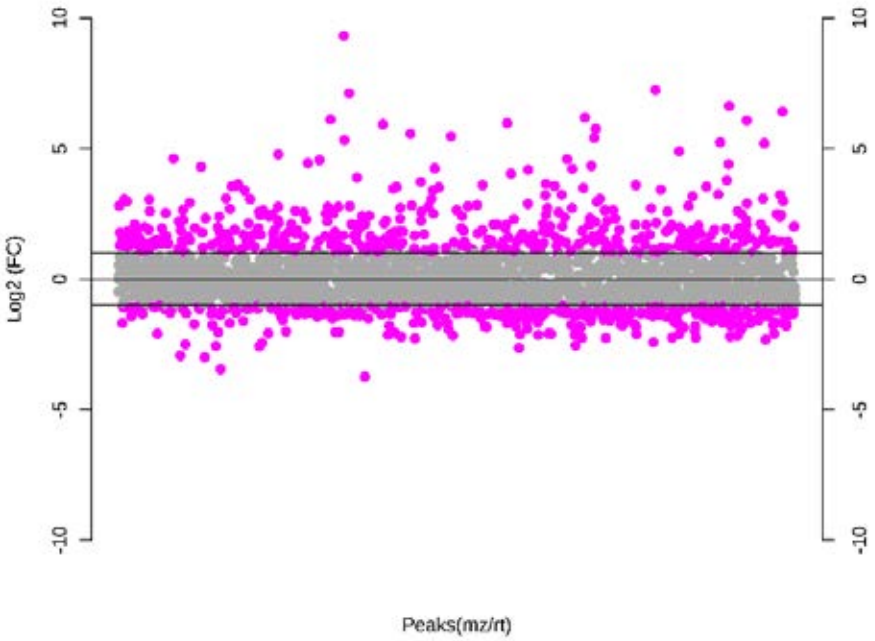


Figure 7.15: Fold change plot for the Control vs Lead 8h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 806 peaks.

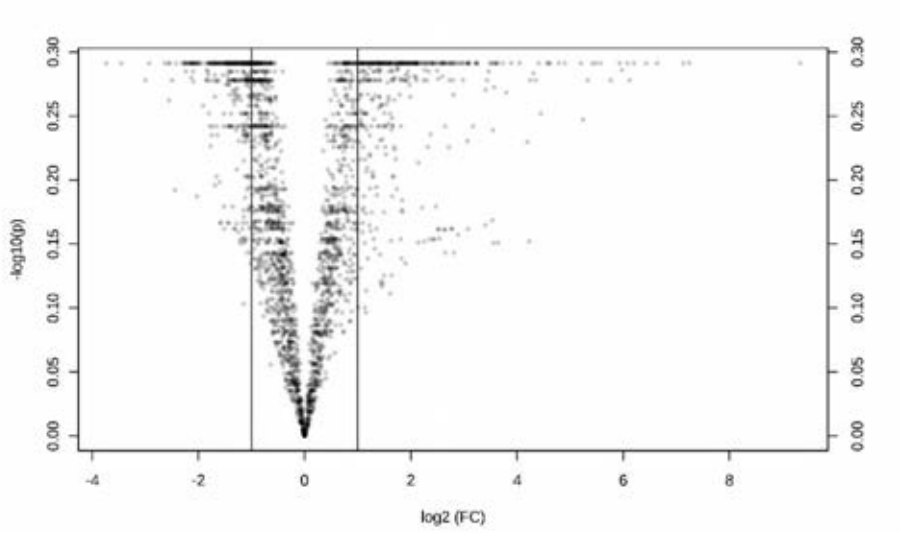


Figure 7.16: Volcano plot for the Control vs Lead 8h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.

7.4. Interpretation

7.4.1. Carbaryl treatment

The volcano plot (see Appendix E) and PLS-DA (Figure 7.7) revealed significantly changing peaks between the Control and Carbaryl treated samples at the 24h time point. A total of 643 peaks were obtained from the volcano plot, and 480 from the PLS-DA. 192 peaks are common between the two statistical analyses (

Figure 7.17). These 192 peaks have been identified as significant, or discriminatory between the experimental classes using different statistical methods that each have their own assumptions about the data. Therefore, these 192 peaks can be considered less likely to be false positives.

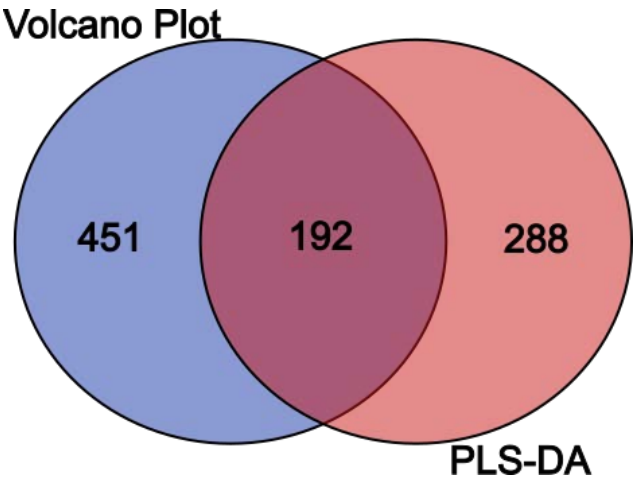


Figure 7.17: Venn diagram showing the overlap between the features obtained from the volcano plot and PLS-DA analysis for the Carbaryl 24h exposure group.

Each of the three groups of peaks are checked against the MI-Pack metabolite annotations mapping them to entries in the KEGG Compound database (section 7.2.3). The hypotheses outlined in Chapter 5 provide predictions of the metabolic effect of the Carbaryl treatment on three levels; KEGG modules, KEGG pathways and areas of metabolism. The same analyses used to form these predictions are repeated for each of three groups of peaks identified during the statistical analysis to assess the accuracy of the predictions. KEGG Mapper (Kanehisa, 2013) is used to identify KEGG modules, KEGG pathways and areas of metabolism that contain the statistically identified and KEGG annotated compounds. The KEGG annotated peaks are passed to the KEGG mapper software to find which peaks are mapped to KEGG modules and KEGG pathways (Table 7.7).

414 (64%) of the 643 peaks obtained from the volcano plot, 457 (95%) of the 480 peaks identified using PLS-DA and 183 (84%) of the 192 peaks common between the two analyses have MI-Pack KEGG annotations.

Table 7.7: The number of peaks from the statistical analysis of the Carbaryl 24h dataset that have KEGG annotations and are mapped to KEGG pathways and modules.

	# Peaks	# Annotations	# in KEGG Module	# in KEGG Pathway
Volcano plot	643	414	9	108
PLS-DA	480	457	12	99
Common	192	183	3	48

7.4.1.1. KEGG modules

For each set of annotated KEGG compounds (Volcano, PLS-DA and the common compounds between them), KEGG Mapper (Kanehisa, 2013) is used to identify KEGG modules that contain the annotated compounds using the same approach that was used to generate the predictions (section 5.3.2). The KEGG modules identified from each set are labelled as *observed*. A total of 18 KEGG modules were predicted to be effected by the Carbaryl exposure (section 5.3.2.1), and these are labelled as *predicted*. The overlap between the predicted and observed KEGG modules is calculated. Table 7.8 records how many KEGG modules are common between the predicted and observed and shows the precision, recall and F-measure of the computationally generated predictions.

For the volcano compounds, 9 (2%) of the 414 annotated peaks are present in 7 KEGG modules, none of which are in common with the 18 predicted KEGG modules.

For the PLS-DA compounds, 12 (3%) of the 457 annotated peaks are present in 7 KEGG modules, with 1 KEGG module being both predicted and observed. This results in; a precision value of 5.56%, a recall value of 14.29% and a F-measure of 8%.

For the compounds that are both in the volcano set and the PLS-DA set, 3 (2%) of the 183 annotated peaks are present in 4 KEGG modules, none of which are in common with the 18 predicted KEGG modules.

Table 7.8: Precision, recall and F-Measure for the assessment of the computationally generated predictions for the Carbaryl exposures in terms of KEGG modules.

KEGG Modules	**Volcano**	**PLS-DA**	**Common**
Predicted	18	18	18
Observed	7	7	4
Overlap	0	1	0
Precision	0%	5.56%	0%
Recall	0%	14.29%	0%
F-Measure	N/A	8.00%	N/A

7.4.1.2. KEGG pathways

For each set of annotated KEGG compounds (Volcano, PLS-DA and the common compounds between them), KEGG Mapper (Kanehisa, 2013) is used to identify KEGG pathways that contain the annotated compounds using the same approach that was used to generate the predictions (section 5.3.2). The KEGG pathways identified from each set are labelled as *observed*. A total of 15 KEGG pathways were predicted to be effected by the Carbaryl exposure (section 5.3.2.1), and these are labelled as *predicted*. The overlap between the predicted and observed KEGG pathways is calculated. Table 7.9 records how many KEGG pathways are common between the predicted and observed and shows the precision, recall and F-measure of the computationally generated predictions.

For the volcano compounds, 108 (26%) of the 414 annotated peaks are present in 32 KEGG pathways, with 8 KEGG pathways being both predicted and observed. This results in; a precision value of 53.33%, a recall value of 25% and a F-Measure of 34.04%.

For the PLS-DA compounds, 99 (22%) of the 457 annotated peaks are present in 32 KEGG pathways, with 9 KEGG pathways being both predicted and observed. This results in; a precision value of 60%, a recall value of 28.13% and a F-Measure of 38.3 %.

For the compounds that are both in the volcano set and the PLS-DA set, 48 (26%) of the 183 annotated peaks are present in 16 KEGG pathways, with 4 KEGG pathways being both predicted and observed. This results in; a precision value of 26.67%, a recall value of 25% and an F-Measure of 25.81%.

Table 7.9: Precision, recall and F-Measure for the assessment of the computationally generated predictions for the Carbaryl exposures in terms of KEGG pathways.

KEGG Pathways	**Volcano**	**PLS-DA**	**Common**
Predicted	15	15	15
Observed	32	32	16
Overlap	8	9	4
Precision	53.33%	60.00%	26.67%
Recall	25.00%	28.13%	25.00%
F-Measure	34.04%	38.30%	25.81%

A hypergeometric distribution is calculated to assess the likelihood of observing an overlap between the predicted and observed KEGG pathways under a null model for the Volcano, PLS-Da and Common compounds. The values are calculated based on the 526 KEGG pathways present in the KEGG database. The probability mass values are 3.518e-7, 1.34562e-8 and 6.0847e-4 respectively, showing that getting these results randomly is unlikely.

7.4.1.3. Areas of metabolism

Each of the areas of metabolism to which the observed KEGG pathways belong are compared with the areas of metabolism that were derived from the predicted KEGG pathways in section 5.3.2.1. A total of 6 areas of metabolism are predicted to be effected by the Carbaryl exposure. The overlap between the predicted and observed areas of metabolism is calculated. Table 7.10 records how many areas of metabolism are common

between the predicted and observed and shows the precision, recall and F-measure of the computationally generated predictions.

For the volcano compounds, 8 areas are observed. There is an overlap of 5 between the 6 areas predicted and the 8 areas observed. This results in; a precision value of 83.33%, recall value of 62.5% and an F-measure of 71.43%.

For the PLS-DA compounds, 8 areas are observed. There is an overlap of 5 between the 6 areas predicted and the 8 areas observed. This results in; a precision value of 83.33%, a recall value of 62.5% and an F-Measure of 71.43 %.

For the compounds that are both in the volcano set and the PLS-DA set, 6 areas are observed. There is an overlap of 4 between the 6 areas predicted and the 6 areas observed. This results in; a precision value of 66.67%, a recall value of 66.67% and a F-Measure of 66.67%.

Table 7.10: Precision, recall and F-Measure for the assessment of the computationally generated predictions for the Carbaryl exposures in terms of Areas of Metabolism

Areas of metabolism	**Volcano**	**PLS-DA**	**Common**
Predicted	6	6	6
Observed	8	8	6
Overlap	5	5	4
Precision	83.33%	83.33$	66.67%
Recall	62.50%	62.50%	66.67%
F-Measure	71.43%	71.43%	66.67%

A hypergeometric distribution is calculated to assess the likelihood of observing an overlap between the predicted and observed KEGG areas of metabolism under a null model for the Volcano, PLS-Da and Common compounds. The values are calculated based on the 12 KEGG areas of metabolism present in the KEGG database. The probability mass values are 0.16317, 0.16317 and 0.183566 respectively. These values

are higher than ideal, but this is a very low granularity way of looking at effects on a metabolic system so should not be considered to be a highly accurate way of assessing the computationally generated hypotheses.

7.4.2. Lead treatment

The statistical analysis of the Lead exposure datasets (section 7.3.2) did not reveal any significantly changing peaks between treatment and control experimental classes. Therefore, no assessment of the accuracy of the computationally generated hypothesis (section 5.3.2.2) can be performed.

7.5. Discussion

This chapter has presented a metabolomics study carried out with the purpose of assessing the validity of the computationally generated predictions that predict how the metabolome of *D. magna* is effected by exposure to two environmental stressors, Carbaryl and Lead (Chapter 5). The experimental design mimicked the experiment that generated the transcriptomics datasets (Orsini et al, 2016) that are used as part of the computational hypothesis generation. These experiments recorded data at the 4h time point, and for the metabolomics study this time point as well as three other later time points (8h, 12h and 24h) are used with the purpose of catching the expected temporal delay in the metabolome response.

HPLC-MS was used to make the metabolomics measurements, and the typical metabolomics mass spectrometry based workflow (Figure 7.1) was followed. Section 7.1 details the sample preparation and data acquisition, and section 7.2 details the data processing. Statistical analysis was carried out on the resultant peak table to identify peaks that are significantly different between treatment and control groups, with sample classes

partitioned so that for each statistical analysis, one treatment is compared to the control group for each time point separately.

PCA analysis failed to reveal substantial intra-class clustering or inter-class separation for all but one of the compared experimental classes, the Carbaryl 24h and the Control 24h groups. The same is true of the PLS-DA analysis, with only the model built using Carbaryl 24h vs Control 24h groups having an acceptable performance. Univariate analyses told the same story, failing to uncover any significant differences between groups apart from the Carbaryl 24h and Control 24h groups.

To investigate this further, the RSDs of features across sample groups is calculated. First the RSDs of features in the QC samples in inspected using the raw peak intensities, Figure 7.18 summarises them in a box plot. The mean RSD is 11.28% which shows that there is a small and acceptable amount of analytical variation in the study.

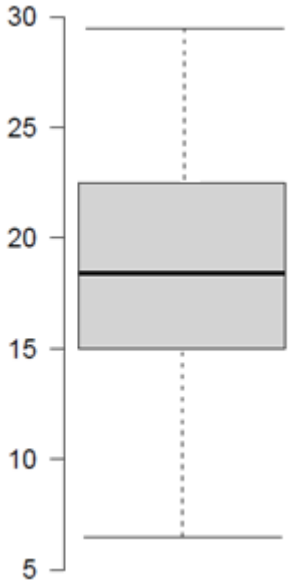


Figure 7.18: Box plot summarising the RSDs of features in the QC samples.

The RSDs of each feature is then calculated across each experimental class. This calculation is performed on the peak matrix that has been processed for univariate

statistics (see Table 7.3). Figure 7.19 shows box plots summarising the distribution of these RSD values.

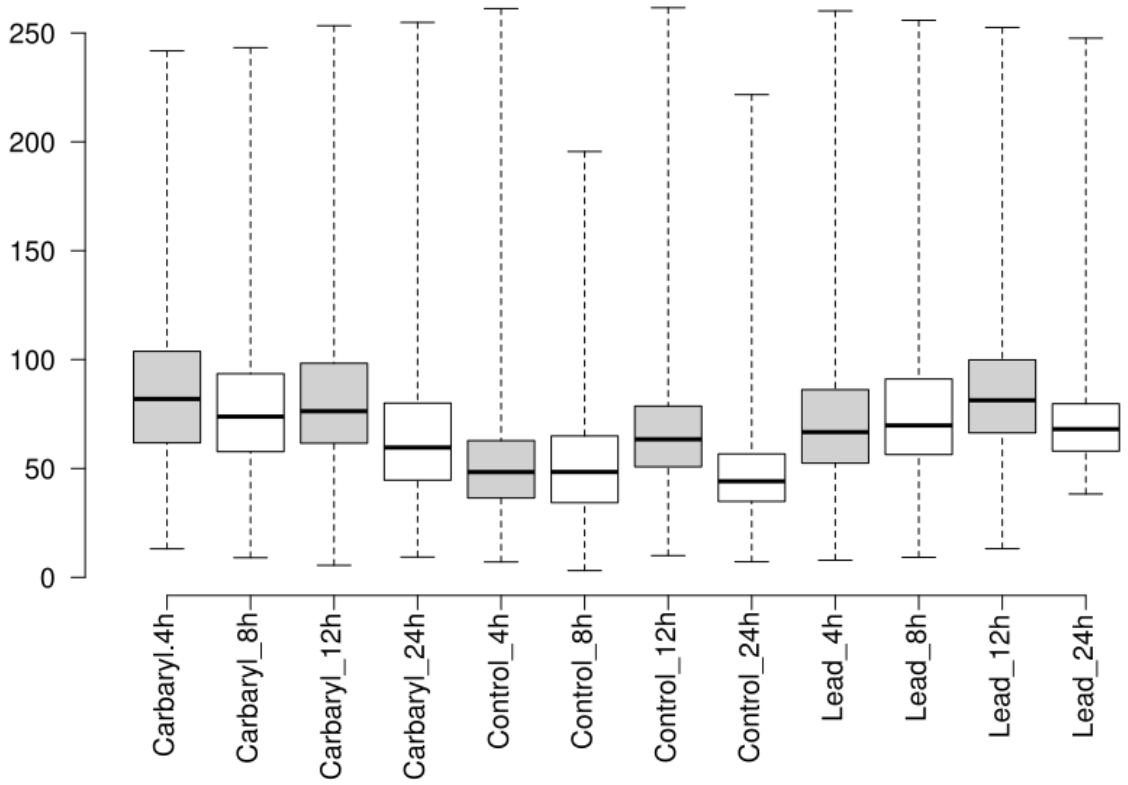


Figure 7.19: Box plots summarising the RSDs of all features across all experimental classes.

The median RSD values are high for all experimental classes, with the lowest values being 44.13% for the Control 24h class and 44.65% for the Carbaryl 24h class. It is interesting to note that these are the two classes whose comparison yielded some statistical significance. The high variability in the experimental classes, particularly the Carbaryl and Lead exposure classes explains why the statistical tests failed to find significant differences between classes, with no assessment possible for the Lead predictions. This suggests issues in the sample preparation stage.

Table 7.11: Median RSDs for all features for each experimental class

Group	Median
Control 4h	48.35
Control 8h	48.36
Control 12h	63.41
Control 24h	44.13
Carbaryl 4h	81.95
Carbaryl 8h	73.79
Carbaryl 12h	76.28
Carbaryl 24h	44.65
Lead 4h	66.67
Lead 8h	69.76
Lead 12h	81.33
Lead 24h	68.06

For the Carbaryl and Control 24h groups' comparison, the statistically significant peaks identified using a volcano plot and PLS-DA were used to assess the quality of the computationally generated predictions. The peaks were first checked to see if they have MI-Pack metabolite annotations. For the 739 unique peaks identified from the volcano plot and PLS-DA model, there are 677 MI-Pack KEGG annotations.

These 677 KEGG compounds are passed to the KEGG Mapper software to find which KEGG modules and KEGG pathways that they participate in. This mirrors the same approach that is used to formulate the predictions using the active modules approach for the computational generation of hypotheses in Chapter 5. The KEGG modules and pathways that contain the 677 KEGG compounds are identified and compared with the predicted KEGG modules and pathways. Only around 2-3% of the KEGG annotations can be linked to KEGG modules, and around 25% can be linked to KEGG pathways.

Table 7.8 summarises the assessment of the KEGG module predictions. With such low representation of the KEGG annotations in the KEGG modules, it is almost meaningless to use KEGG modules to assess the quality of the computationally generated predictions, therefore assessment of the predicted KEGG modules is not possible.

Table 7.9 summarises the assessment of the KEGG pathway predictions by comparing the KEGG pathways that are predicted to be effected and the KEGG pathways that contain compounds that are mapped to peaks that contribute to the statistically significant differences between the Carbaryl and Control 24h classes. Comparing the predicted KEGG pathways with the KEGG pathways linked to peaks identified using the volcano plot and the PLS-DA model results in precision values of 53.33% and 60.00%, recall values of 25% and 28.13% and F-Measures of 34.04% and 38.30% respectively.

The computationally generated predictions are also extended to areas of metabolism (see Figure 5.4). Each KEGG pathway belongs to a more general area of metabolism. The areas of metabolism that contain the predicted KEGG pathways are recorded as part of the computationally generated hypotheses (section 5.3.2) and are compared with the areas of metabolism that the KEGG pathways observed to be effected. Table 7.10 summarises the comparison. The areas of metabolism linked to the peaks identified using the volcano plot and the PLS-DA model both achieve precision values of 83.33%, recall values of 62.5% and F-Measures of 71.43%.

Overall, it is difficult to fully assess the performance of the computationally generated hypotheses. The measured features obtained from the metabolomics study has very high variance for all experimental classes, and only one of the possible eight inter-group comparisons yielded any statistically significant differences. The only possible inter-group comparison was for a 24hr time point. The intention of collecting metabolomics data at multiple time points was to track statistically significant changes in metabolic response across these time points, which was unfortunately not possible. As the transcriptomics data that generated the data used to form the hypotheses was taken at a

4hr time point, caution must be exercised when using only the 24hr time-point metabolomics data to assess the hypotheses.

A possible reason for the lack of significant metabolic responses across the various group comparisons is that the exposures did not induce strong enough metabolic responses. The doses of the Lead and Carbaryl exposures were based on the transcriptomics study (Orsini et al, 2016) which generated the data used to score the GWMR (Section 5.3), and were deemed severe enough to drive significant transcriptional response. (Orsini et al, 2018). The doses were sub-lethal as they aimed to reflect realistic human-induced pollution in inland waters, and this may have resulted in too weak of a metabolic response to be picked up in an untargeted study. A statistical power analysis could be performed to estimate the required sample size to detect a smaller effect size (Blaise et al, 2016) caused by such sub-lethal does.

The fact that only around 25% of the peaks with metabolite annotations can be linked to KEGG pathways also affects the ability to assess the performance thoroughly. The statistical analysis of two experimental classes with the lowest feature variance, Control and Carbaryl 24h, did allow for some assessment to be made in terms of predicted KEGG pathways and areas of metabolism and yielded some positive results. For the pathways, good precision was observed, albeit with lower recall values. For areas of metabolism precision and recall values were good, but this is a very low granularity way of looking at effects on a metabolic system. A more stable dataset in which the features have lower experimental class variation would allow for more robust statistical analysis, which is needed to make a fair assessment of the computationally generated hypotheses.

8. Discussion

The ultimate aim of the research presented in this thesis (see Chapter 2), is to develop, apply and validate a mechanism for using a computational environmental toxicology approach to make *in-silico* predictions of unknown metabolic response of a complex organism of interest to environmental stressors using transcriptomics data. Realising this would allow for toxicology measurements of gene expression to be used to reveal downstream effects on metabolism in a completely *in-silico* way. This hypotheses generation step is of benefit to untargeted metabolomics studies as it adds an element of hypotheses testing to the holistic approach of measuring every metabolite possible using untargeted metabolomics.

The approach used in this thesis differs to some of the traditional ones in that it aims to predict unknown organism response to environmental stressors, with a focus on organisms whose genome sequences are newly sequenced and therefore do not have accurate and curated GWMRs. Other studies have looked at using flux balance analysis with highly curated GWMRs to model known organism responses (Brandes et al, 2012; Colijn et al, 2009; Dreyfuss et al, 2013; Garcia Sanchez et al, 2012; Heavner et al, 2013; van Berlo et al, 2011). Other studies have largely been focussed on predicting cellular growth (Dreyfuss et al, 2013; Feist et al, 2009; Garcia Sanchez et al, 2012; Huang & Fraenkel, 2009; Lee et al, 2012), the production of a specific biologically important metabolite (Varma et al, 1993) or for the analysis of genomics and transcriptomics data (Li et al, 2010; Mols et al, 2007).

The organism of interest used to test this approach is *Daphnia*, an emerging model species in evaluating ecological impact of environmental change, and one that is increasingly

used as the organism of interest for environmental omics studies (section 1.8). The sub-species *D. magna* is used in this research due to the availability of a reference genome, and environmental toxicology transcriptomics resources (Orsini et al, 2016; Orsini et al, 2012).

In order to achieve the goal of this research, a workflow is developed that provides a mechanism for computational hypothesis generation of the metabolic effects of environmental insults using genome-wide metabolic reconstruction (GWMR) to model the metabolome of an organism and to generate computational hypotheses by incorporating transcriptomics data and using a network optimisation technique called active modules (Figure 2.1). GWMR is an *in-silico* modelling technique that aims to represent the metabolic capabilities of an organism at a genomic scale by representing a metabolome as a network of connected nodes. GWMRs provide a platform for analysis, visualisation and contextualisation of omics datasets. The potential for the use of GWMRs in environmental metabolomics based computational toxicology has been highlighted (Kesari, 2017) but to date little has been published in this area (Blais et al, 2017; Kotera & Goto, 2016; Topfer et al, 2015).

The workflow is applied to predict the effects of two environmental stressors on the metabolome of *D. magna*. The two stressors used, Carbaryl and Lead are relevant as they are human-induced because of human-driven pollution caused by agriculture and industry. The computationally generated hypotheses make predictions of the metabolic response of *D.* magna to the two stressors that is previously unknown. To assess the accuracy of these predictions and therefore assess the effectiveness of the proposed workflow, a metabolomics study is performed.

8.1. METRONOME platform

The first step in achieving the overall research goal was to generate a GWMR of *D. magna*. Draft GWMRs can be built in an automated way, and several tools are available for doing this. These tools have limitations when being used to construct draft GWMRs of organisms with newly sequenced genomes as they either require annotated genome sequences in very specific formats or require the organism's genome sequence to be present pathway databases such as KEGG. For *D. magna*, these resources are unavailable so to facilitate the construction of a draft GWMR of *D. magna* the METRONOME platform is developed (Chapter 3).

The METRONOME platform is flexible and lightweight and only requires a genome sequence to construct a draft GWMR. METRONOME makes use of an orthology matching algorithm to assign enzymes to the input sequence. A data mining module then uses these enzymes to extract enzymatic reactions and the associated metabolites to form a network. METRONOMEs flexible architecture allows for multiple databases / data sources to be used by the data mining module. A network merging module makes use of the reaction and metabolite reconciliation resource MetaNetX (Moretti et al, 2016) to merge reactions and metabolites obtained from different sources into a single coherent network. The ability to use multiple sources is a key advantage of the METRONOME platform, and sets it apart from other tools. To date, a sub-module for extracting metabolic reactions from KEGG based on enzyme numbers is implemented along with a sub-module for extracting reactions from an SBML (Systems Biology Mark-up Language). The SBML extraction module in this case is used to extract metabolic reactions from the MetaCyc (Caspi et al, 2014) resource which can be exported as an SBML file.

The performance of METRONOME is evaluated by generating draft GWMRs of two model organisms with highly curated GWMRs, *E. coli* and *S. cerevisiae.* The METRONOME generated GWMRs are then compared to draft GWMRs built using two other automated tools; Pathologic and Model SEED. The comparisons look at how well the draft GWMRs cover the reactions and metabolites contained within well curated GWMRs/database information about the species. METRONOME clearly outperformed Model SEED, and when compared to Pathologic, METRONOME GWMRs recovered a higher proportion of the curated set of reactions but were slightly less precise (section 3.3).

The Pathologic tool makes use of a rule-based pathway inference mechanism when constructing draft GWMR (Karp et al, 2011). This mechanism adds or removes reactions from the generated model if metabolic pathways are nearly complete or mostly incomplete. This offers an explanation as to why Pathologic networks are more precise than METRONOME networks. Network inference could be introduced into the METRONOME approach however when it comes to generating hypotheses using the network (Chapter 5), reaction nodes are scored based on gene expression data. This is possible because each reaction in the METRONOME draft GWMR can be tracked back to one or many genes from the input genome sequence (Figure 1.4). If a pathway inference technique is applied, some reactions will be added to the network based on the presence of other reactions in a pathway meaning that no genes will be directly associated with them. This presents a challenge when scoring the network prior to active module identification.

A key advantage of METRONOME is that the only thing that is required to generate a draft GWMR is an unannotated genome sequence. The enzyme assignment module of the

METRONOME platform essentially performs a genome annotation. Currently the OrthoMCL (Li et al, 2003) algorithm is configured for use with METRONOME. Alternative genome annotation approaches are available (Claudel-Renard et al, 2003; Curtis et al, 2013; Devoid et al, 2013; Li et al, 2003; Romero et al, 2005; Waterhouse et al, 2013; Zhao et al, 2013), and incorporating more of these has the potential to improve the accuracy of the METRONOME draft GWMRs. This can be achieved with relative ease thanks to METRONOME's modular design.

The METRONOME platform does not include subcellular compartmentalisation or the addition of transport reactions. Eukaryotic cells are made up of several membrane bound compartments, each containing different collections of metabolic enzymes. The various compartments are connected metabolically by transport reactions. A result of this is that certain metabolic processes are only possible within certain membrane bound cellular compartments (Klitgord & Segre, 2010). Compartmentalisation of reactions within GWMRs, and the addition of associated transport reactions helps to improve model accuracy and the predictive performance of flux balance analysis (Bekaert, 2012). The addition of compartments and transport reactions forms part of the manual curation stage outlined in the highly detailed protocol for GWMRs (Thiele & Palsson, 2010). The METRONOME platform would benefit from the addition of compartmentalisation, however the AMBIENT algorithm (Bryant et al, 2013b) used to generate the computational hypotheses in this thesis does not take into account the compartmentalisation of metabolic reactions.

8.2. Draft GWMR of *D. magna*

Once the effectiveness of METRONOME platform is evaluated, it is used to build a draft GWMR of *D. magna* (Chapter 4). There is no reported GWMR of *D. magna* so to assess

the reconstruction, the contents of the network are investigated by looking at the coverage of the KEGG reference pathway, three core KEGG pathways and thirteen KEGG pathways that have been highlighted in *D. magna* toxicology studies (section 4.3). Coverage of the core and highlighted KEGG pathways is satisfactory so the *D. magna* draft GWMR is considered acceptable. The generation of a first draft GWMR of *D. magna* has not been previously reported and now the process of manually curating the network based on experimental evidence can begin.

8.3. Computational hypothesis generation

The AMBIENT algorithm (Bryant et al, 2013b), an extension to the active modules approach (Ideker et al, 2002), uses a search heuristic to identify sub-networks, or 'hot-spots', within a metabolic reconstruction that are significantly affected based on a toxicogenomic transcriptomics dataset. These 'hot-spots' reveal areas within the metabolic network that are predicted to be significantly affected by the condition tested in the gene expression data. Chapter 5 details the use of AMBIENT to identify sub-modules within the *D. magna* using a transcriptomics dataset that measures the gene expression response to two environmental stressors, Carbaryl and Lead.

The reactions and metabolites contained within the identified sub-modules are passed to the KEGG mapper software to identify KEGG modules, KEGG pathways and areas of metabolism (see Figure 5.4) that form predictions of how the *D. magna* metabolome is effected at different granularities. These predictions are termed computationally generated hypotheses. This approach of identifying network 'hot-spots' is previously used for discovering underlying mechanisms of known responses (Bryant et al, 2013b; Wang et al, 2013; Wang et al, 2014) which do not require validation. Here the approach is being used in a hypothesis generation context, which can be used to inform untargeted

metabolomics studies and avoid the common criticism of fishing for interesting features within a dataset (Ning & Lo, 2010).

8.4. Closed-loop optimisation of LC-MS analysis

LC-MS is a principal analytical technique for metabolomics and is the analytical platform used in the study conducted to assess the performance of the computationally generated predictions (Chapter 7). Developing LC-MS methods is challenging, and can be formulated as a multi-objective heuristic search problem. Closed loop optimisations are where solutions are evaluated by conducting physical real-world experiments. Typically, these evaluations are costly and time consuming so algorithms that make efficient use of evaluations are required. The MUSCLE platform for closed-loop automated evolutionary multi-objective optimisation of LC-MS analysis (Chapter 6) is developed and used to develop a HPLC-MS method for use in the metabolomics study.

Two optimisation algorithms are implemented in MUSCLE. A modified version of the PESA-II algorithm that uses Latin Hypercube sampling to initially sample the search space is used. A new algorithm, PESA-II-FS, which extends the modified PESA-II to use feature selection is also introduced. PESA-II-FS iteratively uses feature selection to identify the decision variables that have the greatest effect on the multi-objective optimisation. The selected decision variables become the focus of the optimisation, whist the unselected variables have their values set based on values that have previously resulted in good solutions (section 6.2.4.2). This effectively reduces the search space and helps convergence of the algorithm. This is important as for closed-loop optimisations like this, as evaluations are extremely costly in terms of both time and cost.

A comparison of the PESA-II and PESA-II-FS algorithms is presented in section 6.3.1. The optimisation using the PESA-II-FS algorithm performed better than the original PESA-II MUSCLE algorithm, especially in terms of method sensitivity but further testing is required. The PESA-II and the PESA-II-FS algorithms share the first two steps (Figure 6.8 & Figure 6.9), both have an archive set initialised with non-dominated solutions obtained using Latin Hypercube Sampling. The initialisation on the Latin Hypercube is stochastic in nature, and this case, each optimisation used a different Latin Hypercube. In future studies, the same Latin Hypercube can be used for both optimisations to give a fairer comparison between the two algorithms.

MUSCLE is demonstrated by the automated development of several LC-MS method optimisations across a range of analytical systems for both targeted and untargeted analysis. MUSCLE optimisations always resulted in improved LC-MS methods with increased analytical sensitivity and/or shorter analysis times for a number of different analyses across a range of different analytical platforms and analysis types. This closed-loop approach has the potential to benefit many scientific fields that make use of LC-MS including metabolomics, proteomics and pharmacology.

8.5. Prediction validation

To assess the computationally generated predictions (Chapter 5) a metabolomics study is performed. The experimental design mimics the design of the study that produced the transcriptomics datasets used in the prediction process. Extra time points are included to account for the expected temporal delay between gene expression and metabolic changes. The experiment involved *D. magna* clones being exposed to environmentally relevant concentrations of Carbaryl and Lead in controlled laboratory conditions (section 7.1).

PCA, PLS-DA and volcano plots are used to identify statistically different peaks between treatment-control group pairs at each time point (section 7.3). The statistical analyses failed to identify significant difference between all but one of the group comparisons. This result is highly unexpected. To investigate, for each experimental group, the Relative Standard Deviation (RSD) values for each feature is calculated, Figure 7.19 shows boxplots of the distribution of these RSD values. Median RSD values are very high, with the average value across groups being 64%. This high amount of variation seen in the dataset offers an explanation as to why the statistical analyses did not identify significant differences.

Statistically significant differences between the Carbaryl and Control 24h groups were observed using a volcano plot and PLS-DA model. The features that contributed to the statistical differences are matched to MI-Pack (Weber & Viant, 2010) KEGG annotations and these KEGG annotations are used to identify KEGG modules, pathways and areas of metabolism that can be used to assess the accuracy of the predictions using an approach mirrors the approach used during the computational hypotheses generation stage (section 5.3.2). The identified KEGG modules, pathways and areas of metabolism are then compared to the Carbaryl predictions at the three levels of granularity (Figure 5.4).

Assigning annotated peaks to KEGG modules was problematic. Only 2-3% of annotated peaks could be mapped to KEGG modules. Assessing the predictions at this level is therefore meaningless. Comparing predictions to observations at the KEGG pathway level was more successful. Comparing the predicted KEGG pathways with the KEGG pathways linked to the peaks identified using the volcano plot and the PLS-DA model results in precision values of 53.33% and 60.00%, recall values of 25% and 28.13% and F-Measures of 34.04% and 38.30% respectively.

Comparing predictions at the areas of metabolism level is also successful. Comparing the predicted areas of metabolism with the areas of metabolism linked to the peaks identified using the volcano plot and the PLS-DA model both results in precision values of 83.33%, recall values of 62.5% and F-Measures of 71.43%.

Despite the challenges faced in processing the dataset, some positive results are obtained for the predictions of the Carbaryl treatment. At the KEGG pathway level, the predictions had good precision values but with poorer recall values and at the areas of metabolism level both precision and recall values are good. The areas of metabolism level is extremely broad however, so good precision and recall at this level does not necessarily indicate high quality predictions.

It is difficult to make a clear assessment of the quality of the predictions. High variance in the metabolomics dataset meant that only one of the eight possible comparisons is possible. This resulted in only a single statistically relevant comparison being made for one of the two treatments at one of the four time points. The high amount of variation did not allow for any assessment of the predictions for the Lead treatment at all. There is a clear need to test the computational hypotheses generation methodology using a different dataset as only the predicted metabolic effect of the Carbaryl treatment can be assessed. A targeted metabolomics approach may be more suitable for assessment of prediction accuracy.

8.6. Concluding remarks and future work

The aim of this research was to develop a framework for using GWMRs in an environmental computational toxicology setting to form computationally generated hypotheses of the metabolic effects of environmental insults. Using GMWRs in a

predictive way like this has potential benefits to untargeted metabolomics studies as it allows for an element of hypothesis testing to be introduced alongside the hypothesis-generating nature of untargeted metabolomics studies. Using such a computational approach can help in the design of untargeted metabolomics studies and go towards answering a major criticism of untargeted metabolomics, that it is often seen as fishing for hypotheses (Ning & Lo, 2010).

A workflow is introduced that allows for organisms with newly sequenced genomes to be used in this predictive way. Several computational tools and approaches are introduced to achieve this. *D. magna* is an important new model species in environmental research that has seen a lot of attention in environmental omics and toxicology studies. Here, the first reported GWMR of *D. magna* is introduced and used in a predictive way to predict unknown effects on its metabolome to two environmental stressors relevant to human-driven pollution. A metabolomics study is conducted to assess the computational predictions. Due to difficulties in the reproducibility of the dataset, it is difficult to fully assess the accuracy of the predictions. Despite this, some positive results are obtained, further experiments are needed however.

A major assumption of the approach used when generating the computational hypotheses using AMBIENT is that metabolite concentration levels are directly correlated to the transcript levels of the associated enzyme encoding genes. Without measuring the level of enzyme expression this cannot be known for sure. If an enzyme is more active, it suggests that the reactions associated with it are more active, meaning that the metabolic flux through said reaction would be expected to be affected. This does not necessarily mean that the abundance of the metabolites associated with these reactions should be present in higher or lower concentrations as the metabolites are interlinked within the

entire metabolic network through a series of reactions that both consume and produce them (Zelezniak et al, 2014).

The metabolites that are included in AMBIENT modules are selected based on their degree, or connectivity within the network. As a result, it could be hypothesised that their concentrations would be affected in some way by the conditions being tested in the transcriptomics data that produced the AMBIENT modules. A targeted metabolomics study could be carried out to focus specifically on the metabolites that are contained within the modules to investigate how their concentrations change over the time points in the study. This would mitigate the issues with metabolite annotation of untargeted metabolomics data and therefore also aid the statistical analysis. There is a danger in taking this approach however as there could be areas of the metabolome that are significantly changing under the conditions being tested that would not be picked up.

Fluxomics allows for the measurement of reaction rates within an organism (Winter & Kromer, 2013) and could be utilised in this context. A fluxomics study would shift the focus from the metabolites to the reactions contained within the AMBIENT active modules. As it is the reactions nodes in the GWMR that have experimental data directly linked to them this would be beneficial. A requirement for a fluxomics study however is that the metabolic reconstruction being used is of high accuracy and is well curated. This limits the applicability for this approach to be used with newly sequenced organisms such as *D. magna* where the required detailed biological knowledge is not necessarily available.

The fact that a draft GWMR of *D. magna* is used to generate the computational hypotheses likely to have caused uncertainties due to the lack of manual curation of the model. If the draft GWMR were to undergo the full manual curation steps outlined in

(Thiele & Palsson, 2010) there would be more confidence in the predicted metabolic responses. The approach to computationally predict unknown organism response to environmental stressors could be validated using an organism that has a highly curated GWMR such as Human (Swainston et al, 2016), Yeast (Heavner & Price, 2015) or Mouse (Sigurdsson et al, 2010).

References

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C, Gruning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* **44:** W3-w10

Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J (2012) Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol* **8:** e1002518

Allwood JW, Goodacre R (2010) An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochem Anal* **21:** 33-47

Altman T, Travers M, Kothari A, Caspi R, Karp PD (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics* **14:** 112

Altshuler I, Demiri B, Xu S, Constantin A, Yan ND, Cristescu ME (2011) An integrated multi-disciplinary approach for studying multiple stressors in freshwater ecosystems: Daphnia as a model organism. *Integr Comp Biol* **51:** 623-633

Amariei C, Tomita M, Murray DB (2014) Quantifying periodicity in omics data. *Front Cell Dev Biol* **2:** 40

Anderson NL (2010) The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem* **56:** 177-185

Ankley G, Miracle A, Perkins EJ, Daston GP (2007) *Genomics in regulatory ecotoxicology: Applications and challenges*: CRC Press.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil

LK, Paarmann D, Paczian T, Parrello B, Pusch GD et al (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75

Backes C, Rurainski A, Klau GW, Muller O, Stockel D, Gerasch A, Kuntzer J, Maisel D, Ludwig N, Hein M, Keller A, Burtscher H, Kaufmann M, Meese E, Lenhof HP (2012) An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Res* **40:** e43

Bairoch A (1994) The ENZYME data bank. *Nucleic Acids Research* **22:** 3626-3627

Banerjee S, Mazumdar S (2012) Electrospray ionization mass spectrometry: a technique to access the information beyond the molecular weight of the analyte. *Int J Anal Chem* **2012:** 282574

Beckonert O, Keun HC, Ebbels TM, Bundy J, Holmes E, Lindon JC, Nicholson JK (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc* **2:** 2692-2703

Bekaert M (2012) Reconstruction of Danio rerio metabolic model accounting for subcellular compartmentalisation. *PLoS One* **7:** e49903

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

Benton HP, Ivanisevic J, Mahieu NG, Kurczy ME, Johnson CH, Franco L, Rinehart D, Valentine E, Gowda H, Ubhi BK, Tautenhahn R, Gieschen A, Fields MW, Patti GJ, Siuzdak G (2015) Autonomous metabolomics for rapid metabolite identification in global profiling. *Anal Chem* **87:** 884-891

Benton HP, Want EJ, Ebbels TM (2010) Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data. *Bioinformatics* **26:** 2488-2489

Berger B, Peng J, Singh M (2013) Computational solutions for omics data. *Nat Rev Genet* **14:** 333-346

Bernard T, Bridge A, Morgat A, Moretti S, Xenarios I, Pagni M (2014) Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in Bioinformatics* **15:** 123-135

Blais EM, Rawls KD, Dougherty BV, Li ZI, Kolling GL, Ye P, Wallqvist A, Papin JA (2017) Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. *Nat Commun* **8:** 14250

Blaise BJ, Correia G, Tin A, Young JH, Vergnaud AC, Lewis M, Pearce JT, Elliott P, Nicholson JK, Holmes E, Ebbels TM (2016) Power Analysis and Sample Size Determination in Metabolic Phenotyping. *Anal Chem* **88:** 5179-5188

Blankenburg M, Haberland L, Elvers H-D, Tannert C, Jandrig B (2009) High-throughput omics technologies: potential tools for the investigation of influences of EMF on biological systems. *Current genomics* **10:** 86-92

Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. *Science* **277:** 1453-1462

Blum T, Kohlbacher O (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J Comput Biol* **15:** 565-576

Bradbury J, Genta-Jouve G, Allwood JW, Dunn WB, Goodacre R, Knowles JD, He S, Viant MR (2015) MUSCLE: automated multi-objective evolutionary optimization of targeted LC-MS/MS analysis. *Bioinformatics* **31:** 975-977

Brandes A, Lun DS, Ip K, Zucker J, Colijn C, Weiner B, Galagan JE (2012) Inferring carbon sources from gene expression profiles using metabolic flux models. *PLoS One* **7:** e36947

Brooks BW, Lazorchak JM, Howard MD, Johnson MV, Morton SL, Perkins DA, Reavie ED, Scott GI, Smith SA, Steevens JA (2016) Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environ Toxicol Chem* **35:** 6-13

Brown M, Wedge DC, Goodacre R, Kell DB, Baker PN, Kenny LC, Mamas MA, Neyses L, Dunn WB (2011) Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **27:** 1108-1112

Bryant WA, Faruqi AA, Pinney JW (2013a) Analysis of metabolic evolution in bacteria using whole-genome metabolic models. *J Comput Biol* **20:** 755-764

Bryant WA, Sternberg MJ, Pinney JW (2013b) AMBIENT: Active Modules for Bipartite Networks - using high-throughput transcriptomic data to dissect metabolic response. *BMC Systems Biology* **7:** 26

Bundy JG, Davey MP, Viant MR (2009) Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **5:** 3

Bunescu A, Garric J, Vollat B, Canet-Soulas E, Graveron-Demilly D, Fauvelle F (2010) In vivo proton HR-MAS NMR metabolic profile of the freshwater cladoceran Daphnia magna. *Mol Biosyst* **6:** 121-125

Cabusora L, Sutton E, Fulmer A, Forst CV (2005) Differential network expression during drug and stress response. *Bioinformatics* **21:** 2898-2905

Cacciatore S, Tenori L, Luchinat C, Bennett PR, MacIntyre DA (2017) KODAMA: an R package for knowledge discovery and data mining. *Bioinformatics* **33:** 621-623

Campos B, Garcia-Reyero N, Rivetti C, Escalon L, Habib T, Tauler R, Tsakovski S, Pina B, Barata C (2013) Identification of metabolic pathways in Daphnia magna explaining hormetic effects of selective serotonin reuptake inhibitors and 4-nonylphenol using transcriptomic and phenotypic responses. *Environ Sci Technol* **47:** 9434-9443

Carvalho GR, Hughes RN (1983) The effect of food availability, female culture ‐ density and photoperiod on ephippia production in Daphnia magna Straus (Crustacea: Cladocera). *Freshwater Biology* **13:** 37-46

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, Walk TC, Zhang P, Karp PD (2014) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* **36**

Cedergreen N (2014) Quantifying synergy: a systematic review of mixture toxicity studies within environmental toxicology. *PLoS One* **9:** e96580

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30:** 918-920

Chazalviel M, Frainay C, Poupin N, Vinson F, Merlet B, Gloaguen Y, Cottret L, Jourdan F (2017) MetExploreViz: web component for interactive metabolic network visualization. *Bioinformatics*

Cho K, Evans BS, Wood BM, Kumar R, Erb TJ, Warlick BP, Gerlt JA, Sweedler JV (2014) Integration of untargeted metabolomics with transcriptomics reveals active metabolic pathways. *Metabolomics* **2014**

Chowdhury SA, Koyuturk M (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. *Pac Symp Biocomput***:** 133-144

Christie K, Weng S, Balakrishnan R, Costanzo M, Dolinski K, Dwight S, Engel S, Feierbach B, Fisk D, Hirschman J, Hong E, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld C, Andrada R, Binkley G, Dong Q, Lane C et al (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic acids research* **32**

Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3:** 140

Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* **8:** 37-49

Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31:** 6633-6639

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T et al (2011) The ecoresponsive genome of Daphnia pulex. *Science* **331:** 555-561

Colbourne JK, Singan VR, Gilbert DG (2005) wFleaBase: the Daphnia genome database. *BMC Bioinformatics* **6:** 45

Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, Cheng TY, Moody DB, Murray M, Galagan JE (2009) Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol* **5:** e1000489

Consortium U (2015) UniProt: a hub for protein information. *Nucleic Acids Research* **43**

Corne DW, Jerram NR, Knowles JD, Oates MJ (2001) PESA-II: region-based selection in evolutionary multiobjective optimization. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pp 283-290.

Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, Mistrik R, Neumann S, Schymanski EL, Sumner LW (2014) Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **10:** 0

Creek DJ, Jankevics A, Burgess KE, Breitling R, Barrett MP (2012) IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data. *Bioinformatics* **28:** 1048-1049

Croes D, Couche F, Wodak SJ, van Helden J (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res* **33:** W326-330

Croes D, Couche F, Wodak SJ, van Helden J (2006) Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol* **356:** 222-236

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky V, Song H, Williams M, Birney E et al (2014) The Reactome pathway knowledgebase. *Nucleic Acids Research* **42**

Crowson A, Beardah MS (2001) Development of an LC/MS method for the trace analysis of hexamethylenetriperoxidediamine (HMTD).

Cuevas DA, Edirisinghe J, Henry CS, Overbeek R, O'Connell TG, Edwards RA (2016) From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model. *Front Microbiol* **7:** 907

Curtis DS, Phillips AR, Callister SJ, Conlan S, McCue LA (2013) SPOCS: software for predicting and visualizing orthology/paralogy relationships among genomes. *Bioinformatics* **29:** 2641-2642

Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics* **27:** i205-213

David RM, Dakic V, Williams TD, Winter MJ, Chipman JK (2011) Transcriptional responses in neonate and adult Daphnia magna in relation to relative susceptibility to genotoxicants. *Aquat Toxicol* **104:** 192-204

Davidson RL, Weber RJ, Liu H, Sharma-Oates A, Viant MR (2016) Galaxy-M: a Galaxy workflow for processing and analyzing direct infusion and liquid chromatography mass spectrometry-based metabolomics data. *Gigascience* **5:** 10

de Souza LP, Naake T, Tohge T, Fernie AR (2017) From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web-resources for mass spectral plant metabolomics. *Gigascience*

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **6:** 182-197

Deo RC, Hunter L, Lewis GD, Pare G, Vasan RS, Chasman D, Wang TJ, Gerszten RE, Roth FP (2010) Interpreting metabolomic profiles using unbiased pathway models. *PLoS Comput Biol* **6:** e1000692

Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C (2013) Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *Methods Mol Biol* **985:** 17-45

Di Guida R, Engel J, Allwood JW, Weber RJ, Jones MR, Sommer U, Viant MR, Dunn WB (2016) Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **12:** 93

Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24:** i223-231

Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, Lu C, Swainston N, Dunn WB, Fisher P, Hull D, Brown M, Oshota O, Stanford NJ, Kell DB, King RD, Oliver SG, Stevens RD, Mendes P (2010) Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol* **4:** 145

Doma S (1979) Ephippia of Daphnia magna straus — A technique for their mass production and quick revival | SpringerLink. *Hydrobiologia* **67:** 198-188

Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE (2013) Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM. *PLoS Comput Biol* **9:** e1003126

Dunn WB, Broadhurst D, Ellis DI, Brown M, Halsall A, O'Hagan S, Spasic I, Tseng A, Kell DB (2008) A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int J Epidemiol* **37 Suppl 1:** i23-30

Dunn WB, Erban A, Weber RJ, Creek DJ, Brown M, Breitling R, Hankemeier T, Goodacre R, Neumann S, Kopka J (2013) Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9:** 44-66

Durillo JJ, Nebro AJ (2011) jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software* **42:** 760-771

Eads BD, Andrews J, Colbourne JK (2008) Ecological genomics in Daphnia: stress responses and environmental sex determination. *Heredity (Edinb)* **100:** 184-190

Ebert D (2005) *Ecology, Epidemiology, and Evolution of Parasitism in Daphnia*: National Center for Biotechnology Information (US).

Ebert D (2011) A genome for the environment. *Science* **331:** 539-540

Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol* **4:** 133-140

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Vol. 96, pp 226-231.

Famili I, Forster J, Nielsen J, Palsson BO (2003) Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A* **100:** 13134-13139

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* **3:** 121

Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson B (2009) Reconstruction of Biochemical Networks in Microbial Organisms. *Nat Rev Microbiol* **7:** 129-143

Feist AM, Palsson BO (2010) The biomass objective function. *Curr Opin Microbiol* **13:** 344-349

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

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nat Biotechnol* **18:** 1157-1161

Forster M, Pick A, Raitner M, Schreiber F, Brandenburg FJ (2002) The system architecture of the BioPath system. *In Silico Biol* **2:** 415-426

Fortney K, Kotlyar M, Jurisica I (2010) Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. *Genome Biol* **11:** R13

Francke C, Siezen RJ, Teusink B (2005) Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol* **13:** 550-558

Frisch D, Morton PK, Chowdhury PR, Culver BW, Colbourne JK, Weider LJ, Jeyasingh PD (2014) A millennial-scale chronicle of evolutionary responses to cultural eutrophication in Daphnia. *Ecol Lett* **17:** 360-368

Furey A, Moriarty M, Bane V, Kinsella B, Lehane M (2013) Ion suppression; a critical review on causes, evaluation, prevention and applications. *Talanta* **115:** 104-122

Gancedo C, Serrano R (1989) Energy-yielding metabolism. *The yeasts* **3:** 205-259

Garcia Sanchez CE, Vargas Garcia CA, Torres Saez RG (2012) Predictive potential of flux balance analysis of Saccharomyces cerevisiae using as optimization function combinations of cell compartmental objectives. *PLoS One* **7:** e43006

Garcia-Reyero N, Poynton HC, Kennedy AJ, Guan X, Escalon BL, Chang B, Varshavsky J, Loguinov AV, Vulpe CD, Perkins EJ (2009) Biomarker discovery and transcriptomic responses in Daphnia magna exposed to munitions constituents. *Environ Sci Technol* **43:** 4188-4193

Garreta-Lara E, Campos B, Barata C, Lacorte S, Tauler R (2016) Metabolic profiling of Daphnia magna exposed to environmental stressors by GC-MS and chemometric tools. *Metabolomics* **12:** 1

Gassner AL, Weyermann C (2016) LC-MS method development and comparison of sampling materials for the analysis of organic gunshot residues. *Forensic Sci Int* **264:** 47-55

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* **5**

Giacomoni F, Le Corguille G, Monsoor M, Landi M, Pericard P, Petera M, Duperier C, Tremblay-Franco M, Martin JF, Jacob D, Goulitquer S, Thevenot EA, Caron C (2015) Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics* **31:** 1493-1495

Gilligan L, Bradbury J, Taylor A, He S, O'Neil D, Viant M, Foster P (2014) A novel UPLC-MS/MS method to extract and quantify sulphated and non-sulphated oestrogens automatically optimised using MUSCLE software.

Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11:** R86

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science* **274:** 546, 563-547

Gowda GA, Djukovic D (2014) Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods Mol Biol* **1198:** 3-12

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of machine learning research* **3:** 1157-1182

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* **11:** 10-18

Harrigan GG, Goodacre R (2012) *Metabolic profiling: its role in biomarker discovery and gene function analysis*: Springer Science & Business Media.

Harris KD, Bartlett NJ, Lloyd VK (2012) Daphnia as an emerging epigenetic model organism. *Genet Res Int* **2012:** 147892

Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, Muthukrishnan V, Owen G, Turner S, Williams M, Steinbeck C (2012) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* **41**

Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44:** D1214-1219

Haug K, Salek RM, Steinbeck C (2017) Global open data management in metabolomics. *Curr Opin Chem Biol* **36:** 58-63

Heavner BD, Price ND (2015) Comparative Analysis of Yeast Metabolic Network Models Highlights Progress, Opportunities for Metabolic Reconstruction. *PLoS Comput Biol* **11:** e1004530

Heavner BD, Smallbone K, Price ND, Walker LP (2013) Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)* **2013:** bat059

Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R (2009) Gene regulatory network inference: data integration in dynamic models-a review. *Biosystems* **96:** 86-103

Heller S, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **7:** 23

Hodson PV, Qureshi K, Noble CA, Akhtar P, Brown RS (2007) Inhibition of CYP1A enzymes by alpha-naphthoflavone causes both synergism and antagonism of retene toxicity to rainbow trout (Oncorhynchus mykiss). *Aquat Toxicol* **81:** 275-285

Holme P (2009) Model validation of simple-graph representations of metabolism. *J R Soc Interface* **6:** 1027-1034

Hop CE, Chen Y, Yu LJ (2005) Uniformity of ionization response of structurally diverse analytes using a chip-based nanoelectrospray ionization source. *Rapid Commun Mass Spectrom* **19:** 3139-3142

Horgan RP, Kenny LC (2011) 'Omic'technologies: genomics, transcriptomics, proteomics and metabolomics. *The Obstetrician & Gynaecologist* **13:** 189-195

Huang SS, Fraenkel E (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal* **2:** ra40

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr J-H, Hunter PJ et al (2003) The systems biology markup language (SBML): a medium forrepresentation and exchange of biochemical network models. *Bioinformatics* **19:** 524-531

Hwang T, Park T (2009) Identification of differentially expressed subnetworks based on multivariate ANOVA. *BMC Bioinformatics* **10:** 128

Iampolskii LI, Galimov Ia R (2005) Evolutionary genetics of aging in Daphnia. *Zh Obshch Biol* **66:** 416-424

Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1:** S233-240

Institute SR. (2017) XCMS Online.

James G, Witten D, Hastie T, Tibshirani R (2013) *An introduction to statistical learning*, Vol. 112: Springer.

Jansen M, Coors A, Vanoverbeke J, Schepens M, De Voogt P, De Schamphelaere KA, De Meester L (2015) Experimental evolution reveals high insecticide tolerance in Daphnia inhabiting farmland ponds. *Evol Appl* **8:** 442-453

Jansen M, De Meester L, Cielen A, Buser CC, Stoks R (2011) The interplay of past and current stress exposure on the water flea Daphnia. *Funct Ecol* **25:** 974-982

Jenkinson C, Bradbury J, Taylor A, Adams JS, He S, Viant MR, Hewison M (2017) Automated development of an LC-MS/MS method for measuring multiple vitamin D metabolites using MUSCLE software. *Analytical Methods* **9:** 2723-2731

Jenkinson C, Bradbury J, Taylor A, He S, Viant M, Hewison M (2016) Three minute run time LC-MS/MS method for separation and quantifying 25-hydroxyvitamin D from C3-epimers.

Jolliffe IT (2002) Principal component analysis and factor analysis. *Principal component analysis***:** 150-166

Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, Nainala VC, Spicer RA, Williams M, Li X, Salek RM, Griffin JL, Steinbeck C (2016) MetaboLights: An Open-Access Database Repository for Metabolomics Data. *Curr Protoc Bioinformatics* **53:** 14.13.11-18

Kanehisa M (2013) Molecular network analysis of diseases and drugs in KEGG. *Methods Mol Biol* **939:** 263-275

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* **42**

Kanehisa M, Institute for Chemical Research KU, Uji, Kyoto 611-0011, Japan, Goto S, Institute for Chemical Research KU, Uji, Kyoto 611-0011, Japan (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28:** 27-30

Karaman I, Ferreira DL, Boulangé CL, Kaluarachchi MR, Herrington D, Dona AC, Castagné R, Moayyeri A, Lehne B, Loh M (2016) Workflow for Integrated Processing of Multicohort Untargeted 1H NMR Metabolomics Data in Large-Scale Metabolic Epidemiology. *Journal of proteome research* **15:** 4188-4194

Karp P, Latendresse M, Paley S, Krummenacker M, Ong Q, Billington R, Kothari A, Weaver D, Lee T, Subhraveti P, Spaulding A, Fulcher C, Keseler I, Caspi R (2016) Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* **17:** 877-890

Karp PD, Latendresse M, Caspi R (2011) The pathway tools pathway prediction algorithm. *Stand Genomic Sci* **5:** 424-429

Kavlock RJ, Austin CP, Tice RR (2009) Toxicity testing in the 21st century: implications for human health risk assessment. *Risk Anal* **29:** 485-487; discussion 492-487

Kesari KK. (2017) Perspectives in Environmental Toxicology. Springer.

Keseler I, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus R, Johnson A, Krummenacker M, Nolan L, Paley S, Paulsen I, Peralta-Gil M, Santos-Zavaleta A, Shearer A, Karp P (2009) EcoCyc: A comprehensive view of Escherichia coli biology. *Nucleic Acids Research* **37:** D464-D470

Keun HC (2006) Metabonomic modeling of drug toxicity. *Pharmacol Ther* **109:** 92-106

Kim J, Reed JL (2012) RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol* **13:** R78

Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH (2016) PubChem Substance and Compound databases. *Nucleic Acids Res* **44:** D1202-1213

Kirlik G, Sayın S (2014) A new algorithm for generating all nondominated solutions of multiobjective discrete optimization problems. *European Journal of Operational Research* **232:** 479-488

Kitano H (2002) Systems Biology: A Brief Overview.

Klammer M, Godl K, Tebbe A, Schaab C (2010) Identifying differentially regulated subnetworks from phosphoproteomic data. *BMC Bioinformatics* **11:** 351

Klitgord N, Segre D (2010) The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Inform* **22:** 41-55

Knowles J (2006) ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation* **10:** 50-66

Knowles J (2009) Closed-loop evolutionary multiobjective optimization. *IEEE Computational Intelligence Magazine* **4:** 77-91

Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artificial intelligence* **97:** 273-324

Koster RA, Alffenaar JW, Greijdanus B, VanDerNagel JE, Uges DR (2014) Application of sweat patch screening for 16 drugs and metabolites using a fast and highly selective LC-MS/MS method. *Ther Drug Monit* **36:** 35-45

Kotera M, Goto S (2016) Metabolic pathway reconstruction strategies for central metabolism and natural product biosynthesis. *Biophys Physicobiol* **13:** 195-205

Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **84:** 283-289

Kurczy ME, Ivanisevic J, Johnson CH, Uritboonthai W, Hoang L, Fang M, Hicks M, Aldebot A, Rinehart D, Mellander LJ, Tautenhahn R, Patti GJ, Spilker ME, Benton HP, Siuzdak G (2015) Determining conserved metabolic biomarkers from a million database queries. *Bioinformatics* **31:** 3721-3724

Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Ogiwara A, Meissen J, Showalter M, Takeuchi K, Kind T, Beal P, Arita M, Fiehn O (2018) Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* **15:** 53-56

Lampert W, Kinne O (2011) *Daphnia: development of a model organism in ecology and evolution*, Vol. 21: International Ecology Institute Oldendorf/Luhe.

Laszlo E (1996) The Systems View of the World a Holistic Vision for Our Time.

Le TH, Lim ES, Hong NH, Lee SK, Shim YS, Hwang JR, Kim YH, Min J (2013) Proteomic analysis in Daphnia magna exposed to As(III), As(V) and Cd heavy metals and their binary mixtures for screening potential biomarkers. *Chemosphere* **93:** 2341-2348

Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, Kell DB, Mendes P, Swainston N (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Syst Biol* **6:** 73

Lemetre C, Zhang Q, Zhang ZD (2013) SubNet: a Java application for subnetwork extraction. *Bioinformatics* **29:** 2509-2511

Lenz EM, Wilson ID (2007) Analytical strategies in metabonomics. *J Proteome Res* **6:** 443-458

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13:** 2178-2189

Li S, Pozhitkov A, Ryan RA, Manning CS, Brown-Peterson N, Brouwer M (2010) Constructing a fish metabolic network model. *Genome Biol* **11:** R115

Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, Neumann S, Trausinger G, Sinner F, Pieber T, Magnes C (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16:** 118

Lindon JC, Nicholson JK, Holmes E (2011) *The handbook of metabonomics and metabolomics*: Elsevier.

Ma H, Zeng AP (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19:** 270-277

Martens L, Chambers M, Sturm M, Kessner D, Levander F, Shofstahl J, Tang WH, Rompp A, Neumann S, Pizarro AD, Montecchi-Palazzi L, Tasman N, Coleman M, Reisinger F, Souda P, Hermjakob H, Binz PA, Deutsch EW (2011) mzML--a community standard for mass spectrometry data. *Mol Cell Proteomics* **10:** R110.000133

Martins J, Oliva Teles L, Vasconcelos V (2007) Assays with Daphnia magna and Danio rerio as alert systems in aquatic toxicology. *Environment International* **33:** 414-425

Marx V (2013) Biology: The big challenges of big data. *Nature* **498:** 255-260

McCarty LS (2012) Model validation in aquatic toxicity testing: implications for regulatory practice. *Regul Toxicol Pharmacol* **63:** 353-362

McCarty LS (2013) Are we in the dark ages of environmental toxicology? *Regul Toxicol Pharmacol* **67:** 321-324

McCarty LS, Arnot JA, Mackay D (2013) Evaluation of critical body residue data for acute narcosis in aquatic organisms. *Environ Toxicol Chem* **32:** 2301-2314

McClure RS, Overall CC, McDermott JE, Hill EA, Markillie LM, McCue LA, Taylor RC, Ludwig M, Bryant DA, Beliaev AS (2016) Network analysis of transcriptomics expands regulatory landscapes in Synechococcus sp. PCC 7002. *Nucleic Acids Res* **44:** 8810-8825

McKay MD, Beckman RJ, Conover WJ (2000) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **42:** 55-61

Menni C, Zierer J, Valdes AM, Spector TD (2017) Mixing omics: combining genetics and metabolomics to study rheumatic diseases. *Nature Reviews Rheumatology* **13:** 174-181

Meurant G (2012) *Cellular energy metabolism and its regulation*: Elsevier.

Meyer VR (2013) *Pitfalls and Errors of HPLC in Pictures*: John Wiley & Sons.

Miner BE, De Meester L, Pfrender ME, Lampert W, Hairston NG, Jr. (2012) Linking genes to communities and ecosystems: Daphnia as an ecogenomic model. *Proc Biol Sci* **279:** 1873-1882

Mitra K, Carvunis AR, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* **14:** 719-732

Mo ML, Palsson BO, Herrgard MJ (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* **3:** 37

Mols M, de Been M, Zwietering MH, Moezelaar R, Abee T (2007) Metabolic capacity of Bacillus cereus strains ATCC 14579 and ATCC 10987 interlinked with comparative genomics. *Environ Microbiol* **9:** 2933-2944

Moreno-Sanchez R, Saavedra E, Rodriguez-Enriquez S, Olin-Sandoval V (2008) Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways. *J Biomed Biotechnol* **2008:** 597913

Moretti S, Martin O, Bridge A, Morgat A, Pagni M (2016) MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Research* **44**

Morgat A, Axelsen KB, Lombardot T, Alcántara R, Aimo L, Zerara M, Niknejad A, Belda E, Hyka-Nouspikel N, Coudert E, Redaschi N, Bougueleret L, Steinbeck C, Xenarios I, Bridge A (2015) Updates in Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Research* **43**

Morrison N, Bearden D, Bundy JG, Collette T, Currie F, Davey MP, Haigh NS, Hancock D, Jones OA, Rochfort S (2007) Standard reporting requirements for biological samples in metabolomics experiments: environmental context. *Metabolomics* **3:** 203-210

Nacu S, Critchley-Thorne R, Lee P, Holmes S (2007) Gene expression network analysis and applications to immunology. *Bioinformatics* **23:** 850-858

Nagato EG, D'Eon J C, Lankadurai BP, Poirier DG, Reiner EJ, Simpson AJ, Simpson MJ (2013) (1)H NMR-based metabolomics investigation of Daphnia magna responses to sub-lethal exposure to arsenic, copper and lithium. *Chemosphere* **93:** 331-337

Nguyen N, Huang H, Oraintara S, Vo A (2010) Mass spectrometry data processing using zero-crossing lines in multi-scale of Gaussian derivative wavelet. *Bioinformatics* **26:** i659-665

Ning M, Lo EH (2010) Opportunities and challenges in omics. *Transl Stroke Res* **1:** 233-237

O'Hagan S, Dunn WB, Brown M, Knowles JD, Kell DB (2005) Closed-loop, multiobjective optimization of analytical instrumentation: gas chromatography/time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal Chem* **77:** 290-303

O'Hagan S, Dunn WB, Knowles JD, Broadhurst D, Williams R, Ashworth JJ, Cameron M, Kell DB (2007) Closed-loop, multiobjective optimization of two-dimensional gas chromatography/mass spectrometry for serum metabolomics. *Anal Chem* **79:** 464-476

Oberhardt MA, Palsson B, Papin JA (2009) Applications of genome-scale metabolic reconstructions. In *Mol Syst Biol* Vol. 5, p 320.

Offem BO, Ayotunde EO (2008) Toxicity of lead to freshwater invertebrates (Water fleas; Daphnia magna and Cyclop sp) in fish ponds in a tropical floodplain. *Water, air, and soil pollution* **192:** 39-46

Orsini L, Brown JB, Shams Solari O, Li D, He S, Podicheti R, Stoiber MH, Spanier KI, Gilbert D, Jansen M, Rusch DB, Pfrender ME, Colbourne JK, Frilander MJ, Kvist J, Decaestecker E, De Schamphelaere KAC, De Meester L (2018) Early transcriptional response pathways in Daphnia magna are coordinated in networks of crustacean-specific genes. *Mol Ecol* **27:** 886-897

Orsini L, Gilbert D, Podicheti R, Jansen M, Brown JB, Solari OS, Spanier KI, Colbourne JK, Rusch DB, Decaestecker E, Asselman J, Schamphelaere KACD, Ebert D, Haag CR, Kvist J, Laforsch C, Petrusek A, Beckerman AP, Little TJ, Chaturvedi A et al (2016) Daphnia magna transcriptome by RNA-Seq across 12 environmental stressors. *Scientific Data, Published online: 10 May 2016; | doi:101038/sdata201630*

Orsini L, Spanier KI, L DEM (2012) Genomic signature of natural and anthropogenic stress in wild populations of the waterflea Daphnia magna: validation in space, time and experimental evolution. *Mol Ecol* **21:** 2160-2175

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* **7:** 535

Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* **28:** 245-248

Otte KA, Frohlich T, Arnold GJ, Laforsch C (2014) Proteomic analysis of Daphnia magna hints at molecular pathways involved in defensive plastic responses. *BMC Genomics* **15:** 306

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L et al (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33:** 5691-5702

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**

Parsons HM, Ludwig C, Günther UL, Viant MR (2007) Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics* **8:** 234

Pestana JL, Loureiro S, Baird DJ, Soares AM (2010) Pesticide exposure and inducible antipredator responses in the zooplankton grazer, Daphnia magna Straus. *Chemosphere* **78:** 241-248

Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA (2002a) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359:** 572-577

Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA (2002b) Clinical proteomics: translating benchside promise into bedside reality. *Nat Rev Drug Discov* **1:** 683-695

Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res* **33:** 1399-1409

Pitt JJ (2009) Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry. In *Clin Biochem Rev* Vol. 30, pp 19-34.

Pivato A, Vanin S, Raga R, Lavagnolo MC, Barausse A, Rieple A, Laurent A, Cossu R (2016) Use of digestate from a decentralized on-farm biogas plant as fertilizer in soils: an ecotoxicological study for future indicators in risk and life cycle assessment. *Waste management* **49:** 378-389

Popovska-Gorevski M, Dubocovich ML, Rajnarayanan RV (2017) Carbamate Insecticides Target Human Melatonin Receptors. *Chem Res Toxicol* **30:** 574-582

Poynton HC, Lazorchak JM, Impellitteri CA, Blalock BJ, Rogers K, Allen HJ, Loguinov A, Heckman JL, Govindasmawy S (2012) Toxicogenomic responses of nanotoxicity in Daphnia magna exposed to silver nitrate and coated silver nanoparticles. *Environ Sci Technol* **46:** 6288-6296

Poynton HC, Taylor NS, Hicks J, Colson K, Chan S, Clark C, Scanlan L, Loguinov AV, Vulpe C, Viant MR (2011) Metabolomics of microliter hemolymph samples enables an improved understanding of the combined metabolic and transcriptional responses of Daphnia magna to cadmium. *Environ Sci Technol* **45:** 3710-3717

Qiu YQ, Zhang S, Zhang XS, Chen L (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinformatics* **11:** 26

Qu RJ, Wang XH, Feng MB, Li Y, Liu HX, Wang LS, Wang ZY (2013) The toxicity of cadmium to three aquatic organisms (Photobacterium phosphoreum, Daphnia magna and Carassius auratus) under different pH levels. *Ecotoxicol Environ Saf* **95:** 83-90

Rainville LC, Carolan D, Varela AC, Doyle H, Sheehan D (2014) Proteomic evaluation of citrate-coated silver nanoparticles toxicity in Daphnia magna. *Analyst* **139:** 1678-1686

Rajagopalan D, Agarwal P (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* **21:** 788-793

Ramakrishna R, Edwards JS, McCulloch A, Palsson BO (2001) Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *Am J Physiol Regul Integr Comp Physiol* **280:** R695-704

Ransohoff DF (2005) Lessons from controversy: ovarian cancer screening and serum proteomics. *J Natl Cancer Inst* **97:** 315-319

Reisfeld B, Mayeno AN (2012) What is computational toxicology? *Methods Mol Biol* **929:** 3-7

Rivetti C, Campos B, Faria M, De Castro Catala N, Malik A, Munoz I, Tauler R, Soares AM, Osorio V, Perez S, Gorga M, Petrovic M, Mastroianni N, de Alda ML, Masia A, Campo J, Pico Y, Guasc H, Barcelo D, Barata C (2015) Transcriptomic, biochemical and individual markers in transplanted Daphnia magna to characterize impacts in the field. *Sci Total Environ* **503-504:** 200-212

Roberts LD, Souza AL, Gerszten RE, Clish CB (2012) Targeted Metabolomics. *Curr Protoc Mol Biol* **CHAPTER:** Unit30 32

Robertson DG (2005) Metabonomics in toxicology: a review. *Toxicol Sci* **85:** 809-822

Rocca-Serra P, Salek RM, Arita M, Correa E, Dayalan S, Gonzalez-Beltran A, Ebbels T, Goodacre R, Hastings J, Haug K, Koulman A, Nikolski M, Oresic M, Sansone SA, Schober D, Smith J, Steinbeck C, Viant MR, Neumann S (2016) Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **12:** 14

Rohrs HW (2006) LC/MS: A Practical User's Guide. *Journal of the American Society for Mass Spectrometry* **17:** 1193-1193

Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6:** R2

Rusyn I, Daston GP (2010) Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ Health Perspect* **118:** 1047-1050

Salek RM, Neumann S, Schober D, Hummel J, Billiau K, Kopka J, Correa E, Reijmers T, Rosato A, Tenori L, Turano P, Marin S, Deborde C, Jacob D, Rolin D, Dartigues B, Conesa P, Haug K, Rocca-Serra P, O'Hagan S et al (2015) COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access. *Metabolomics* **11:** 1587-1597

Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* **11:** 213

Scheltema RA, Jankevics A, Jansen RC, Swertz MA, Breitling R (2011) PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis. *Anal Chem* **83:** 2786-2793

Scott J, Ideker T, Karp RM, Sharan R (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol* **13:** 133-144

Scott RT, Jr., Treff NR (2010) Assessing the reproductive competence of individual embryos: a proposal for the validation of new "-omics" technologies. *Fertil Steril* **94:** 791-794

Segal E, Wang H, Koller D (2003) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19 Suppl 1:** i264-271

Segre D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* **99:** 15112-15117

Selivanov VA, Benito A, Miranda A, Aguilar E, Polat IH, Centelles JJ, Jayaraman A, Lee PW, Marin S, Cascante M (2017) MIDcor, an R-program for deciphering mass interferences in mass spectra of metabolites enriched in stable isotopes. *BMC Bioinformatics* **18:** 88

Shaffer RE (2002) Multi‐and Megavariate Data Analysis. Principles and Applications, I. Eriksson, E. Johansson, N. Kettaneh‐Wold and S. Wold, Umetrics Academy, Umeå, 2001, ISBN 91‐973730‐1‐X, 533pp. *Journal of Chemometrics* **16:** 261-262

Shaw J, Pfrender M, Eads B, Klaper R, Callaghan A, Sibly R, Colson I, Jansen B, Gilbert D, Colbourne J (2008) Daphnia as an emerging model for toxicological genomics. In

Comparative Toxicogenomics Vol. 2, Advances in Experimental Biology, pp 165-328. Elsevier

Shumway RH, Stoffer DS (2010) *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*: Springer.

Sigurdsson MI, Jamshidi N, Steingrimsson E, Thiele I, Palsson BO (2010) A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst Biol* **4:** 140

Small BG, McColl BW, Allmendinger R, Pahle J, López-Castejón G, Rothwell NJ, Knowles J, Mendes P, Brough D, Kell DB (2011) Efficient discovery of anti-inflammatory small-molecule combinations using evolutionary computing. *Nature Chemical Biology* **7:** 902-908

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

Smith TC, Frank E (2016) Introducing Machine Learning Concepts with WEKA. *Methods Mol Biol* **1418:** 353-378

Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27:** 431-432

Sohler F, Hanisch D, Zimmer R (2004) New methods for joint analysis of biological networks and expression data. *Bioinformatics* **20:** 1517-1521

Stanstrup J, Gerlich M, Dragsted LO, Neumann S (2013) Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. *Anal Bioanal Chem* **405:** 5037-5048

Steuer R, Morgenthal K, Weckwerth W, Selbig J (2007) A gentle guide to the analysis of metabolomic data. *Methods Mol Biol* **358:** 105-126

Stollewerk A (2010) The water flea Daphnia--a 'new' model system for ecology and evolution? *J Biol* **9:** 21

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis:

a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102:** 15545-15550

Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Jr., Murphy RC, Raetz CR, Russell DW, Subramaniam S (2007) LMSD: LIPID MAPS structure database. *Nucleic Acids Res* **35:** D527-532

Sud M, Fahy E, Cotter D, Dennis EA, Subramaniam S (2012) LIPID MAPS-Nature Lipidomics Gateway: An Online Resource for Students and Educators Interested in Lipids. *J Chem Educ* **89:** 291-292

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3:** 211-221

Sumpter JP, Jobling S (2013) The occurrence, causes, and consequences of estrogens in the aquatic environment. *Environ Toxicol Chem* **32:** 249-251

Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, Zielinski DC, Ang KS, Gardiner NJ, Gutierrez JM, Kyriakopoulos S, Lakshmanan M, Li S, Liu JK, Martinez VS, Orellana CA, Quek LE, Thomas A, Zanghellini J, Borth N et al (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12:** 109

Swainston N, Smallbone K, Mendes P, Kell D, Paton N (2011) The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform* **8:** 186

Szymanska E, Saccenti E, Smilde AK, Westerhuis JA (2012) Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **8:** 3-16

Tautenhahn R, Böttcher C, Neumann S (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **9:** 504

Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G (2012) An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol* **30:** 826-828

Taylor A, Bancos I, Chortis V, Lang K, O'Neil D, Hughes B, Jenkinson C, Deeks J, Shackleton C, Biehl M (2015) Further advances in diagnosis of adrenal cancer: a high-throughput urinary steroid profiling method using liquid chromatography tandem mass spectrometry (LC-MS/MS).

Taylor N, Weber R, Southam A, Payne T, Hrydziuszko O, Arvanitis T, Viant M (2008) A new approach to toxicity testing in Daphnia magna : application of high throughput FT-ICR mass spectrometry metabolomics. *Metabolomics, doi:101007/s11306-008-0133-3*

Taylor NS, Weber RJ, White TA, Viant MR (2010) Discriminating between different acute chemical toxicities via changes in the daphnid metabolome. *Toxicol Sci* **118:** 307-317

Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* **5:** 93-121

Topfer N, Kleessen S, Nikoloski Z (2015) Integration of metabolomics data into metabolic networks. *Front Plant Sci* **6:** 49

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* **17:** 520-525

Tsui MT, Wang WX (2006) Acute toxicity of mercury to Daphnia magna under different conditions. *Environ Sci Technol* **40:** 4025-4030

Unger S, Li W, Flarakos J, Tse FL, Patel S, Huang QM, Jian W, Edom R, Weng N, Bansal SK (2013) *Handbook of LC-MS Bioanalysis: Best Practices, Experimental Protocols, and Regulations*: John Wiley & Sons.

Valavanidis A, Vlahogianni T, Dassenakis M, Scoullos M (2006) Molecular biomarkers of oxidative stress in aquatic organisms in relation to toxic environmental pollutants. *Ecotoxicol Environ Saf* **64:** 178-189

van Berlo RJ, de Ridder D, Daran JM, Daran-Lapujade PA, Teusink B, Reinders MJ (2011) Predicting metabolic fluxes using gene expression differences as constraints. *IEEE/ACM Trans Comput Biol Bioinform* **8:** 206-216

Varma A, Boesch BW, Palsson BO (1993) Biochemical production capabilities of Escherichia coli. *Biotechnol Bioeng* **42:** 59-73

Vatansever B, Munoz A, Klein CL, Reinert K (2017) Development and optimisation of a generic micro LC-ESI-MS method for the qualitative and quantitative determination of

30-mer toxic gliadin peptides in wheat flour for food analysis. *Anal Bioanal Chem* **409:** 989-997

Vazquez A (2010) Frontiers in NeuroscienceProtein Interaction Networks. In *Neuroproteomics*, Alzate O (ed). Boca Raton (FL): CRC Press/Taylor & FrancisLlc.

Viant MR (2008) Recent developments in environmental metabolomics. *Mol Biosyst* **4:** 980-986

Viant MR, Bearden DW, Bundy JG, Burton IW, Collette TW, Ekman DR, Ezernieks V, Karakach TK, Lin CY, Rochfort S, de Ropp JS, Teng Q, Tjeerdema RS, Walter JA, Wu H (2009) International NMR-based environmental metabolomics intercomparison exercise. *Environ Sci Technol* **43:** 219-225

Viant MR, Sommer U (2013) Mass spectrometry based environmental metabolomics: a primer and review. *Metabolomics* **9:** 144-158

Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O (2012) A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. In *Metabolites* Vol. 2, pp 775-795.

Vinayavekhin N, Saghatelian A (2010) Untargeted metabolomics. *Curr Protoc Mol Biol* **Chapter 30:** Unit 30.31.31-24

Vo TD, Greenberg HJ, Palsson BO (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J Biol Chem* **279:** 39532-39540

Wang J, Chen L, Tian X, Gao L, Niu X, Shi M, Zhang W (2013) Global metabolomic and network analysis of Escherichia coli responses to exogenous biofuels. *J Proteome Res* **12:** 5302-5312

Wang Y, Eddy JA, Price ND (2012) Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol* **6:** 153

Wang Y, Ludwig C, Lilley DM, Gunther UL, Clore M, Campbell S, Han X, Neidle S, Postle T, Topliff C (2007) *Metabolomics, metabonomics and metabolite profiling*: Royal Society of Chemistry.

Wang Y, Shi M, Niu X, Zhang X, Gao L, Chen L, Wang J, Zhang W (2014) Metabolomic basis of laboratory evolution of butanol tolerance in photosynthetic Synechocystis sp. PCC 6803. *Microb Cell Fact* **13:** 151

Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV (2013) OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* **41:** D358-365

Weber RJ, Viant MR (2010) MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. *Chemometrics and Intelligent Laboratory Systems* **104:** 75-82

Weber RJM, Lawson TN, Salek RM, Ebbels TMD, Glen RC, Goodacre R, Griffin JL, Haug K, Koulman A, Moreno P, Ralser M, Steinbeck C, Dunn WB, Viant MR (2017) Computational tools and workflows in metabolomics: An international survey highlights the opportunity for harmonisation through Galaxy. *Metabolomics* **13:** 12

Weinberg GM (2011) *An introduction to general systems thinking*: New York: Wiley.

Weininger D (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci* **28:** 31-36

Widmer L, Watson S, Schlatter K, Crowson A (2002) Development of an LC/MS method for the trace analysis of triacetone triperoxide (TATP). *Analyst* **127:** 1627-1632

Winter G, Kromer JO (2013) Fluxomics - connecting 'omics analysis and phenotypes. *Environ Microbiol* **15:** 1901-1916

Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F et al (2012) HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* **41**

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems* **58:** 109-130

Worley B, Powers R (2013) Multivariate Analysis in Metabolomics. *Curr Metabolomics* **1:** 92-107

Wrzodek C, Drager A, Zell A (2011) KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. *Bioinformatics* **27:** 2314-2315

Xia J, Sinelnikov IV, Han B, Wishart DS (2015) MetaboAnalyst 3.0--making metabolomics more meaningful. *Nucleic Acids Res* **43:** W251-257

Yeh T, Chang T-H, Miller RC (2009) Sikuli: using GUI screenshots for search and automation. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pp 183-192.

Zelena E, Dunn WB, Broadhurst D, Francis-McIntyre S, Carroll KM, Begley P, O'Hagan S, Knowles JD, Halsall A, Wilson ID, Kell DB, Consortium H (2009) Development of a Robust and Repeatable UPLC-MS Method for the Long-Term Metabolomic Study of Human Serum. *Analytical Chemistry* **81:** 1357-1364

Zelezniak A, Sheridan S, Patil KR (2014) Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. *PLoS Comput Biol* **10:** e1003572

Zhang J, Gonzalez E, Hestilow T, Haskins W, Huang Y (2009) Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr Genomics* **10:** 388-401

Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, Bonanno JB, Hillerich BS, Seidel RD, Babbitt PC, Almo SC, Sweedler JV, Gerlt JA, Cronan JE, Jacobson MP (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* **502:** 698-702

Zhao XM, Wang RS, Chen L, Aihara K (2008) Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res* **36:** e48

Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS (2003) Detection of cancer-specific markers amid massive mass spectral data. *Proc Natl Acad Sci U S A* **100:** 14666-14671

Zonaras V, Alexis M, Koupparis M (2016) Development and validation of an LC–MS method for the simultaneous determination of sulfadiazine, trimethoprim, and N4-acetyl-sulfadiazine in muscle plus skin of cultured fish. http://dxdoiorg/101080/1082607620161169425

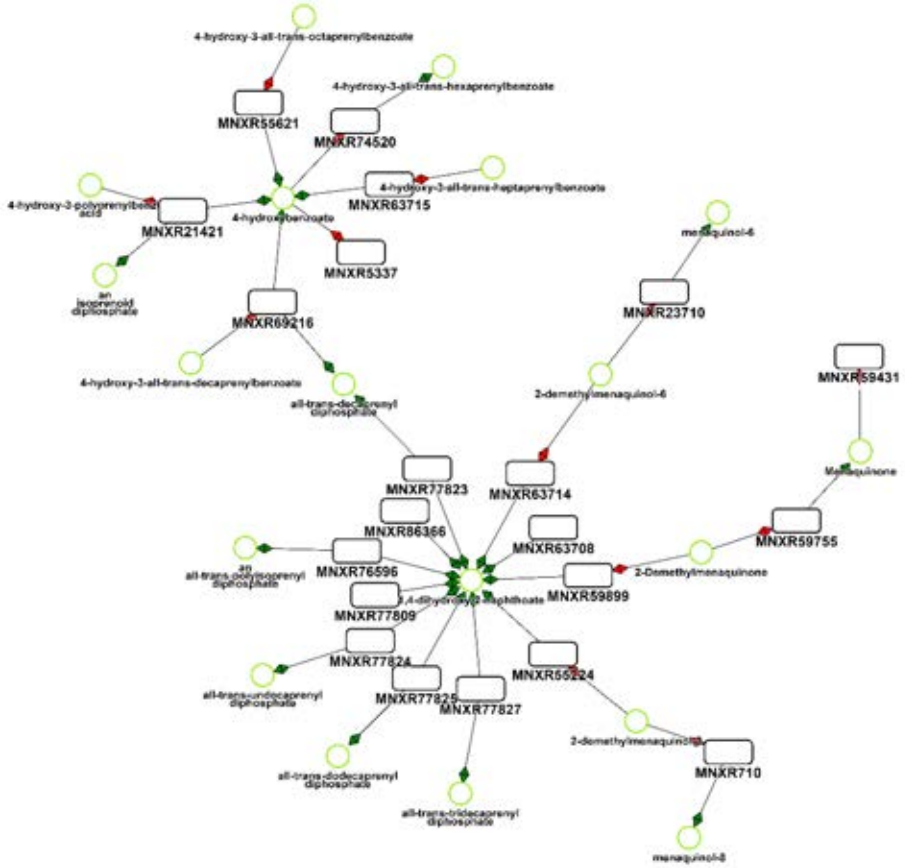
Zur H, Ruppin E, Shlomi T (2010) iMAT: an integrative metabolic analysis tool. *Bioinformatics* **26:** 3140-3142

Črepinšek M, Liu S-H, Mernik M (2013) Exploration and exploitation in evolutionary algorithms: A survey. *ACM Computing Surveys (CSUR)* **45:** 35

9. Appendix A – AMBIENT active modules

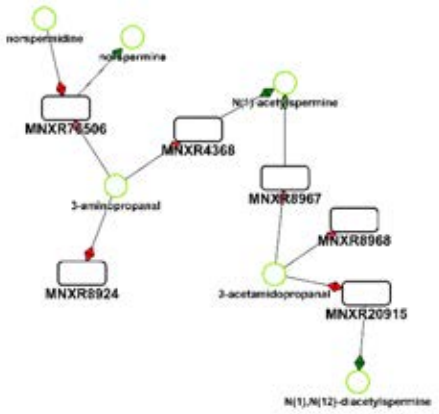
Below, the active modules generated using the AMBIENT algorithm (see section 5.3 and Table 5.2) are visualised with the metabolites contained within each of them listed in tables that show the name, formula, KEGG ID and MetaCyc ID of each metabolite.

Carbaryl treatment Module 1



NAME	FORMULA	KEGG ID	METACYC ID
1,4-dihydroxy-2-naphthoate	C11H7O4	C03657	DIHYDROXYNAPHTHOATE
2-demethylmenaquinol-6	C40H56O2	n/a	CPD-12116
2-demethylmenaquinol-8	C50H72O2	n/a	CPD-12115, CPD0-2129
2-Demethylmenaquinone	C15H14O2(C5H8)n	C05818	n/a
4-hydroxy-3-all-trans-decaprenylbenzoate	C57H85O3	n/a	CPD-9864
4-hydroxy-3-all-trans-heptaprenylbenzoate	C42H61O3	n/a	CPD-9852
4-hydroxy-3-all-trans-hexaprenylbenzoate	C37H53O3	C13425	3-HEXAPRENYL-4-HYDROXYBENZOATE
4-hydroxy-3-all-trans-octaprenylbenzoate	C47H69O3	C05809	3-OCTAPRENYL-4-HYDROXYBENZOATE
4-hydroxy-3-polyprenylbenzoic acid	(C5H8)nC7H6O3	n/a	4-Hydroxy-3-polyprenylbenzoates
4-hydroxybenzoate	C7H5O3	C00156	4-hydroxybenzoate
all-trans-decaprenyl diphosphate	C50H81O7P2	C17432	CPD-9610
all-trans-dodecaprenyl diphosphate	C60H97O7P2	n/a	CPD-9650
all-trans-tridecaprenyl diphosphate	C65H105O7P2	n/a	CPD-9972
all-trans-undecaprenyl diphosphate	C55H89O7P2	n/a	CPD-9649
an all-trans-polyisoprenyl diphosphate	n/a	n/a	TRANS-POLYISOPRENYL-PP
an isoprenoid diphosphate	n/a	n/a	Polyisoprenyl-Diphosphates
menaquinol-6	C41H58O2	n/a	CPD-12124
menaquinol-8	C51H74O2	n/a	REDUCED-MENAQUINONE
Menaquinone	C16H16O2(C5H8)n	C00828	n/a

Carbaryl treatment Module 2

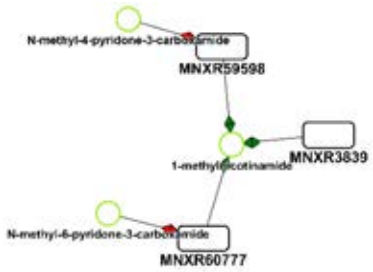


NAME	FORMULA	KEGG ID	METACYC ID
3-acetamidopropanal	C5H9NO2	C18170	CPD-10687
3-aminopropanal	C3H8NO	C05665	CPD-6082
N(1),n(12)-diacetylspermine	C14H32N4O2	C03413	CPD-11268
N(1)-acetylspermine	C12H31N4O	C02567	N1-ACETYLSPERMINE
Norspermidine	C6H20N3	C03375	NORSPERMIDINE
Norspermine	C9H28N4	n/a	CPD-10689

Carbaryl treatment Module 3

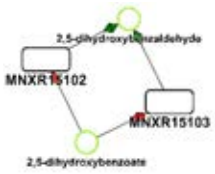
NAME	FORMULA	KEGG ID	METACYC ID
Bilirubin	C33H34N4O6	C00486	BILIRUBIN
Bilirubin-bisglucuronoside	C45H50N4O18	C05787	BILIRUBIN-BISGLUCURONOSIDE
Biliverdin	C33H32N4O6	C00500	BILIVERDINE
Fe(2+)	Fe	C14818	FE+2
Carbon monoxide(1+)	CHO	C00237, D09706, D03398	CARBON-MONOXIDE

Carbaryl treatment Module 4



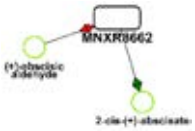
NAME	FORMULA	KEGG ID	METACYC ID
1-methylnicotinamide	C7H9N2O	C02918	CPD-396
n-methyl-6-pyridone-3-carboxamide	C7H8N2O2	C05842	n/a
n-methyl-4-pyridone-3-carboxamide	C7H8N2O2	C05843	n/a

Carbaryl treatment Module 5



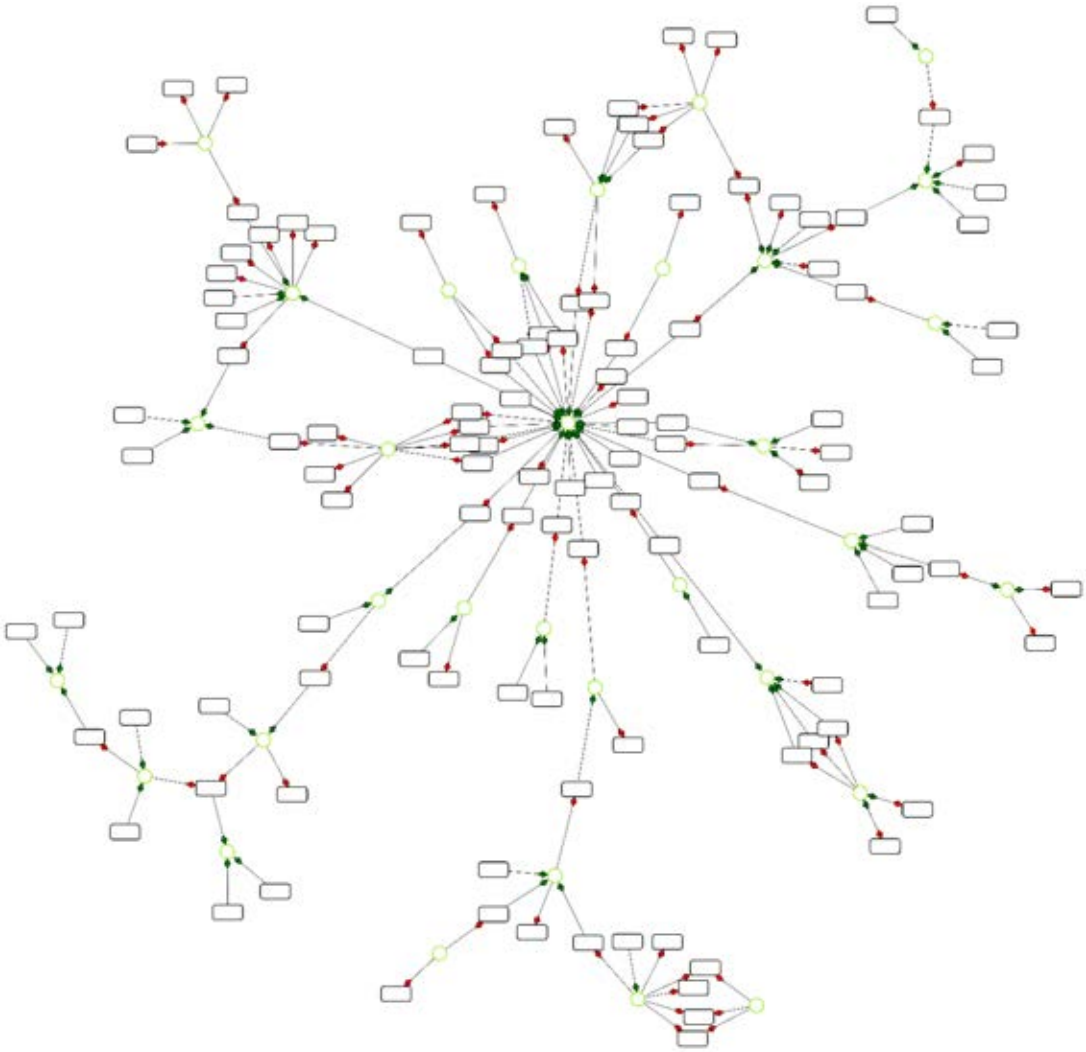
NAME	FORMULA	KEGG ID	METACYC ID
2,5-dihydroxybenzaldehyde	C7H6O3	C05585	CPD-16722
2,5-dihydroxybenzoate	C7H5O4	C00628	CPD-633

Carbaryl treatment Module 6



NAME	FORMULA	KEGG ID	METACYC ID
2-cis-(+)-abscisate	C15H19O4	C06082, C11060	CPD-693, CPD-7731
(+)-abscisic aldehyde	C15H20O3	C13455	CPD-14385, CPD-692

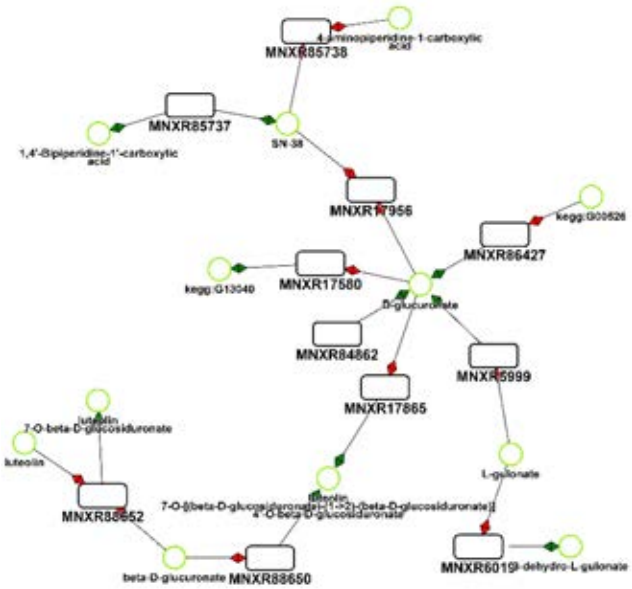
Carbaryl treatment Module 7



NAME	FORMULA	KEGG ID	METACYC ID
Alpha-maltotriose	C18H32O16	C01835	MALTOTRIOSE
Linear maltodextrin	(C12H20O10)n	C01935, C00718, D02329, G10495	n/a
Beta-d-galactosyl-(1->4)-beta-d-glucosyl-(11)-n-acylsphing-4-enine	C31H56NO13R	C01290, G00092	Lactosyl-Ceramides
N-acetyl-beta-d-galactosaminyl-(1->4)-beta-d-galactosyl-(1->4)-beta-d-glucosyl-(11')-n-acylsphing-4-enine	C39H69N2O18R	C06135, G00123	n/a
Beta-d-galactose	C6H12O6	C00962	GALACTOSE
A glycogen	C204H342O171	C00182	Glycogens
Oligoglycosylglucose	(C12H20O10)n	C03018, C00369, C00721, D00084, D06507, G10545	CPD-8556

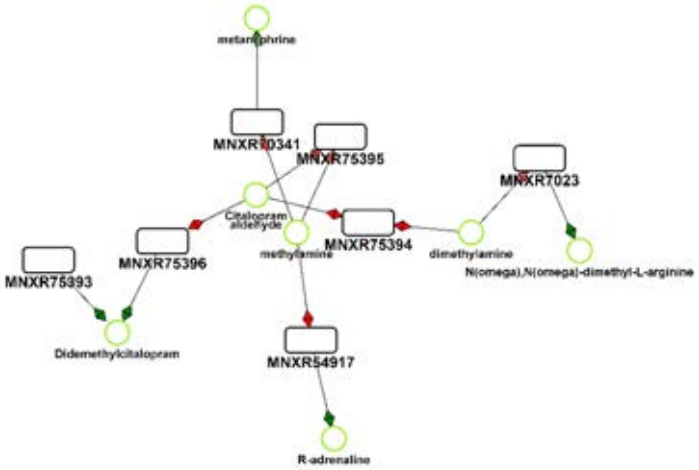
Xxxg xyloglucan oligosaccharide	C39H66O33	n/a	CPD-13375
Melibiose	C12H22O11	C05402, G01275	MELIBIOSE
Alpha,alpha-trehalose	C12H22O11	C01083, G00293	TREHALOSE
Cis-beta-d-glucosyl-2-hydroxycinnamate	C15H17O8	C05839	CPD-7417
Stachyose	C24H42O21	C01613, G00278	CPD-170
Cellobiose	C12H22O11	C00185, G00289, C06422	CELLOBIOSE, CPD-15975
Amylose	C14H26O11	n/a	1-4-alpha-D-Glucan
N-acetyl-beta-d-galactosaminyl-(1->4)-[alpha-n-acetylneuraminosyl-(2->3)]-beta-d-galactosyl-(1->4)-beta-d-glucosyl-(11)-n-acylsphing-4-enine	C50H85N3O26R	C04884, G00109	CPD-1100
Galabiose	C12H22O11	C00760, D00093, G10481	CELLULOSE
A debranched alpha-limit dextrin	C60H102O51	C02492, G10532	CPD0-1027
Alpha-d-glucose 6-phosphate	C6H11O9P	C00668	ALPHA-GLC-6-P
Isomaltose	C12H22O11	C00252, G01318	Isomaltose
Lactose	C12H22O11	C00243, D00046, G10504	CPD-15972
Beta-d-glucosyl-(11)-n-acylsphing-4-enine	C25H46NO8R	C01190, G10238	n/a
D-glucose	C6H12O6	C00031, D00009	Glucopyranose
1d-myo-inositol 3-phosphate	C6H11O9P	C04006	1-L-MYO-INOSITOL-1-P
Beta-d-glucose 6-phosphate	C6H11O9P	C01172	GLC-6-P
Starch	n/a	n/a	Starch
Glucose	C6H12O6	n/a	Glucose
Alpha-maltohexaose	C36H62O31	C01936, G00755	MALTOHEXAOSE
Sucrose	C12H22O11	C00089, D00025, G00370	SUCROSE
D-glucopyranose 6-phosphate	C6H11O9P	C00092	D-glucopyranose-6-phosphate
Alpha-maltopentaose	C30H52O26	C06218, G00546	CPD-13237, MALTOPENTAOSE
An alpha-limit dextrin	C156H262O131	n/a	CPD0-971
Alpha-d-galactose	C6H12O6	C00984, D04291	ALPHA-D-GALACTOSE
N-acetyl-d-galactosamine	C8H15NO6	C01132	N-acetyl-D-galactosamine

Carbaryl treatment Module 8



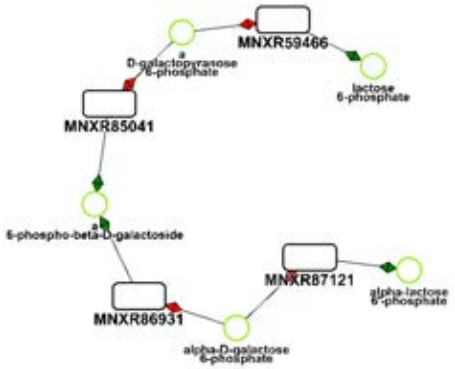
NAME	FORMULA	KEGG ID	METACYC ID
4-aminopiperidine-1-carboxylic acid	C6H12N2O2	C16837	n/a
1,4'-bipiperidine-1'-carboxylic acid	C11H20N2O2	C16836	n/a
beta-d-glucuronate	C6H9O7	C08350	CPD-12521
3-dehydro-l-gulonate	C6H9O7	C00618	3-KETO-L-GULONATE
luteolin 7-o-beta-d-glucosiduronate	C21H16O12	C03515	LUTEOLIN-7-O-BETA-D-GLUCURONIDE
d-glucuronate	C6H9O7	C00191	D-Glucopyranuronate
luteolin	C15H9O6	C01514	5734-TETRAHYDROXYFLAVONE
sn-38	C22H20N2O5	C11173	n/a
l-gulonate	C6H11O7	C00800	CPD-16836, L-GULONATE
kegg:g13040	n/a	G13040	n/a
kegg:g00526	n/a	G00526	n/a
luteolin 7-o-[(beta-d-glucosiduronate)-(1->2)-(beta-d-glucosiduronate)] 4'-o-beta-d-glucosiduronate	C33H30O24	C04900	LUTEOLIN-7-O-BETA-D-GLUCURONOSYL-1-2

Carbaryl treatment Module 9



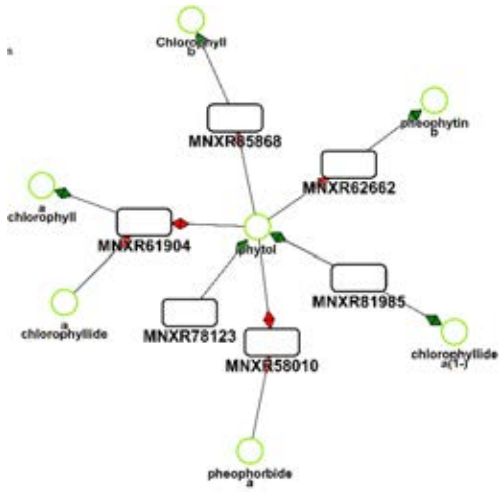
NAME	FORMULA	KEGG ID	METACYC ID
Metanephrine	C10H16NO3	C05588	CPD-11877
Methylamine	CH6N	C00218	A-METHYLATED-AMINE, METHYLAMINE
Dimethylamine	C2H8N	C00543	DIMETHYLAMINE
Citalopram aldehyde	C18H14FNO2	C16612	n/a
Didemethylcitalopram	C18H18FN2O	C16609	n/a
N(omega),n(omega)-dimethyl-l-arginine	C8H19N4O2	C03626	CPD-596
R-adrenaline	C9H14NO3	C00788, D00095, D05688	L-EPINEPHRINE

Carbaryl treatment Module 10



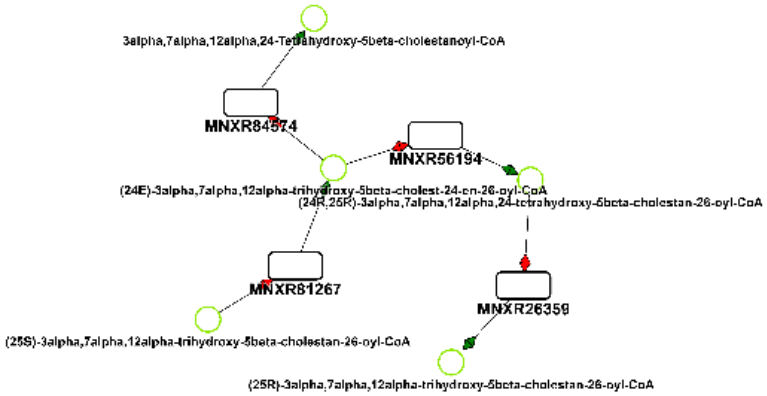
NAME	FORMULA	KEGG ID	METACYC ID
Lactose 6-phosphate	C12H21O14P	C05396, G10517	CPD-15973
Alpha-lactose 6'-phosphate	C12H21O14P	n/a	LACTOSE-6P
Alpha-d-galactose 6-phosphate	C6H11O9P	n/a	CPD-1241
A D-GALACTOPYRANOSE 6-PHOSPHATE	C6H11O9P	C01113	D-galactopyranose-6-phosphate
A 6-PHOSPHO-BETA-D-GALACTOSIDE	C6H10O9PR	C03847	6-Phospho-b-D-galactosides

Carbaryl treatment Module 11



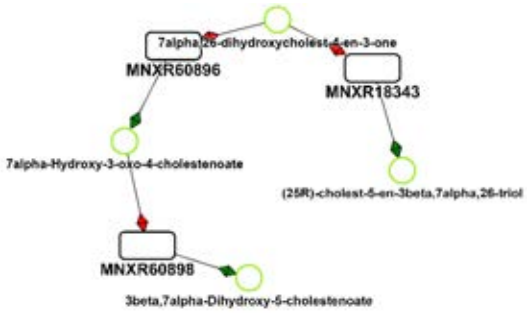
NAME	FORMULA	KEGG ID	METACYC ID
Phytol	C20H40O	C01389	PHYTOL
Chlorophyll b	C55H70MgN4O6	C05307	CHLOROPHYLL-B
Chlorophyllide a(1-)	C35H33MgN4O5	C02139	CHLOROPHYLLIDE-A
Pheophytin b	C55H72N4O6	n/a	CPD-8178
A chlorophyllide	.	n/a	Chlorophyllides
A chlorophyll	.	C01793	Chlorophylls
Pheophorbide a	C35H35N4O5	C18021	CPD-7061

Carbaryl treatment Module 12



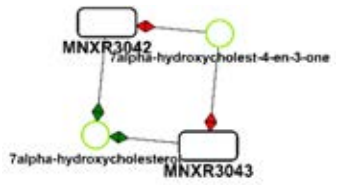
NAME	FORMULA	KEGG ID	METACYC ID
3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestanoyl-coa	C48H76N7O21P3S	C05450	n/a
(24r,25r)-3alpha,7alpha,12alpha,24-tetrahydroxy-5beta-cholestan-26-oyl-coa	C48H76N7O21P3S	C15614	CPD-7275
(24e)-3alpha,7alpha,12alpha-trihydroxy-5beta-cholest-24-en-26-oyl-coa	C48H74N7O20P3S	C05460	CPD-7243
(25r)-3alpha,7alpha,12alpha-trihydroxy-5beta-cholestan-26-oyl-coa	C48H76N7O20P3S	C15613	CPD-71
(25s)-3alpha,7alpha,12alpha-trihydroxy-5beta-cholestan-26-oyl-coa	C48H76N7O20P3S	C17343	CPD-10505

Carbaryl treatment Module 13



NAME	FORMULA	KEGG ID	METACYC ID
3beta,7alpha-dihydroxy-5-cholestenoate	C27H43O4	C17335	n/a
7alpha-hydroxy-3-oxo-4-cholestenoate	C27H41O4	C17337	n/a
(25r)-cholest-5-en-3beta,7alpha,26-triol	C27H46O3	C06341	7-ALPHA27-DIHYDROXYCHOLESTEROL
7alpha,26-dihydroxycholest-4-en-3-one	C27H44O3	C17336	n/a

Carbaryl treatment Module 14



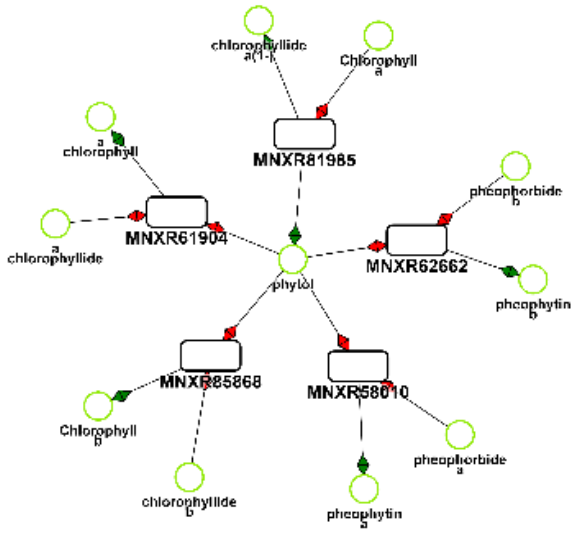
NAME	FORMULA	KEGG ID	METACYC ID
7alpha-hydroxycholest-4-en-3-one	C27H44O2	C05455	CPD-1087
7alpha-hydroxycholesterol	C27H46O2	C03594	CPD-266

Lead treatment Module 1

NAME	FORMULA	KEGG ID	METACYC ID
4-hydroxy-2-nonenal-glutathione conjugate	C19H32N3O8S	n/a	CPD-14704
4-hydroxy-2-nonenal-[cys-gly] conjugate	C14H26N2O5S	n/a	CPD-14705
leukotriene d4	C25H39N2O6S	C05951	CPD66-21
ubiquinone	C14H18O4(C5H8)n	C00399	n/a
l-cysteinylglycine	C5H10N2O3S	C01419	CYS-GLY
benzo[a]pyrene-7,8-diol	C20H14O2	C14852	n/a
mandelonitrile	C8H7NO	C00561	CPD-12702
methylselenol	CH4Se	C05703	n/a
n(5)-alkyl-l-glutamine residue	C5H7N2O2R	C03636	Protein-N5-alkylglutamines
n(2)-acetyl-l-ornithine	C7H14N2O3	C00437	N-ALPHA-ACETYLORNITHINE
an s-substituted l-cysteine	C3H6NO2SR	C02882, C05726	S-Substituted-L-Cysteines
glutathione	C10H16N3O6S	C00051, D00014	GLUTATHIONE
aflatoxin b1 exo-8,9-epoxide	C17H12O7	C19586	n/a
5-l-glutamyl amino acid	C7H10N2O5R	C03363	5-L-GLUTAMYL-AMINO-ACID
15h-11,12-eeta	C20H31O4	C14781	n/a
r-s-alanylglycine	C5H9N2O3SR	C05729	n/a
chloride	Cl	C00698, C01327, D02057	CL-, HCL
bromobenzene-2,3-oxide	C6H5BrO	C14840	n/a
se-methyl-l-selenocysteine	C4H9NO2Se	C05689	CPD-12024
trichloroacetaldehyde	C2HCl3O	C14866	n/a
5-oxo-l-proline	C5H6NO3	C01879	5-OXOPROLINE
l-gamma-glutamyl-l-cysteine	C8H13N2O5S	C00669	L-GAMMA-GLUTAMYLCYSTEINE
an l-amino acid	C2H4NO2R	C00151	L-Amino-Acids
hydrogen cyanide	CHN	C00177, C01326	CPD-13584, HCN
l-ornithine	C5H13N2O2	C00077, D08302, C01602	L-ORNITHINE
11h-14,15-eeta	C20H31O4	C14813	n/a
urea	CH4N2O	C00086, D00023, D01749, D10496	UREA
methylglyoxal	C3H4O2	C00546	CPD-10807, METHYL-GLYOXAL
cyanohydrin	C2HNOR2	C05712	n/a

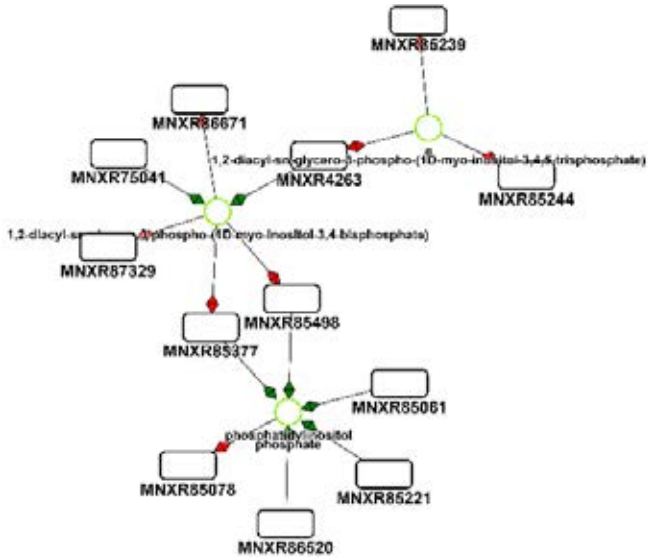
5-oxoprolyl-peptide	C7H9N2O4R(C2H2NOR)n	C02805	5-OXOPROLYL-PEPTIDE
a l-gamma-glutamyl-l-amino acid	C7H10N2O5R	C03740	5-L-GLUTAMYL-L-AMINO-ACID
a nitrile	CNR	C00726	Nitriles
l-glutaminyl-peptide	C7H12N3O4R(C2H2NOR)n	C02986	Protein-L-glutamine
(5s,6e,8z,11z,14z)-5-hydroperoxyicosa-6,8,11,14-tetraenoate	C20H31O4	C05356	6E8Z11Z14Z-5S-5-HYDROPEROXYCOSA-6
bromobenzene-3,4-oxide	C6H5BrO	C14839	n/a
1-nitronaphthalene-7,8-oxide	C10H7NO3	C14802	n/a
a reduced electron-transfer flavoprotein	n/a	C04570	ETF-Reduced
r-s-glutathione	C10H16N3O6SR	C02320	S-Substituted-Glutathione
(5z,8z,11z,13e,15s)-15-hydroperoxyicosa-5,8,11,13-tetraenoate	C20H31O4	C05966	5Z8Z11Z13E-15S-15-HYDROPEROXYICOS
aldophosphamide	C7H15Cl2N2O3P	C07645	n/a
trichloroethene	C2HCl3	C06790	TRICHLOROETHENE
1-nitronaphthalene-5,6-oxide	C10H7NO3	C14800	n/a
s-(hydroxymethyl)glutathione	C11H18N3O7S	C14180	S-HYDROXYMETHYLGLUTATHIONE
(r)-lactate	C3H5O3	C00256	D-LACTATE
a ketone	COR2	C01450	LONG-CHAIN-KETONE
(gamma-l-glutamyl) n-terminal alpha-amino-acid residue	C7H10N2O4R	C03193	5-L-GLUTAMYL-PEPTIDE
ferricytochrome c	C42H44FeN8O8S2R4	C00125	n/a

Lead treatment Module 2



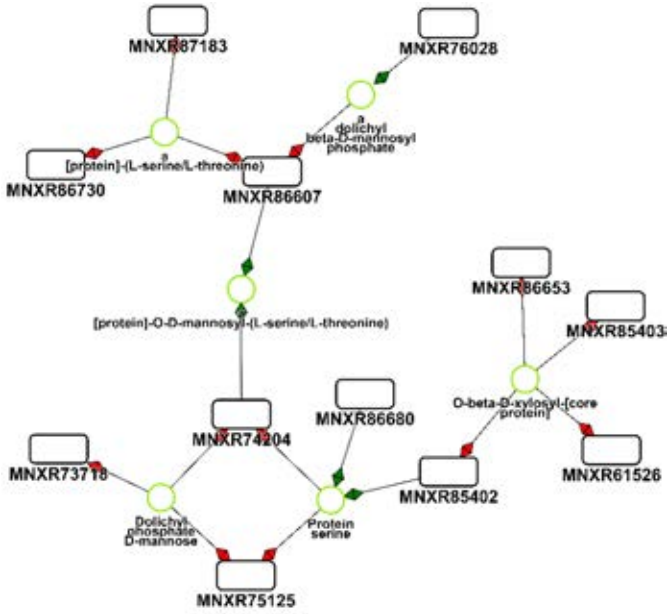
NAME	FORMULA	KEGG ID	METACYC ID
Phytol	C20H40O	C01389	PHYTOL
Chlorophyll a	C55H72MgN4O5	C05306	CHLOROPHYLL-A
Chlorophyll b	C55H70MgN4O6	C05307	CHLOROPHYLL-B
Chlorophyllide a(1-)	C35H33MgN4O5	C02139	CHLOROPHYLLIDE-A
Pheophorbide b	C35H33N4O6	n/a	CPD-7062
Pheophytin a	C55H74N4O5	C05797	CPD-10334, CPD-8155
A chlorophyllide	.	n/a	Chlorophyllides
Chlorophyllide b	C35H31MgN4O6	C16541	CPD-7014
A chlorophyll	.	C01793	Chlorophylls
Pheophytin b	C55H72N4O6	n/a	CPD-8178
Pheophorbide a	C35H35N4O5	C18021	CPD-7061

Lead treatment Module 3



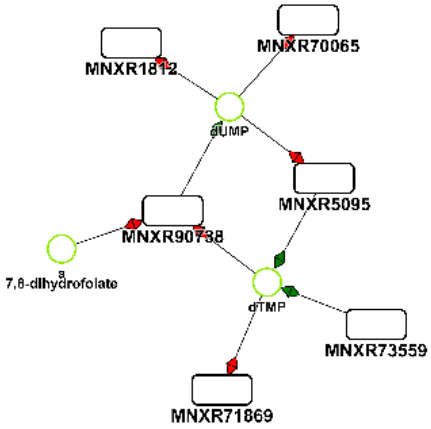
NAME	FORMULA	KEGG ID	METACYC ID
A 1,2-diacyl-sn-glycero-3-phospho-(1d-myo-inositol-3,4,5-trisphosphate)	C11H13O22P4R2	C05981	PHOSPHATIDYLINOSITOL-345-TRIPHOSPHATE
Phosphatidylinositol phosphate	C11H18O16P2R2	C01277, C04021	n/a
1,2-diacyl-sn-glycero-3-phospho-(1d-myo-inositol-3,4-bisphosphate)	C11H14O19P3R2	C11554	1-PHOSPHATIDYL-1D-MYO-INOSITOL-34-BISPH

Lead treatment Module 4



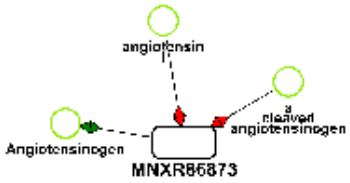
NAME	FORMULA	KEGG ID	METACYC ID
A [protein]-o-d-mannosyl-(l-serine/l-threonine)	C10H16N2O8R2	C02863	O-D-MANNOSYL-PROTEIN
O-beta-d-xylosyl-[core protein]	C9H14N2O7R2	C02399, G00154	Core-Protein-L-Ser-Xyl
Dolichyl phosphate d-mannose	C26H47O9P(C5H8)n	C03862, G10617	n/a
A [protein]-(l-serine/l-threonine)	n/a	n/a	Protein-L-serine-or-L-threonine
Protein serine	C4H6N2O3R2	C02189, C06395	Protein-L-serines
A dolichyl beta-d-mannosyl phosphate	C86H142O9P	n/a	CPD-171

Lead treatment Module 5



NAME	FORMULA	KEGG ID	METACYC ID
dTMP	C10H13N2O8P	C00364	TMP
A 7,8-dihydrofolate	n/a	n/a	DIHYDROFOLATE-GLU-N
dUMP	C9H11N2O8P	C00365	DUMP

Lead treatment Module 6

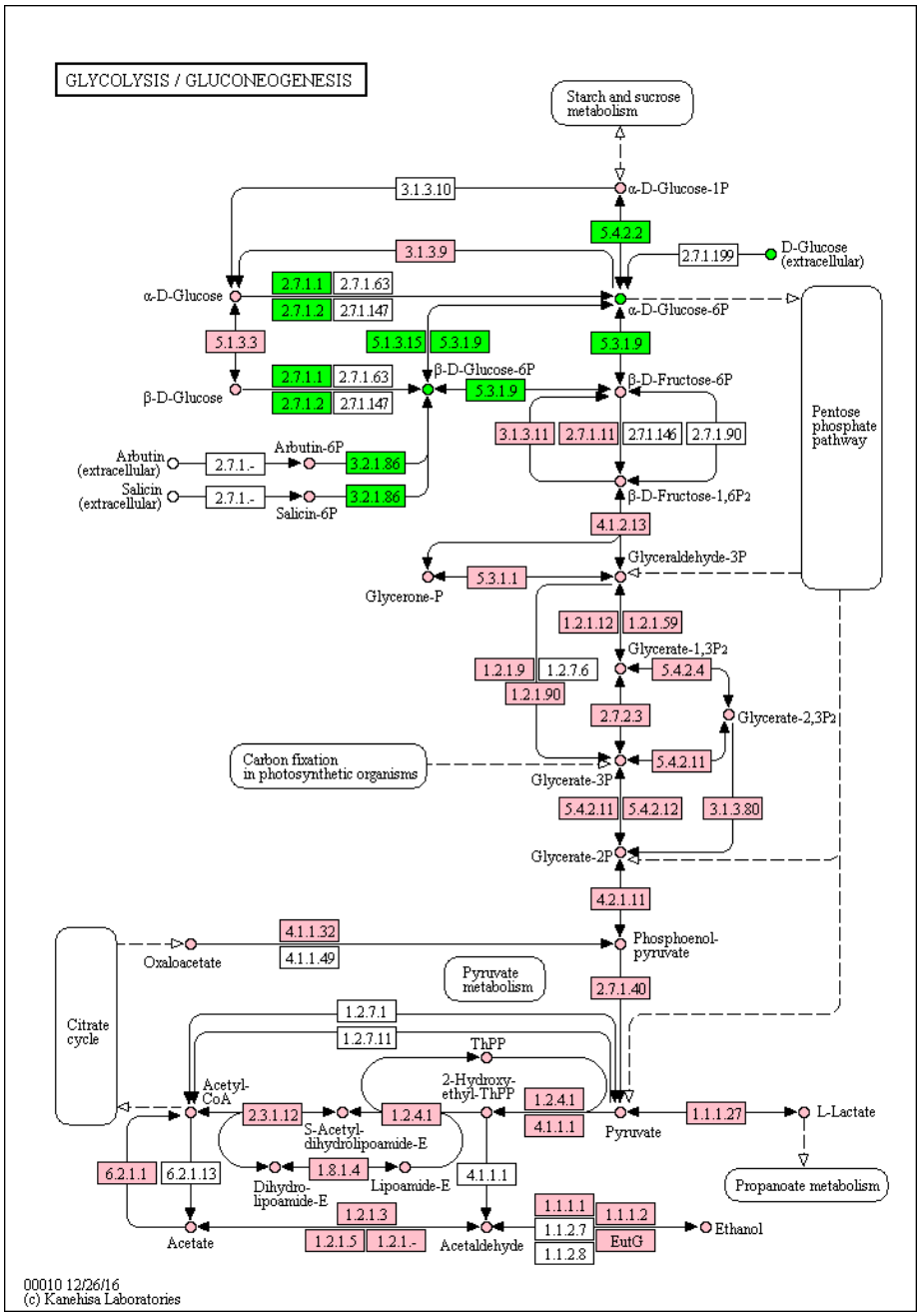


NAME	FORMULA	KEGG ID	METACYC ID
Angiotensin I	C62H89N17O14	C00873	CPD-13004
A cleaved angiotensinogen	n/a	n/a	Cleaved-Angiotensinogen
Angiotensinogen	n/a	C02246	Angiotensinogens

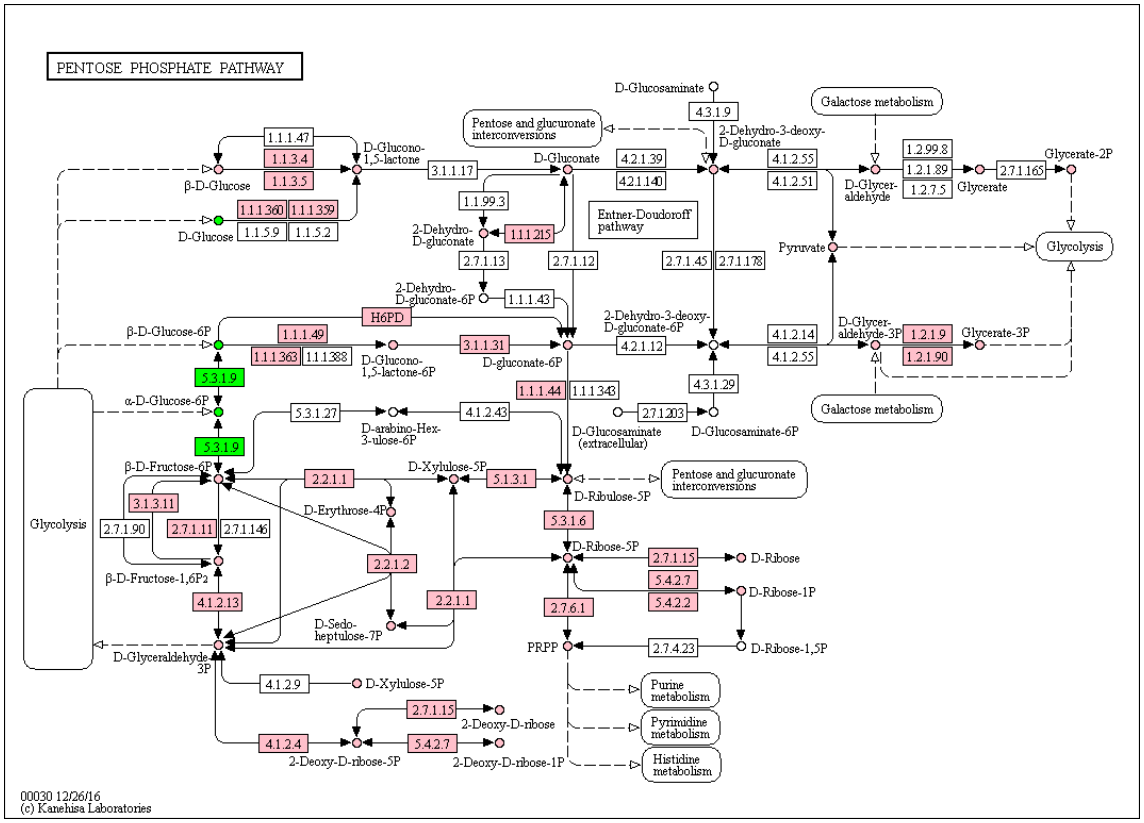
10. Appendix B – KEGG pathway analysis

The figures below show KEGG pathways that are predicted to be affected by the Carbaryl treatment (see section 5.3.2.1). Reactions and metabolites are coloured pink if they are present in the *D. magna* draft GWMR and green if they are in an identified AMBIENT active module for the Carbaryl STRESSFLEA dataset

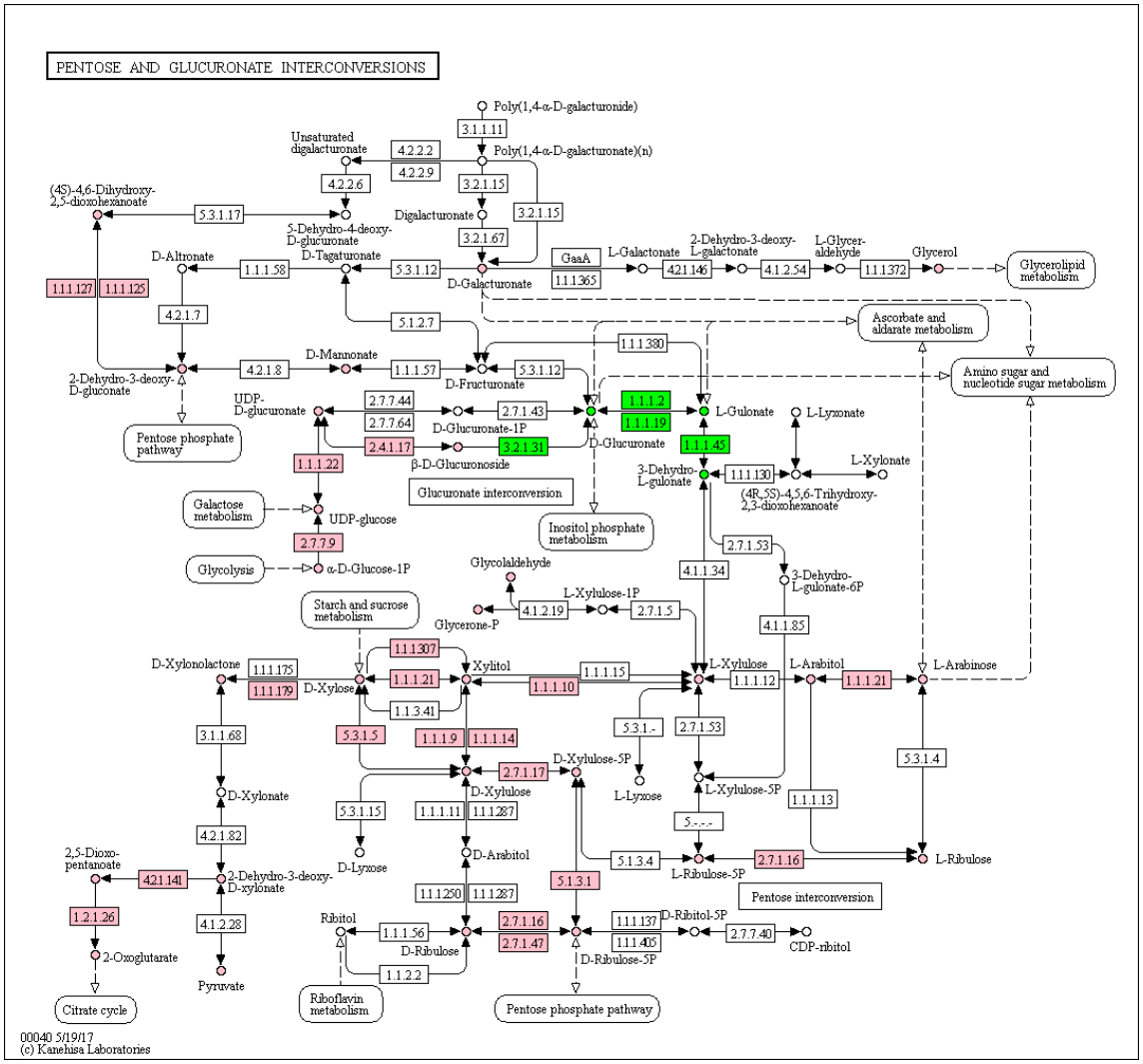
Glycolysis / gluconeogenesis



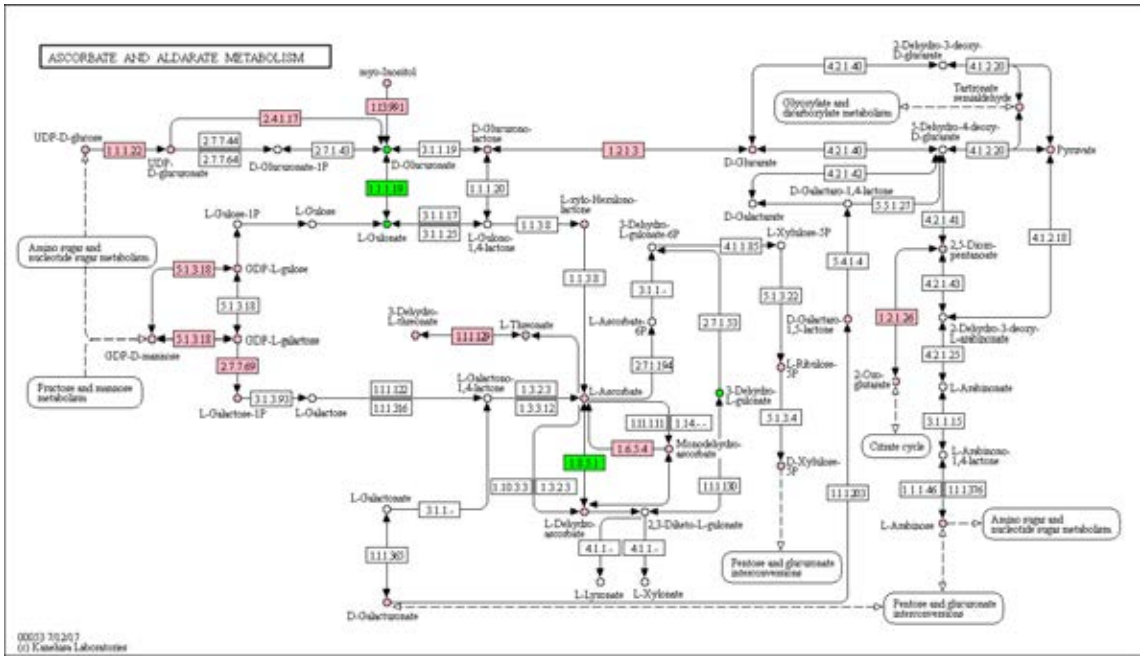
Pentose phosphate pathway



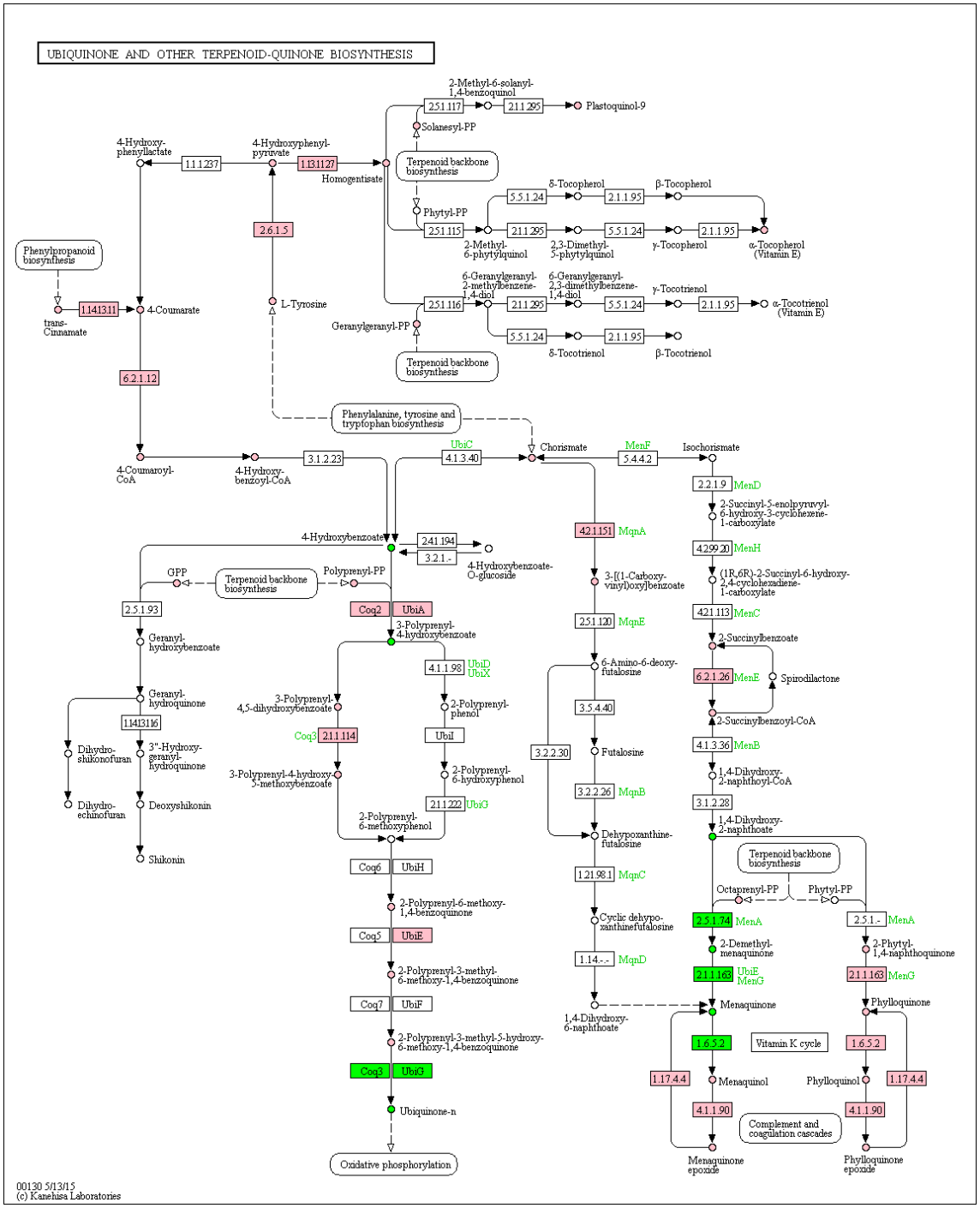
Pentose and glucuronate interconversions



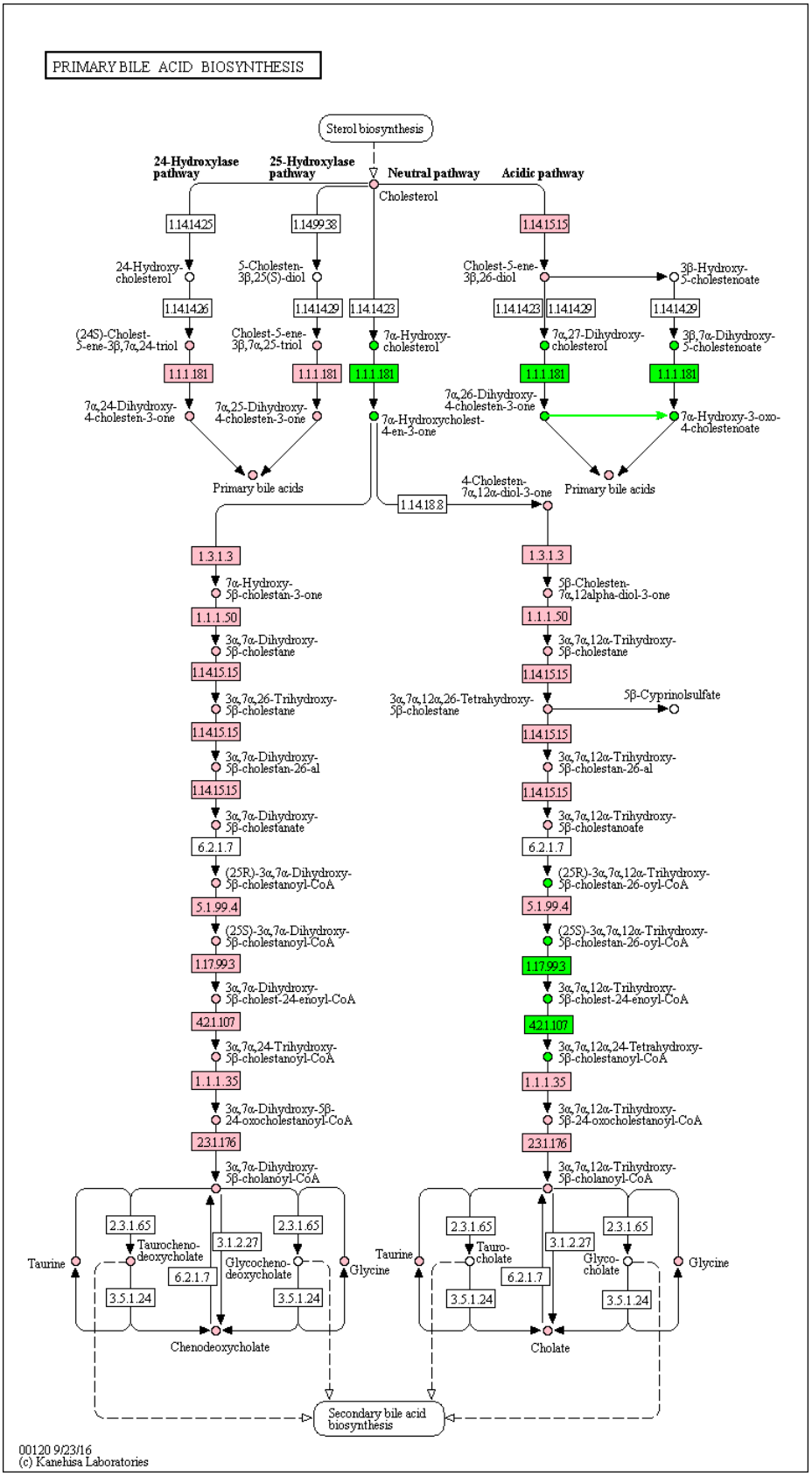
Ascorbate and aldarate metabolism



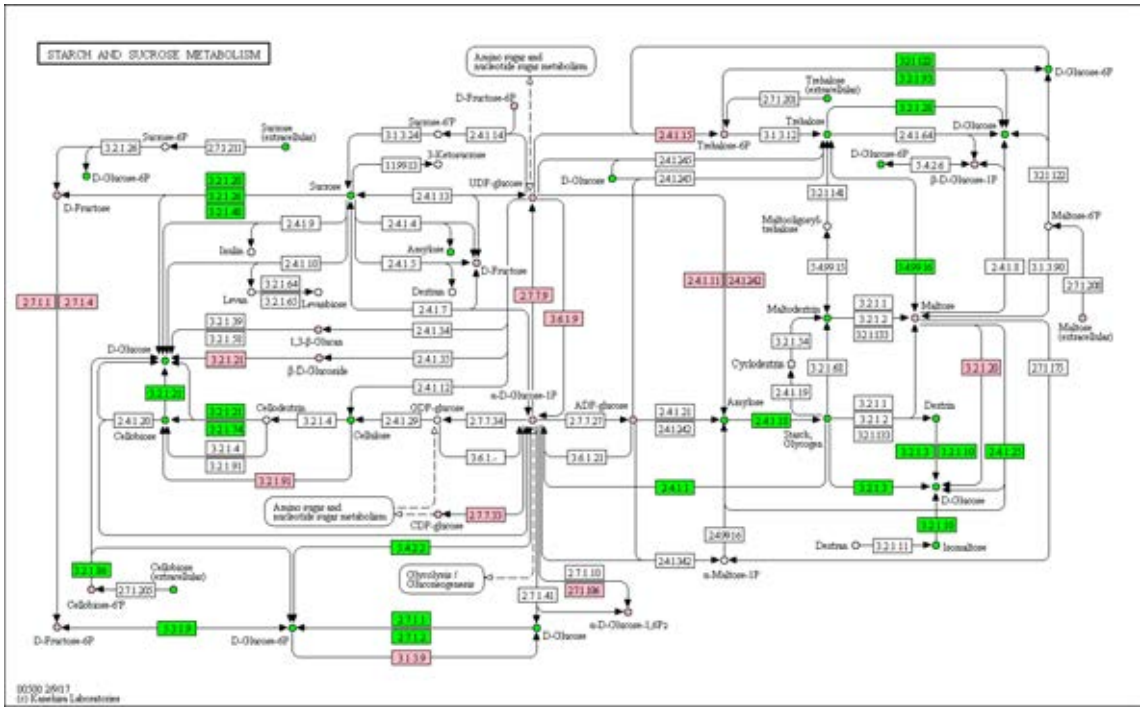
Ubiquinone and other terpenoid-quinone biosynthesis



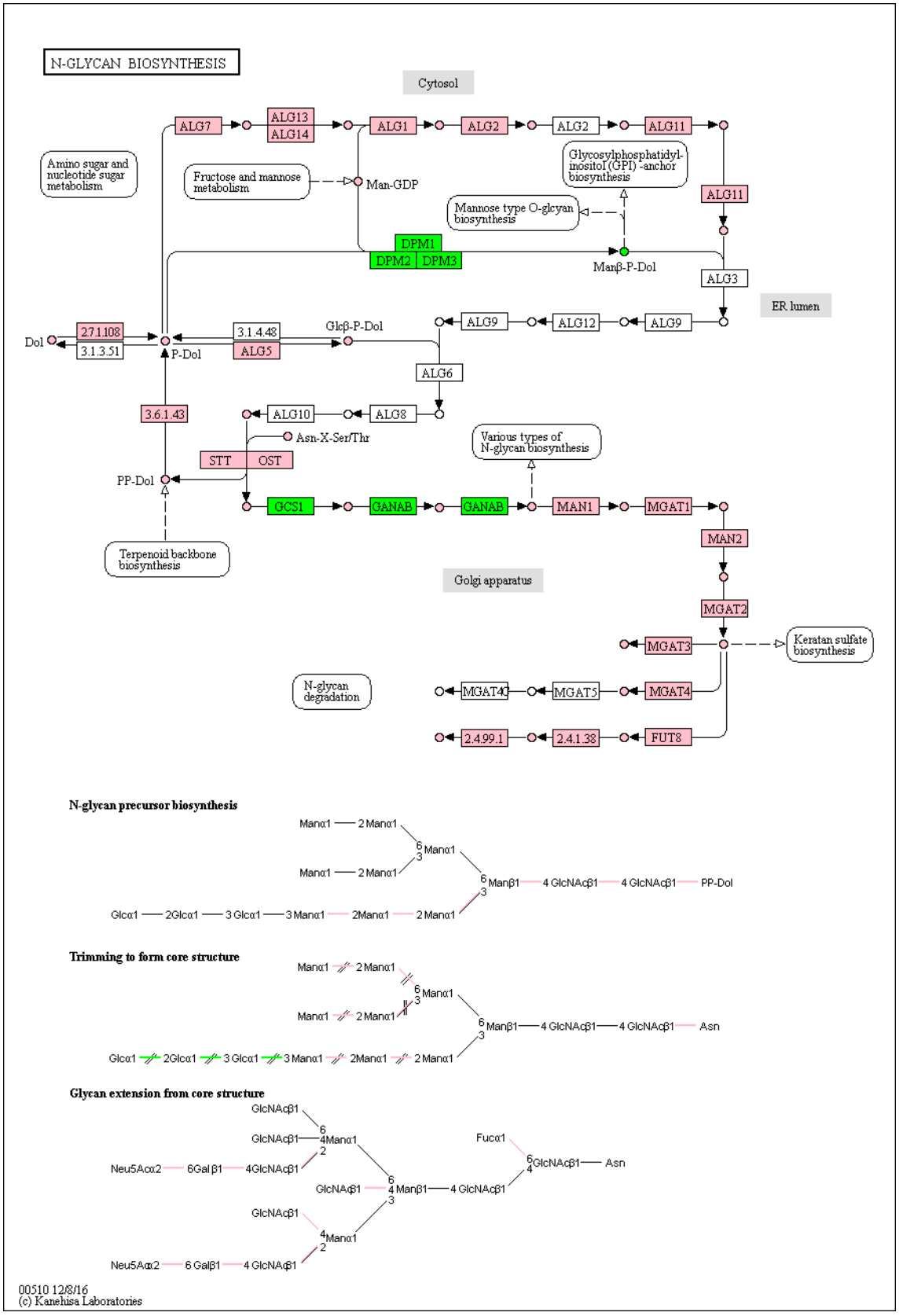
Primary bile acid biosynthesis



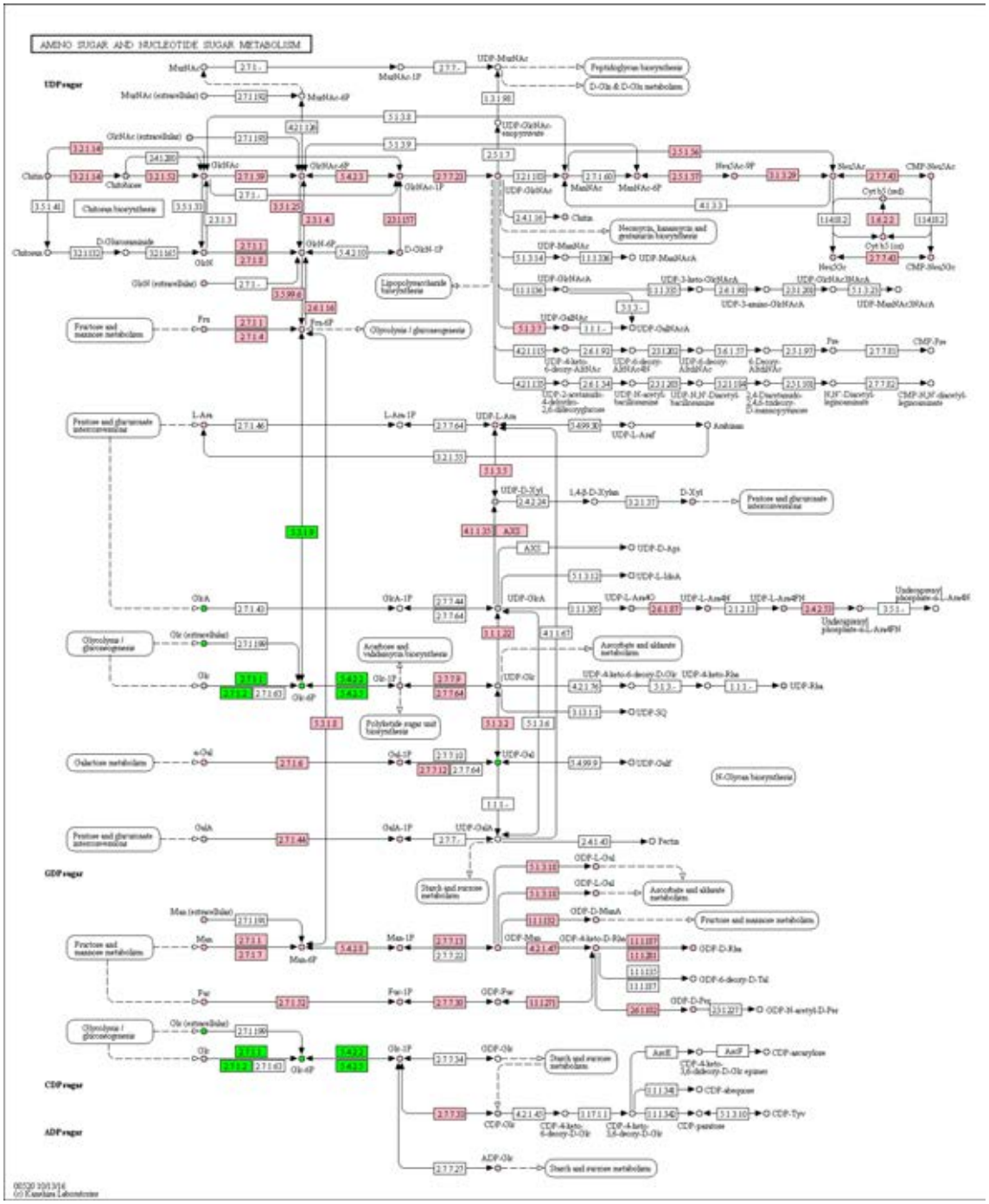
Starch and sucrose metabolism



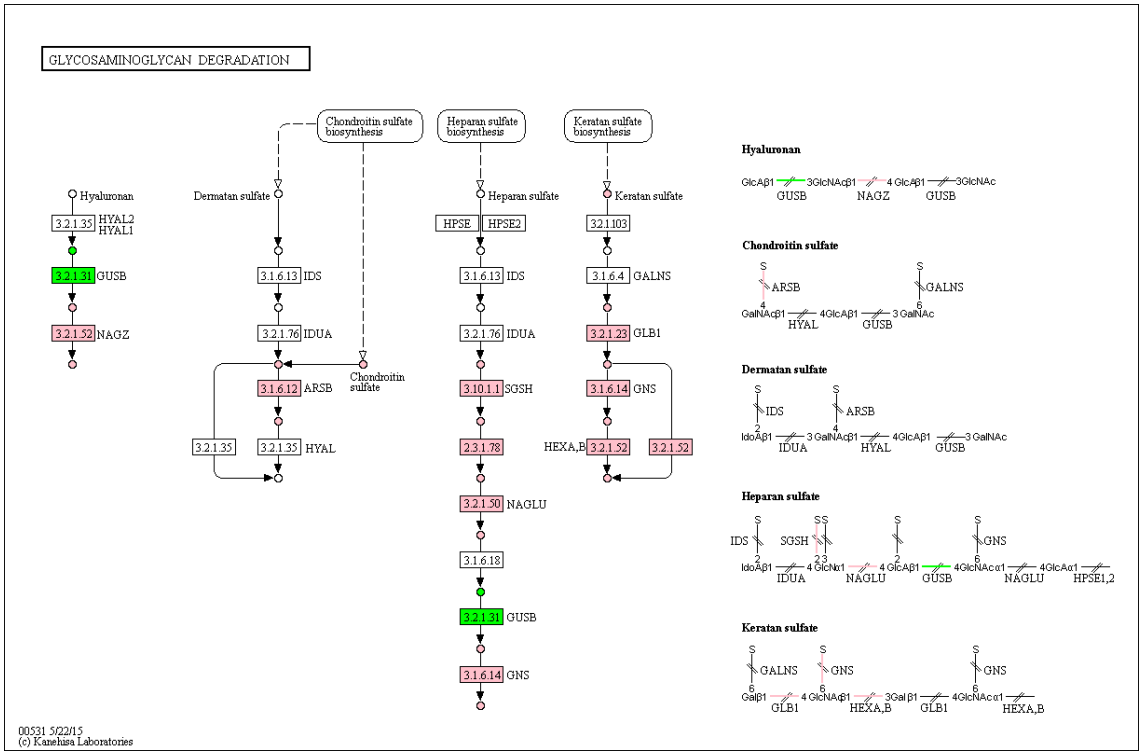
N-Glycan biosynthesis



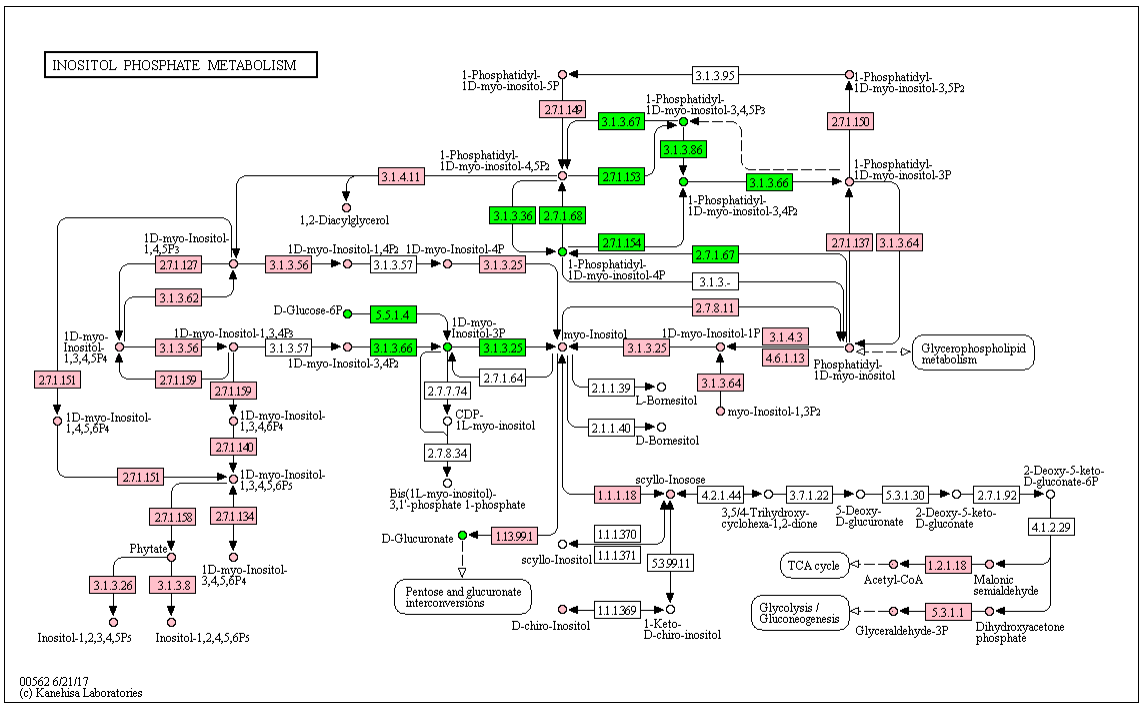
Amino sugar and nucleotide sugar metabolism



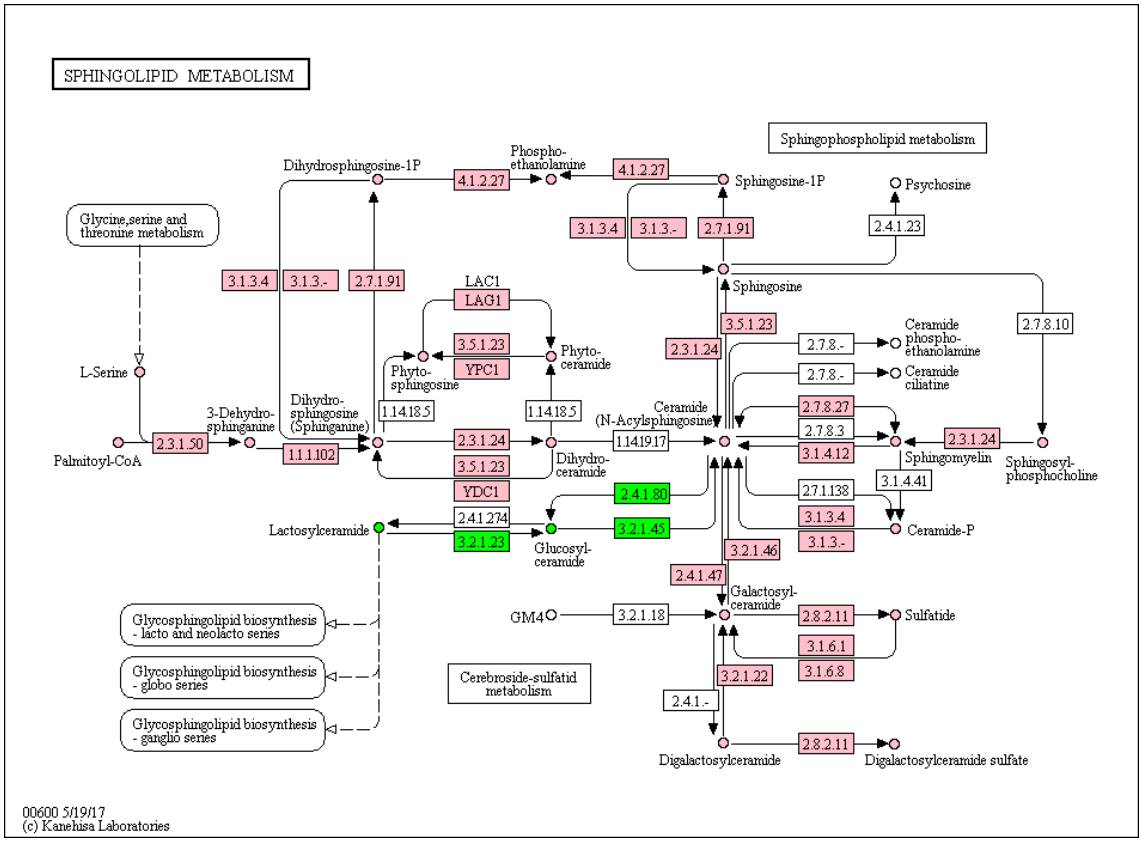
Glycosaminoglycan degradation



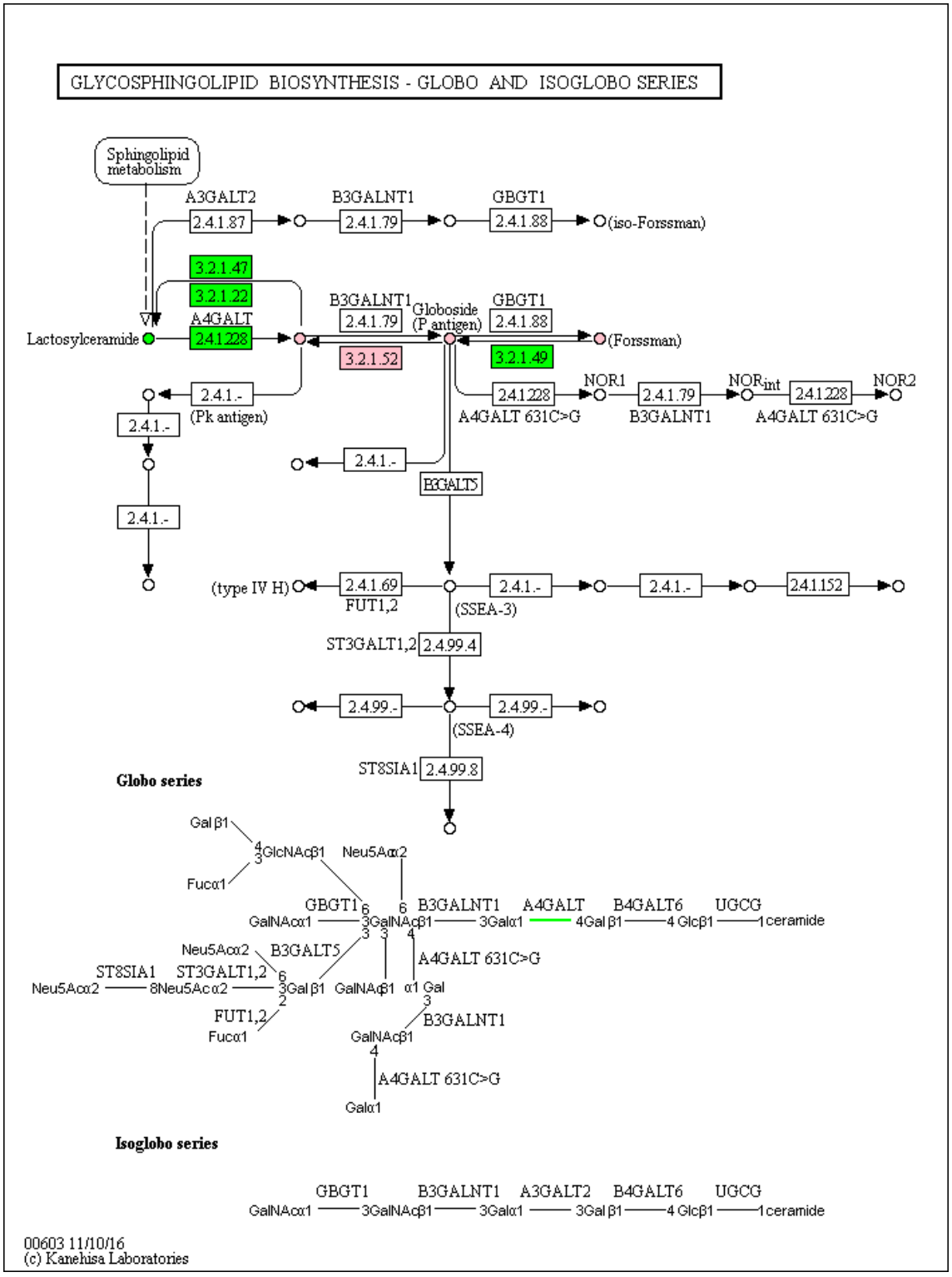
Inositol phosphate metabolism



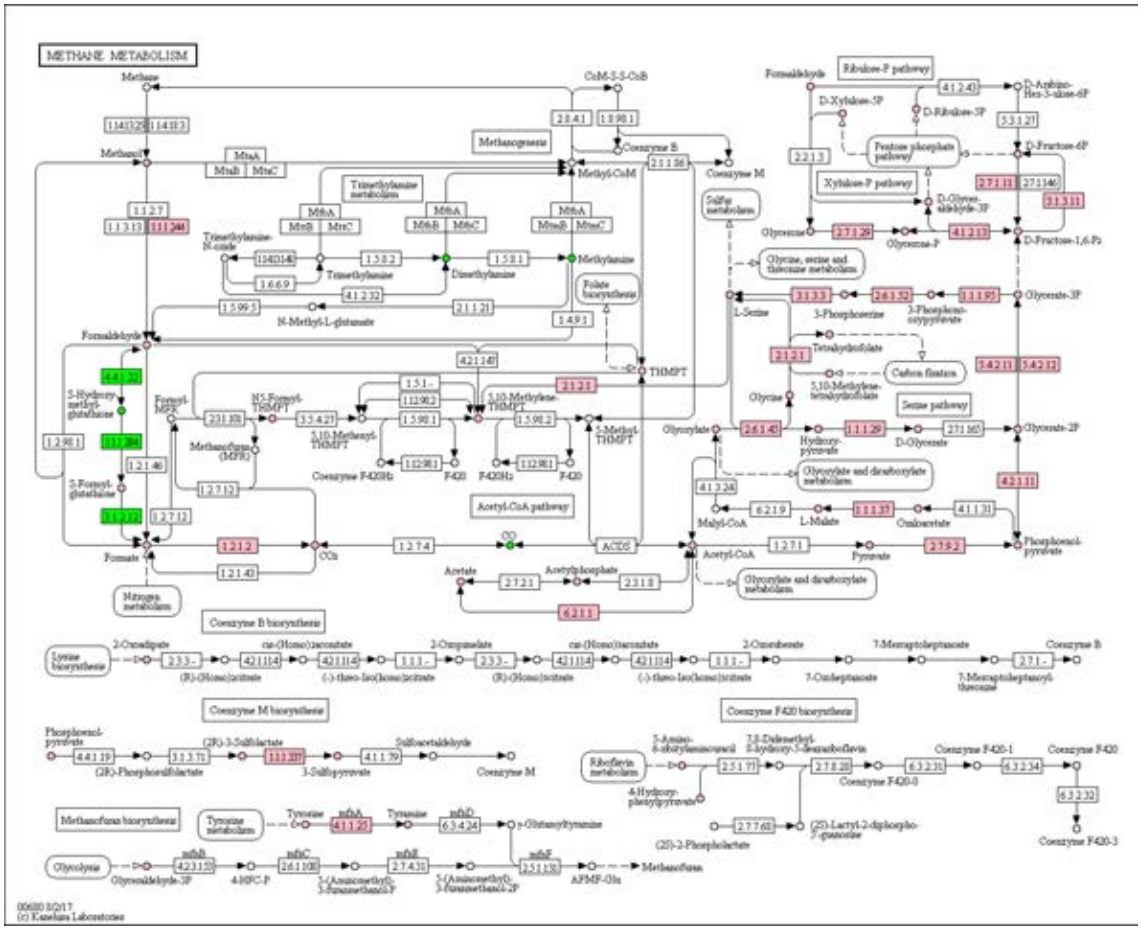
Sphingolipid metabolism



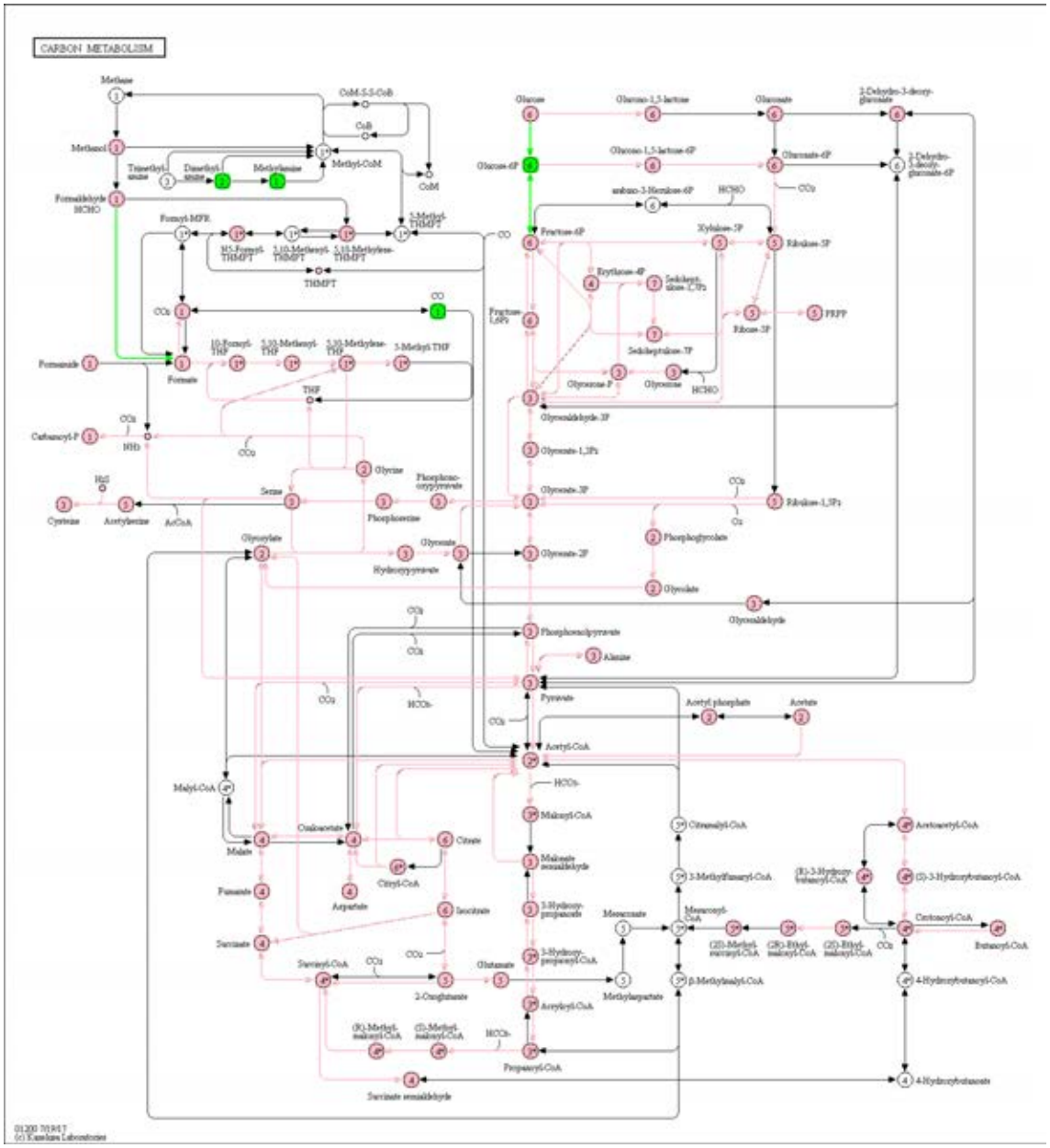
Glycosphingolipid biosynthesis – Globo and isoglobo series



Methane metabolism

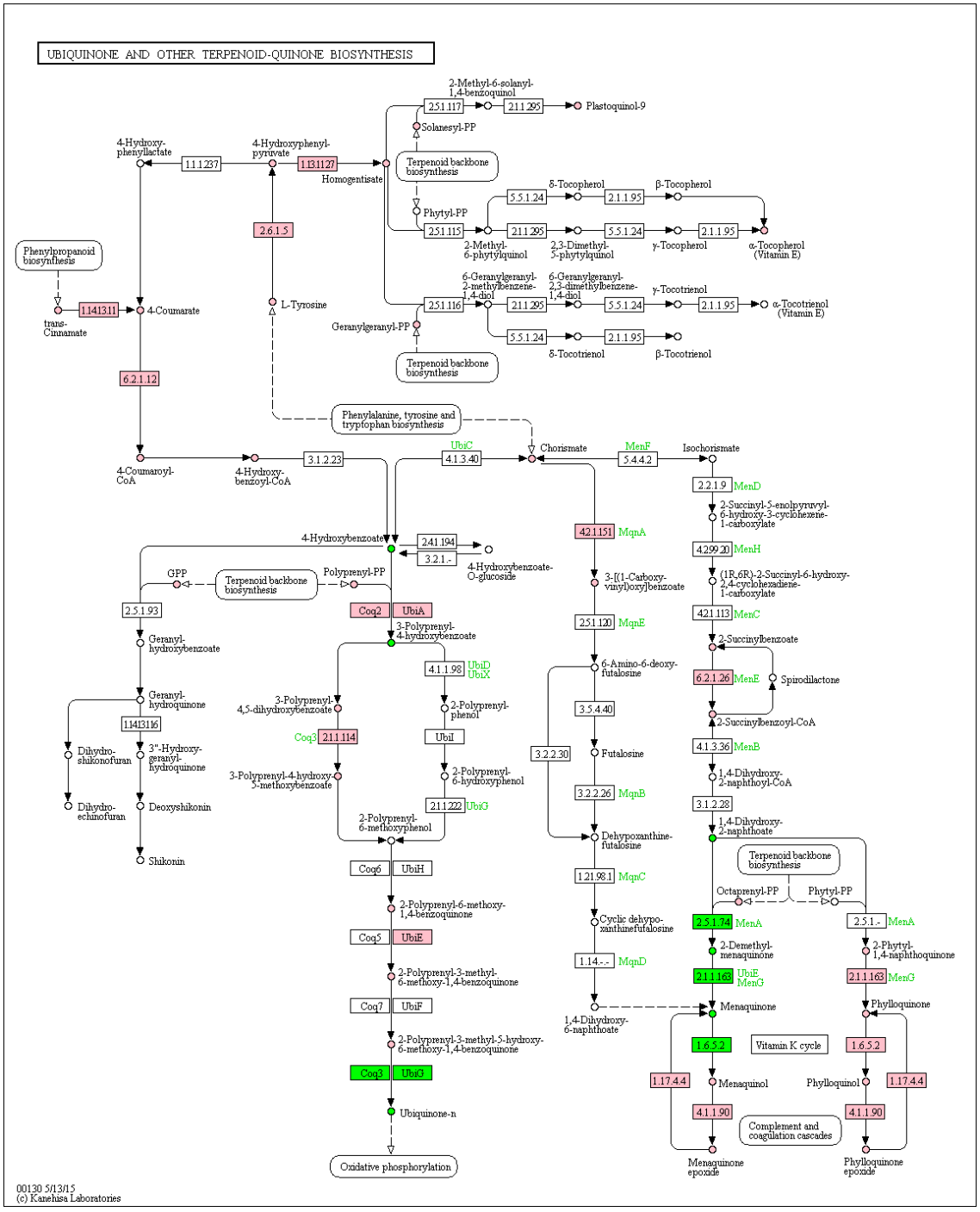


Carbon metabolism

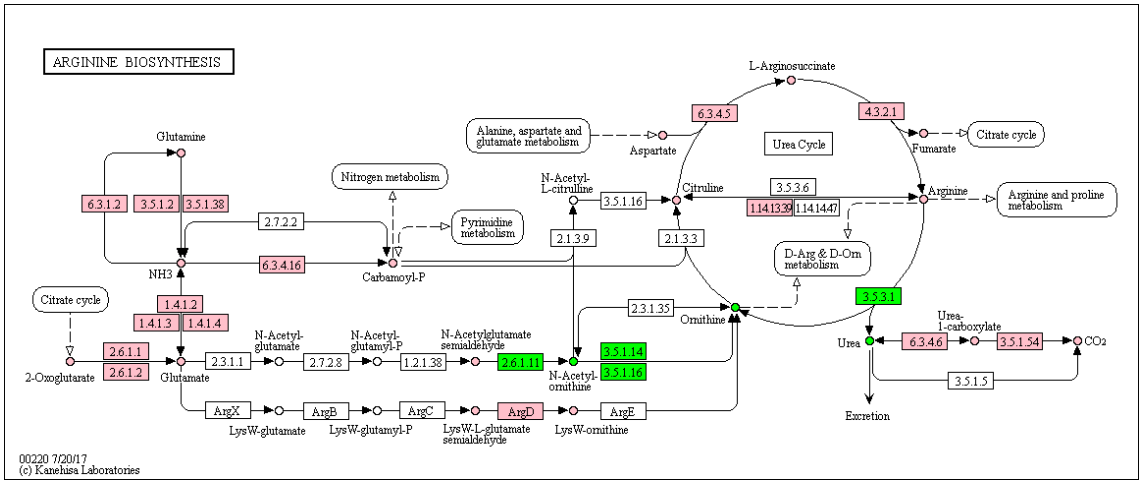


The figures below show KEGG pathways that are predicted to be affected by the Lead treatment (see section 5.3.2.2). Reactions and metabolites are coloured pink if they are present in the *D. magna* draft GWMR and green if they are in an identified AMBIENT active module for the Lead STRESSFLEA dataset

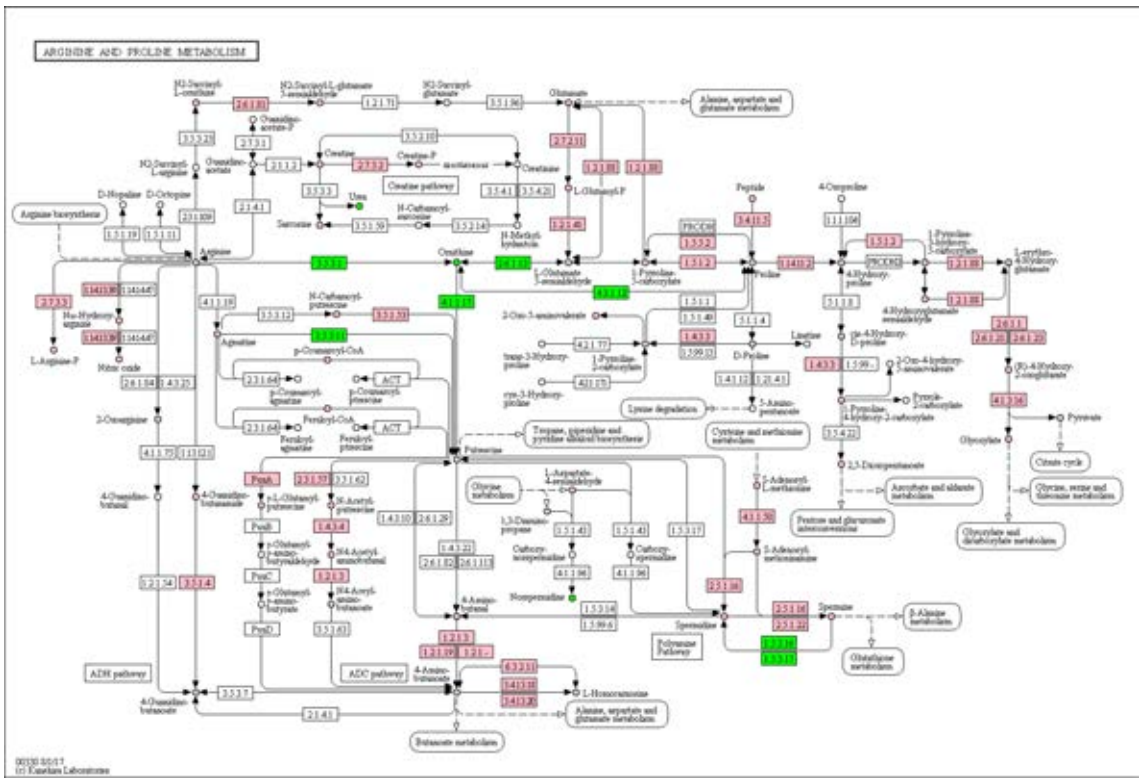
Ubiquinone and other terpenoid-quinone biosynthesis



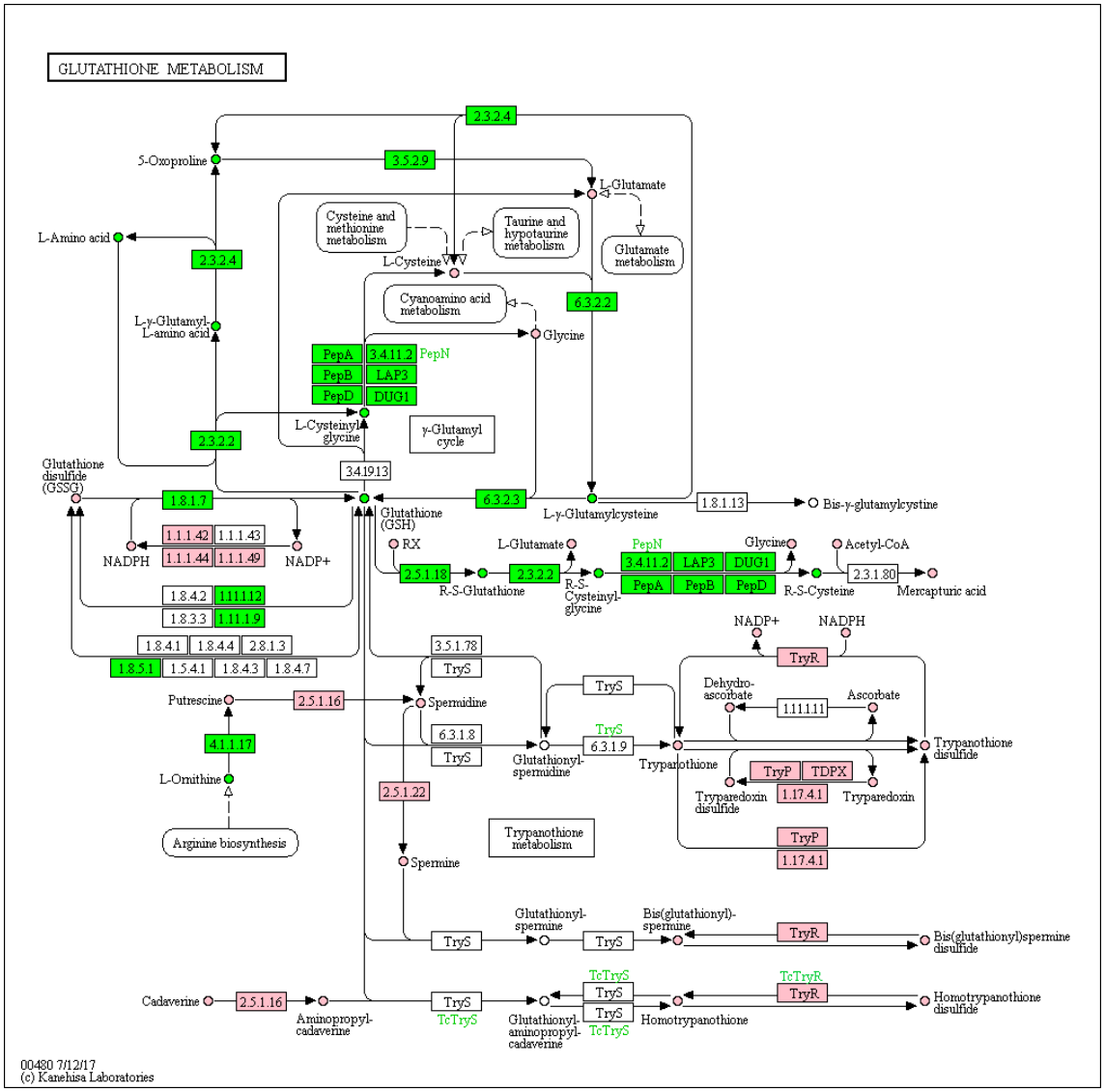
Arginine biosynthesis



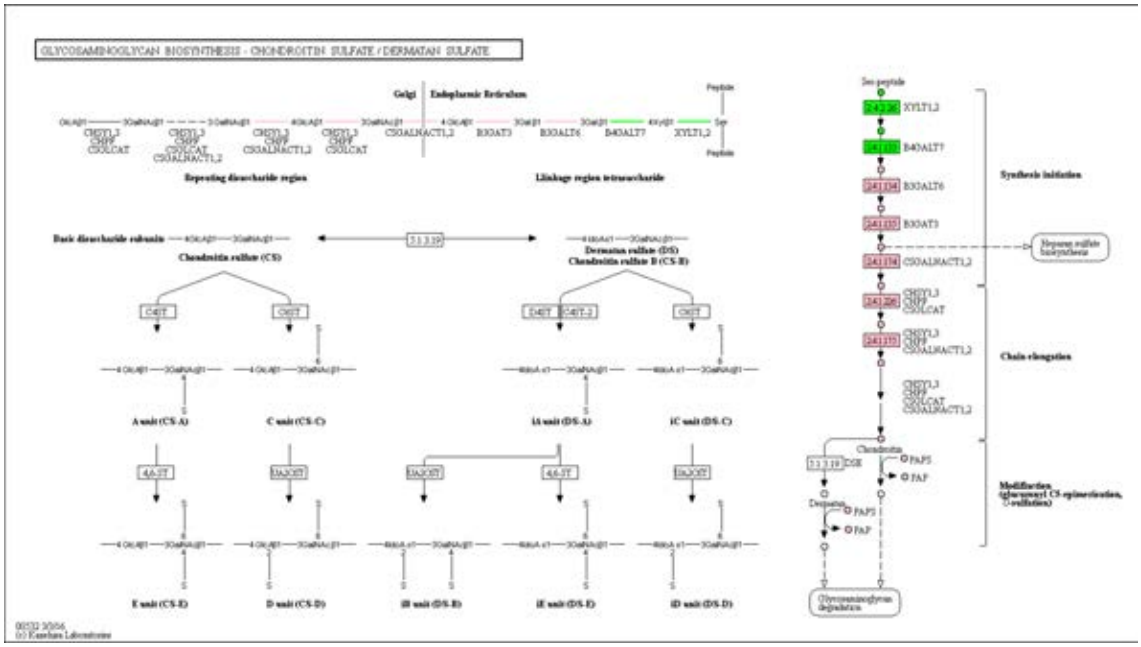
Arginine and proline metabolism



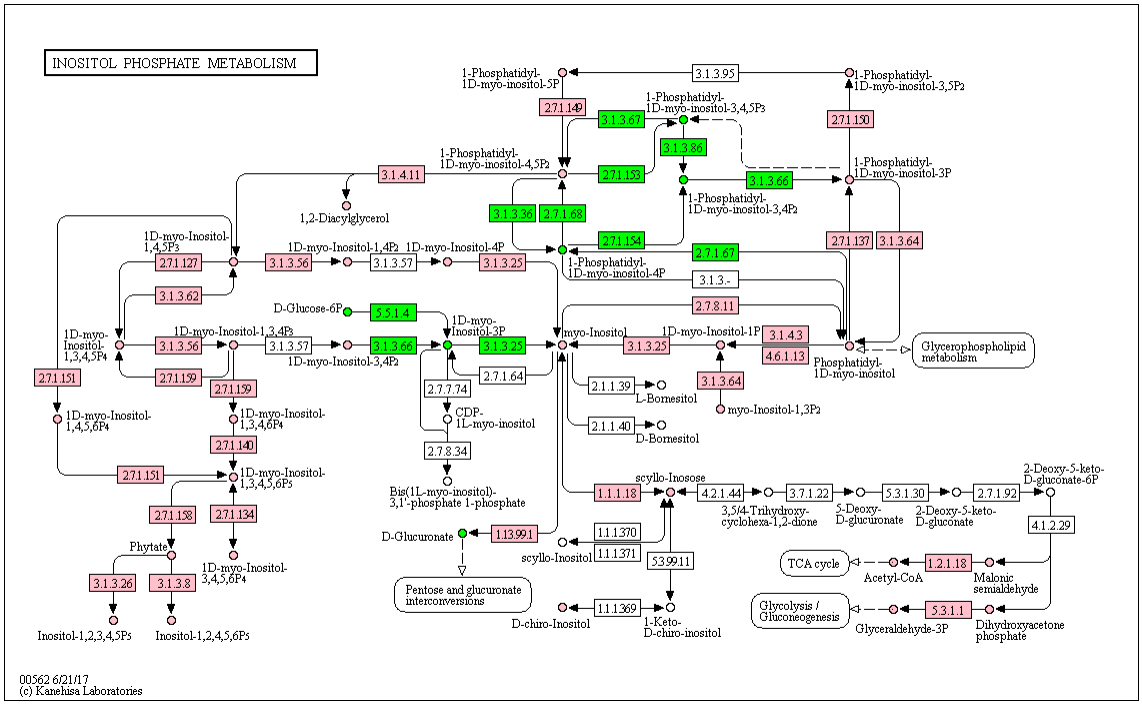
Glutathione metabolism



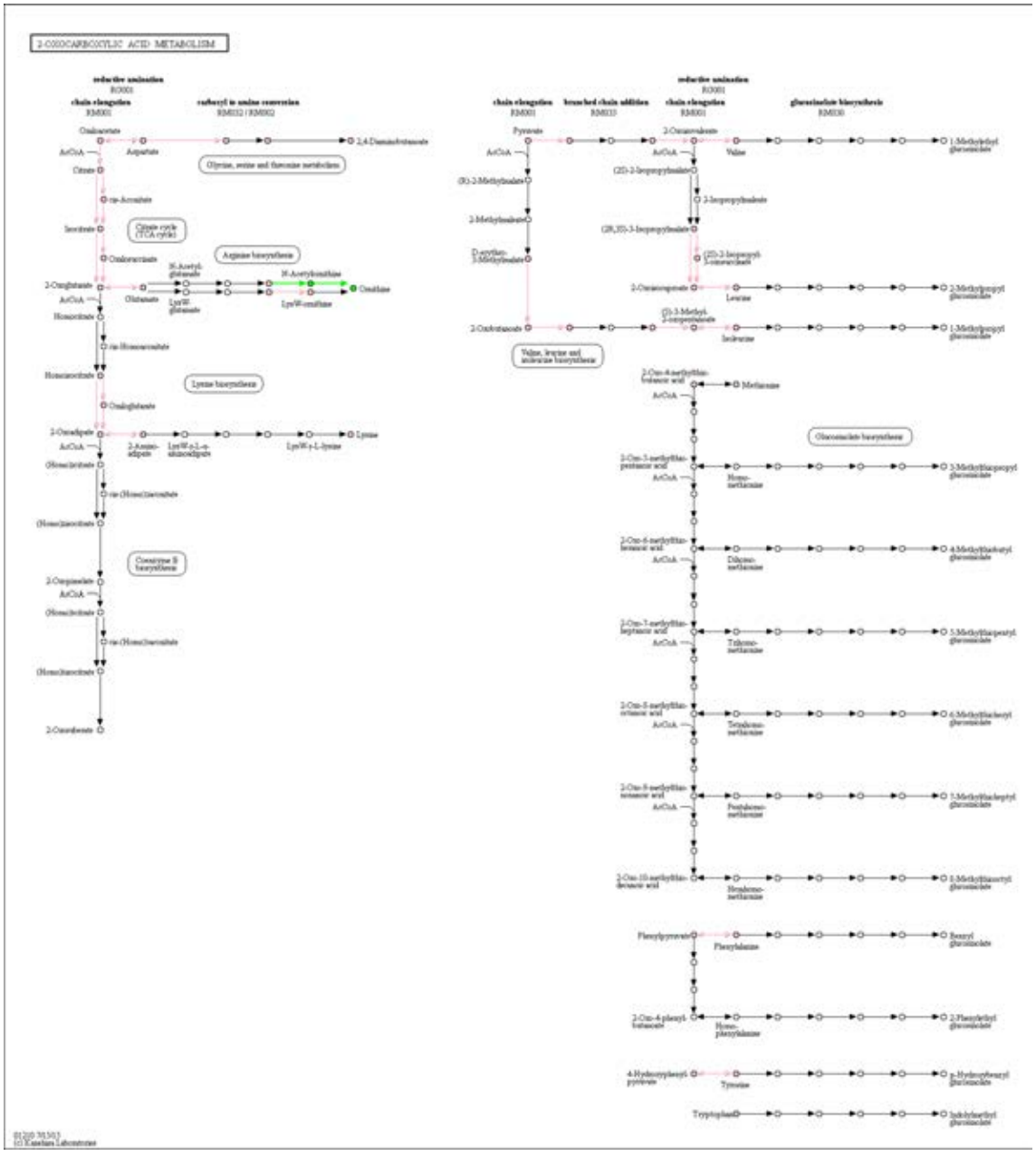
Glycosaminoglycan biosynthesis – Chondroitin sulfate / dermatan sulfate



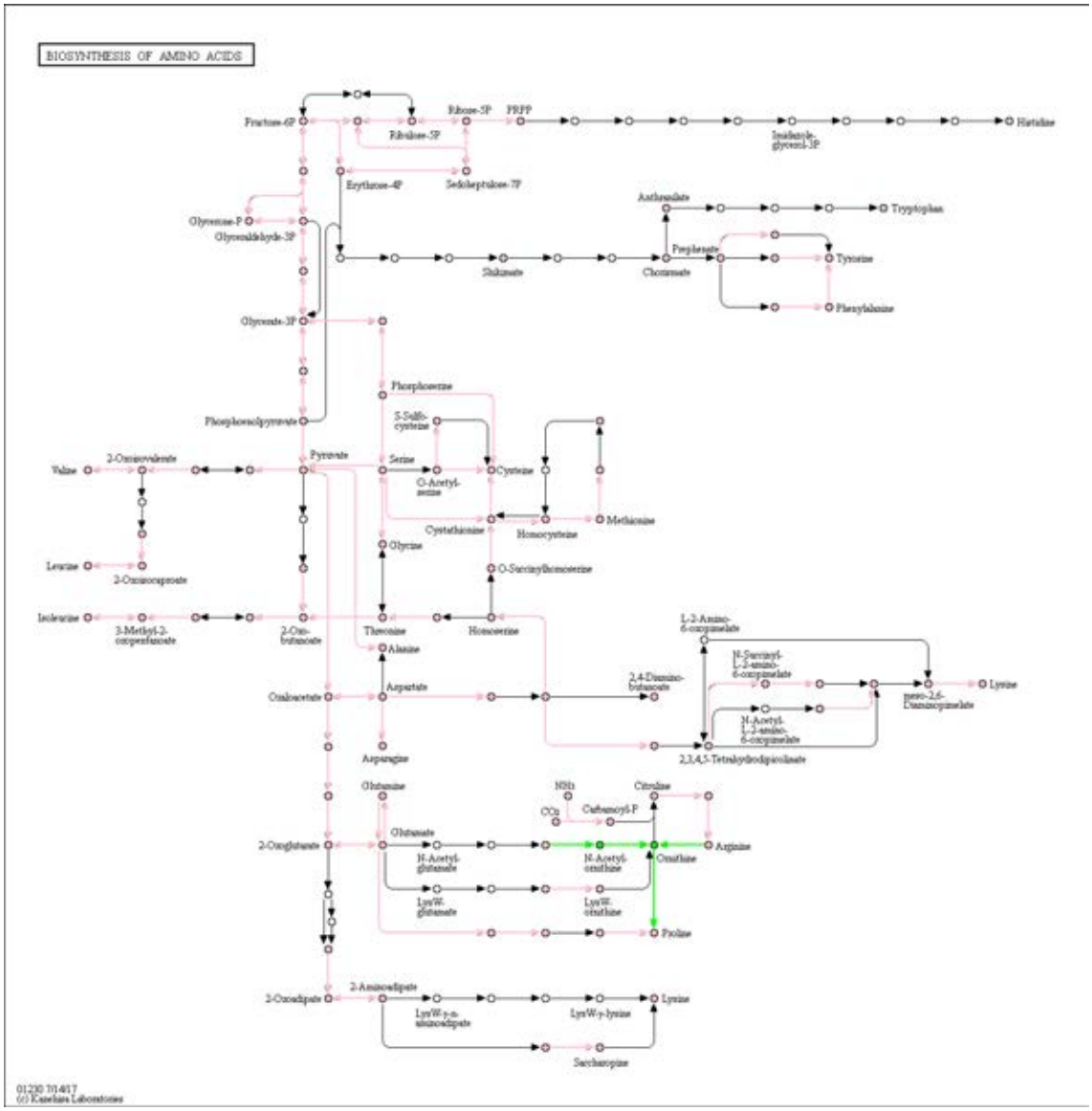
Inositol phosphate metabolism



2-Oxocarboxylic acid metabolism



Biosynthesis of amino acids



11. Appendix C – MUSCLE Paper 1

Bioinformatics, 31(6), 2015, 975–977
doi: 10.1093/bioinformatics/btu740
Advance Access Publication Date: 11 November 2014
Applications Note

Systems biology

MUSCLE: automated multi-objective evolutionary optimization of targeted LC-MS/MS analysis

James Bradbury[1], Grégory Genta-Jouve[2], J. William Allwood[2], Warwick B. Dunn[2], Royston Goodacre[3,4], Joshua D. Knowles[5], Shan He[1,*] and Mark R. Viant[2,*]

[1]School of Computer Science and [2]School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK, [3]Manchester Institute of Biotechnology, [4]School of Chemistry and [5]School of Computer Science, The University of Manchester, Manchester M13 9JD, UK

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on June 4, 2014; revised on October 20, 2014; accepted on November 5, 2014

Abstract

Summary: Developing liquid chromatography tandem mass spectrometry (LC-MS/MS) analyses of (bio)chemicals is both time consuming and challenging, largely because of the large number of LC and MS instrument parameters that need to be optimized. This bottleneck significantly impedes our ability to establish new (bio)analytical methods in fields such as pharmacology, metabolomics and pesticide research. We report the development of a multi-platform, user-friendly software tool MUSCLE (multi-platform unbiased optimization of spectrometry via closed-loop experimentation) for the robust and fully automated multi-objective optimization of targeted LC-MS/MS analysis. MUSCLE shortened the analysis times and increased the analytical sensitivities of targeted metabolite analysis, which was demonstrated on two different manufacturer's LC-MS/MS instruments.
Availability and implementation: Available at http://www.muscleproject.org.
Contact: info@muscleproject.org
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Liquid chromatography mass spectrometry (LC-MS) is widely used in analytical laboratories for measuring a range of (bio)chemicals and as the principal technology for metabolomics and proteomics. Developing new LC-MS methods and transferring existing methods between instruments and laboratories are time consuming and challenging, mostly because of the large number of LC and MS parameters that require optimization. Varying all these parameters systematically to optimize the analysis of selected chemicals is generally regarded as impossible because of the large search space.

Previously, a fully automated closed-loop strategy was reported that successfully optimized gas chromatography (GC)-MS and LC-MS methods for non-targeted metabolite analyses, resulting in increased analytical sensitivity (O'Hagan *et al.*, 2005, 2007; Zelena *et al.*, 2009). Although highlighting the value of closed-loop optimization in mass spectrometry, each of these implementations was for a specific manufacturer's analytical platform. Extending this to further instruments would have required extensive reprogramming, therefore significantly limiting the deployability of this approach.

Here, we present MUSCLE (multi-platform unbiased optimization of spectrometry via closed-loop experimentation), a software tool for robust and fully automated optimization of targeted LC-MS/MS analyses. MUSCLE is instrument-manufacturer independent and requires no knowledge of computer programming to operate. Using a process called visual scripting, users create a set of configuration scripts, which instruct MUSCLE how to operate an LC-MS/MS. These scripts can be imported and exported from MUSCLE to facilitate sharing and re-use across laboratories.

We demonstrate MUSCLE by optimizing the analyses of six steroids on two different manufacturer's LC-MS/MS instruments.

2 Methods and implementation

MUSCLE is a stand-alone desktop application and has been tested on Windows XP, 7 and 8; see system diagram in Supplementary Figure S1. User-defined visual scripts imitate the keyboard and mouse commands that an analyst would use to manually change parameters and launch an LC-MS/MS analysis, enabling MUSCLE to control multiple LC and MS parameters on any instrument (Section 2.1). An experiment configuration contains all the information MUSCLE requires to run an automated optimization, including the user-defined LC and MS parameters to optimize and the (bio)chemicals to be analysed (Section 2.2). A multi-objective genetic algorithm (GA) optimizes the values of the LC and MS parameters, based upon the fitness of user-defined objective functions that measure, e.g. analytical sensitivity and analysis time (Section 2.3). On completion, MUSCLE presents a set of best solutions that the analyst can inspect and then select their preferred solution, a set of LC/MS parameters achieving both fast and sensitive analysis.

2.1 Visual scripting

Visual scripting enables direct visual references to be made to objects displayed on the screen, e.g. a 'File' menu item and allows MUSCLE to mimic the keyboard and mouse actions that a user would make. Here we use the Sikuli Java library (Yeh *et al.*, 2009), providing a powerful and flexible API to allow users to create visual scripts that can: click/double click on selected objects on the screen, enter text into text fields and press selected keyboard keys, e.g. Enter. These visual scripts can be saved and later reused or modified for reuse on different analytical instruments and can be shared between laboratories using an import and export function.

2.2 Experiment configuration

Once visual scripts are set up to control a particular instrument, an experiment configuration can be created, which contains all the information MUSCLE requires for an automated optimization study. This includes: (i) details of the target list of (bio)chemicals to be analysed, including *m/z* values of the parent and fragment ions (e.g. Supplementary Fig. S2 and Table S1), (ii) settings for the GA including the user-defined objective functions (see Section 2.3 and Supplementary Table S2) and (iii) user-defined list of LC and MS parameters to be optimized, where each parameter has an associated visual script and minimum, maximum and step size values (e.g. Supplementary Table S3). Further details of the experiment configuration are provided in Supplementary Material.

2.3 Closed-loop evolutionary optimization

Closed-loop evolutionary optimization is a probabilistic search heuristic, whereby potential solutions are evaluated by conducting physical experiments (Knowles, 2009), which in the case of MUSCLE corresponds to LC-MS/MS analyses. Each solution represents a set of control parameters for the LC-MS instrument and is generated using a GA. Typically, GAs evaluate tens of thousands of solutions *in silico* during an optimization process. Because of the time constraints on evaluating each solution in an LC-MS/MS study, closed-loop optimization requires the GA to perform well when limited to just a few tens or hundreds of evaluations. To evaluate each solution, a fitness value is calculated for each of the objectives, where each objective measures the quality of the LC-MS/MS spectra obtained using the selected instrument settings. For targeted LC-MS/MS analysis, the user-selected objectives include (i) minimizing the analysis time (measured as the retention time of the last eluting target analyte, not the total analysis time); (ii) maximizing the number of analytes detected from the target list and (iii) maximizing the total peak area of these analytes. Fitness values are calculated based on the results of a custom peak detection algorithm (see Supplementary Material), which processes mzML files (Supplementary Fig. S1). This enables MUSCLE to analyse results from any LC-MS/MS instrument following conversion of the vendor specific data format to mzML. Because the three objectives are in conflict, a multi-objective GA must be used, which can efficiently find a set of Pareto optimal solutions. Typically, a large number of optimization experiments are required to achieve a highly optimized search method, which due to cost implications of conducting LC-MS/MS analyses was not feasible. We therefore opted to use the PESA-II multi-objective GA (Corne *et al.*, 2001) as it is widely used, and we are familiar with configuring this algorithm for the optimization of mass spectrometry analyses. The Java library implementation of the algorithm, jMetal (Durillo *et al.*, 2010), was used in this case. The values of each LC and MS parameter in the first *n* runs (where *n* is user defined) are chosen randomly. For each subsequent run, the GA decides the LC and MS parameters based upon the evaluation of previous LC-MS/MS analyses, favouring parameters that produced high fitness values. The GA maintains a set of the best solutions in an archive set and from these solutions decides on the next set of parameters by applying selection, crossover and mutation operators. If suboptimal parameters are selected by the GA, a low-quality chromatogram will result with low fitness values, which will not be added to the archive set.

The user selects the maximum number of runs to be performed, which fixes the overall time and cost of the optimization. Because of the limited number of runs, the optimization algorithm will never realistically reach the global optima but instead has a high likelihood of converging towards a local optima. The user is shown real-time results of the optimization and if the convergence towards an optimum set of parameters seems to be complete, they have the ability to pause or completely stop the optimization.

3 Results and discussion

We have demonstrated MUSCLE in two common laboratory scenarios, using two manufacturers' LC-MS/MS instruments and associated software, to optimize the targeted analysis of a mixture of six steroids (Supplementary Fig. S2). First, we used MUSCLE to improve an LC-MS/MS analysis that had previously been optimized manually by an experienced analytical chemist, using a Thermo Scientific UHPLC Ultimate 3000 TSQ Vantage running under Xcalibur software V2.0.7. Second, we transferred this manually optimized method from the Thermo Scientific instrument to a Waters ACQUITY UPLC Xevo TQ LC-MS/MS running under MassLynx software V4.1 and used MUSCLE to re-optimize the LC and MS parameters.

In the first study, the user selected minimum and maximum values and step sizes for each of 10 LC and MS parameters to be optimized (Supplementary Table S3 and Fig. S3) along with settings for the GA (Supplementary Table S2). Following an ~48-h optimization, comprising 200 LC-MS/MS analyses, this fully automated approach discovered an improved set of parameters that provided a faster (34.5%) and more sensitive (10.0%) analysis than achieved manually (Supplementary Table S4). Figure 1a shows the final

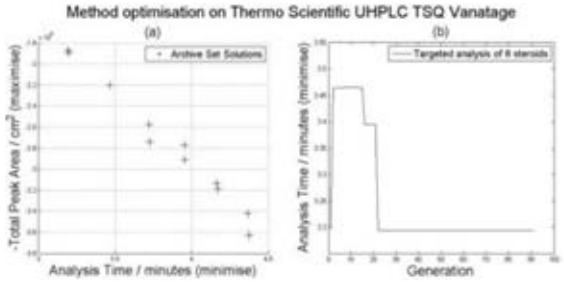


Fig. 1. (**a**) Pareto front. The total peak area axis has been reversed for readability, each cross represents a solution with the corresponding objective values. (**b**) Generation-by-generation lowest run time in the archive set (considering the set of best solutions for which six out of six peaks are detected). The first generation was 20 randomized runs and each subsequent generation consisted of two runs

Pareto front after 200 analyses. Figure 1b shows how the LC analysis time decreases through the optimization process, considering the set of best solutions for which six out of six steroids are detected, plateauing around generation 20.

Following transfer of the manually optimized method from the Thermo Scientific to Waters instrument, the minimum and maximum values and step sizes for nine LC and MS parameters were selected (Supplementary Table S5) and then optimized during an ~48-h fully automated LC-MS/MS study. Again, MUSCLE was able to discover an improved set of parameters that provided a faster (18.5%) and much more sensitive (104%) analysis (Supplementary Table S6). Supplementary Figure S4a shows the final Pareto front after 200 analyses. Supplementary Figure S4b shows how the total peak area increases through the optimization.

One limitation of MUSCLE is that an optimization is prone to finding local rather than global optimum solutions. Also the convergence of some optimizations may plateau before the maximum number of runs has been reached. To combat this, the user has the ability to view the results of the optimization generation by generation, and if they feel that MUSCLE is no longer improving the quality of the analysis, the optimization can be paused or stopped completely. A further limitation is the peak detection procedure, which is based on a relatively simple algorithm that is designed to work for data derived from a range of mass spectrometers, as described in Section 1.1.3 (Supplementary Material). However, MUSCLE has been programmed to enable the user to add alternative peak detection algorithms from a drop-down box, should the current implementation not work well for a particular dataset.

In conclusion, MUSCLE shortened the analysis times and increased the analytical sensitivities of the targeted analysis of multiple steroids on two manufacturer's LC-MS/MS instruments in a fully automated manner and is anticipated to benefit several fields including pharmacology, metabolomics and proteomics.

Acknowledgement

We thank Steve O'Hagan for technical advice.

Funding

This work was supported by UK Biotechnology and Biological Sciences Research Council [BB/I024085/1].

Conflict of interest: none declared.

References

Corne,D. *et al.* (2001) PESA-II: region-based selection in evolutionary multiobjective optimization. In: *Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco - July 2001*, pp. 283–290. Morgan Kaufmann Publishers, San Francisco.

Durillo,J.J. and Nebro,A.J. (2011) jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software*, **42**, 760–771.

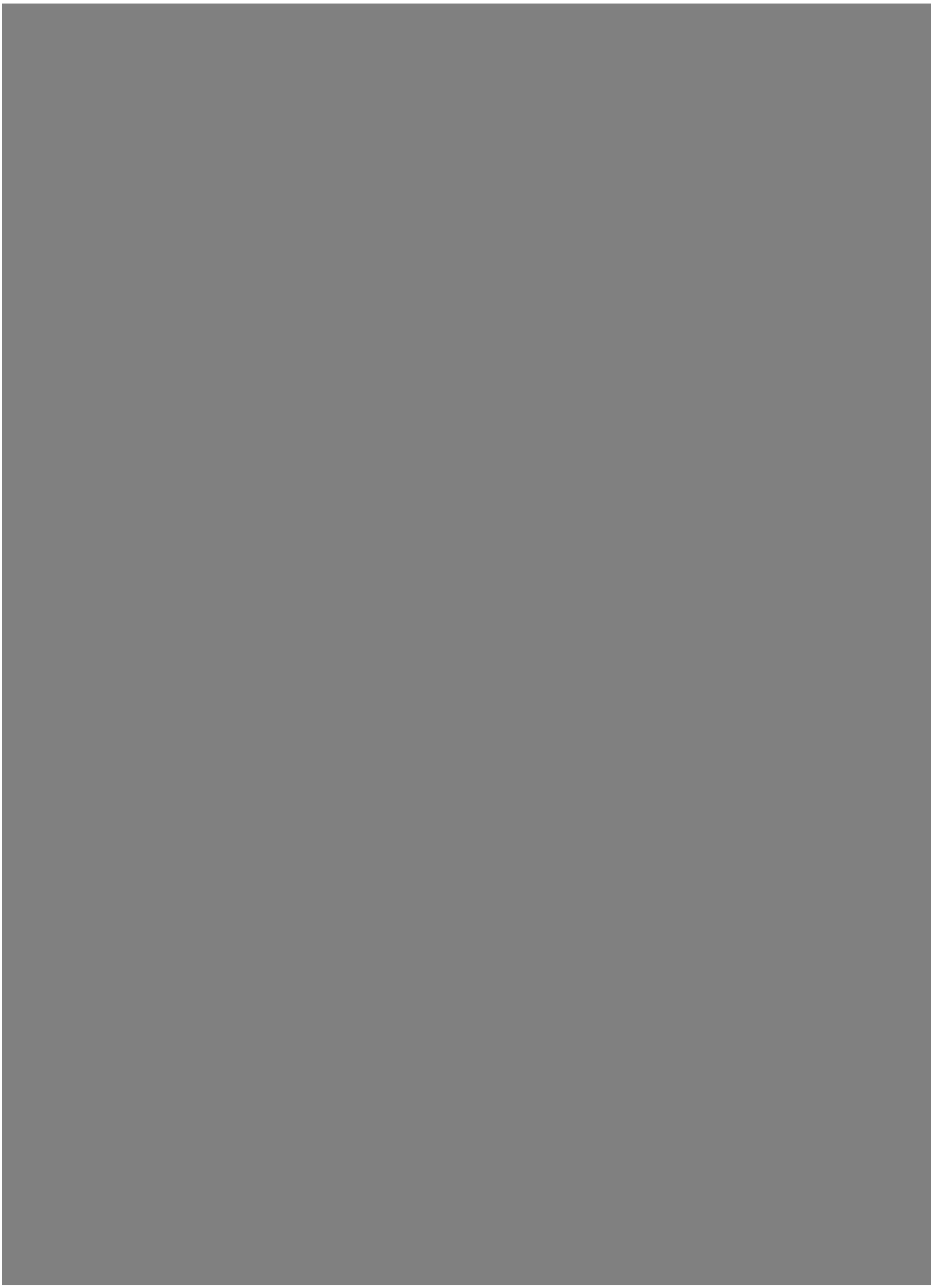
Knowles,J. (2009) Closed-loop evolutionary multiobjective optimization. *IEEE Comput. Intell. Mag.*, **4**, 77–91.

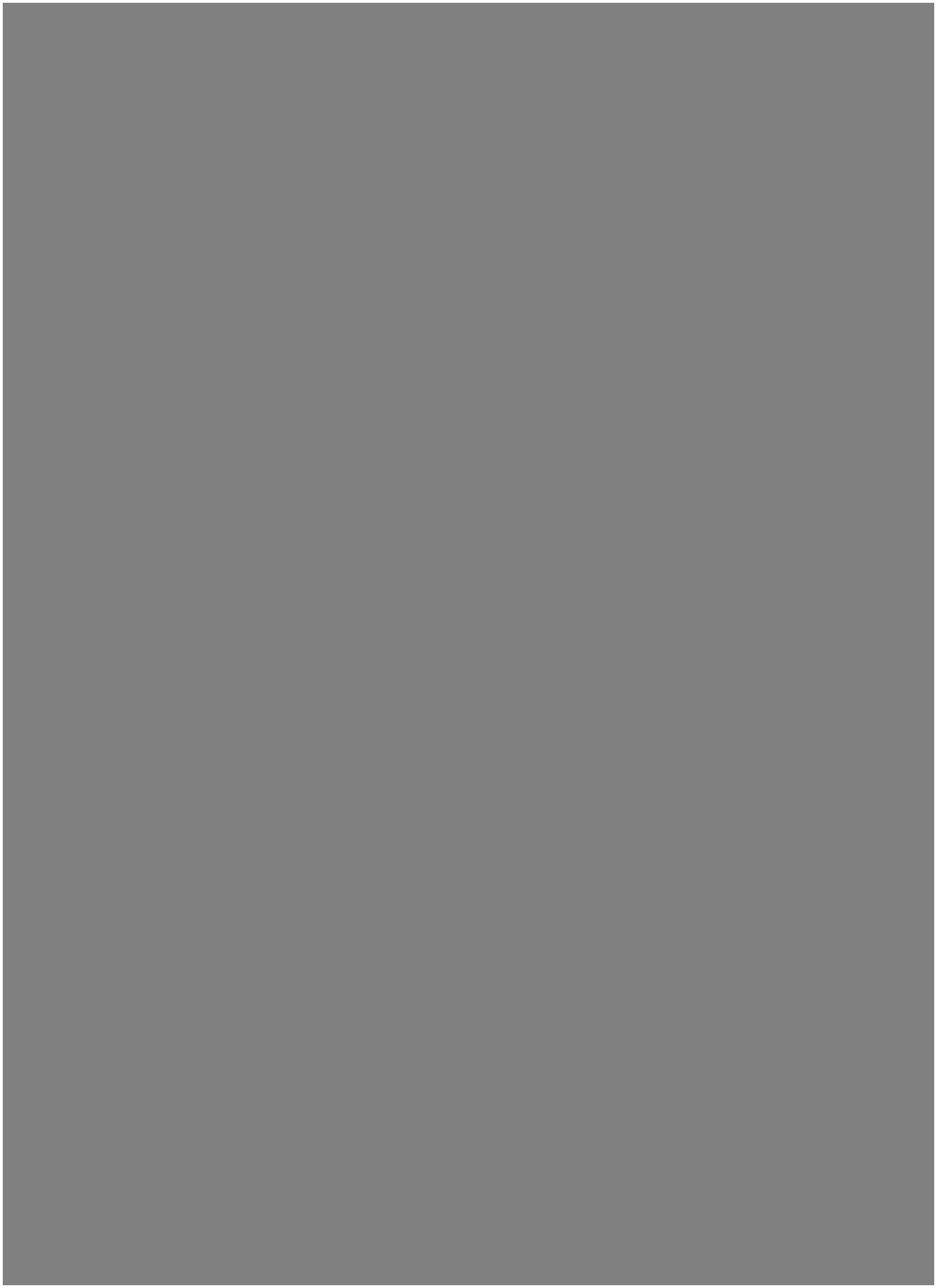
O'Hagan,S. *et al.* (2005) Closed-loop, multiobjective optimization of analytical instrumentation: gas chromatography/time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. *Anal. Chem.*, **77**, 290–303.

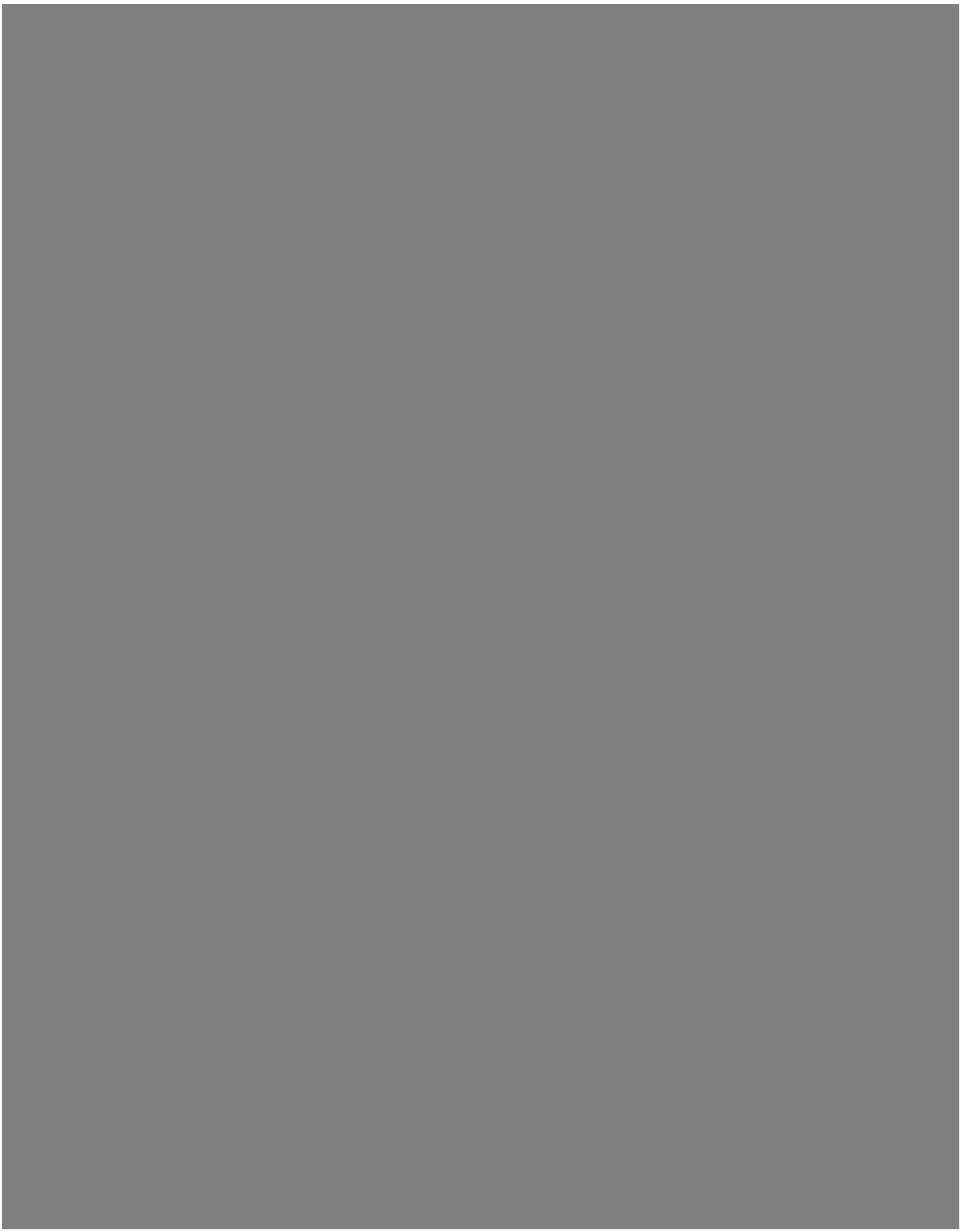
O'Hagan,S. *et al.* (2007) Closed-loop, multiobjective optimization of two-dimensional gas chromatography/mass spectrometry for serum metabolomics. *Anal. Chem.*, **79**, 464–476.

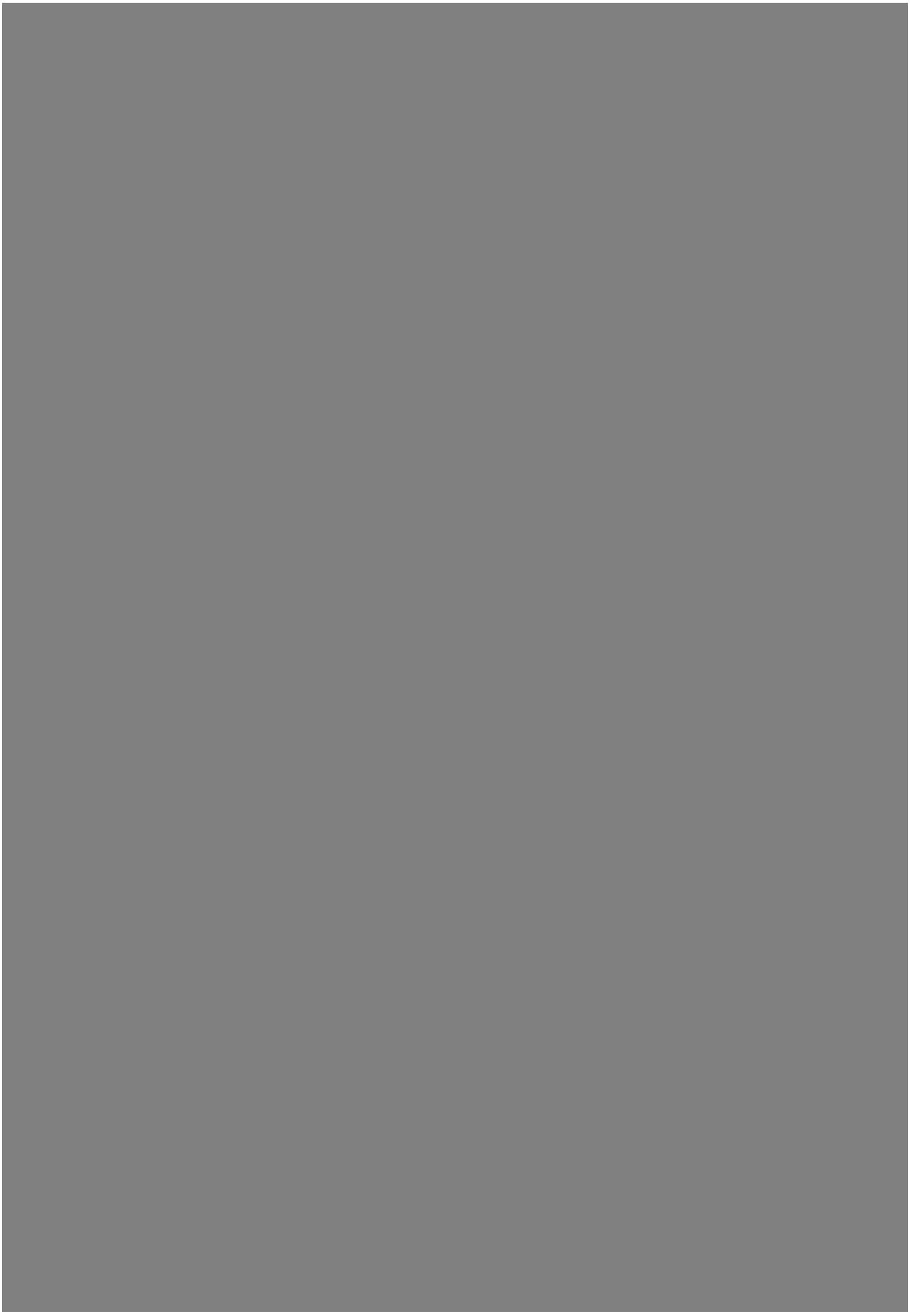
Yeh,T. *et al.* (2009) Sikuli: using GUI screenshots for search and automation. In: *Proceedings of the 22nd annual ACM symposium on User interface software and technology, Victoria, BC, Canada, October 4–7, 2009*, pp. 183–192. ACM, New York.

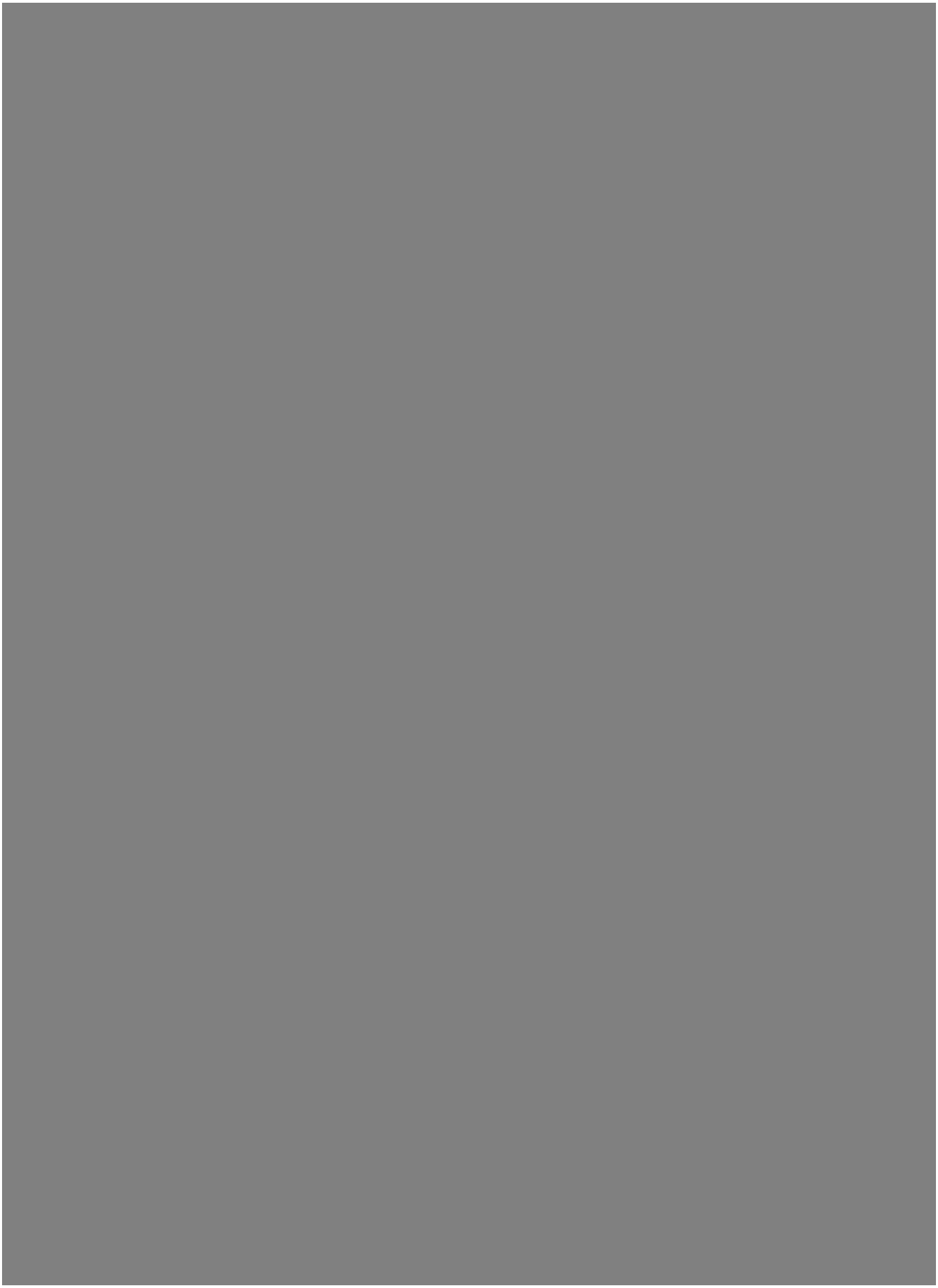
Zelena,E. *et al.* (2009) Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal. Chem.*, **81**, 1357–1364.

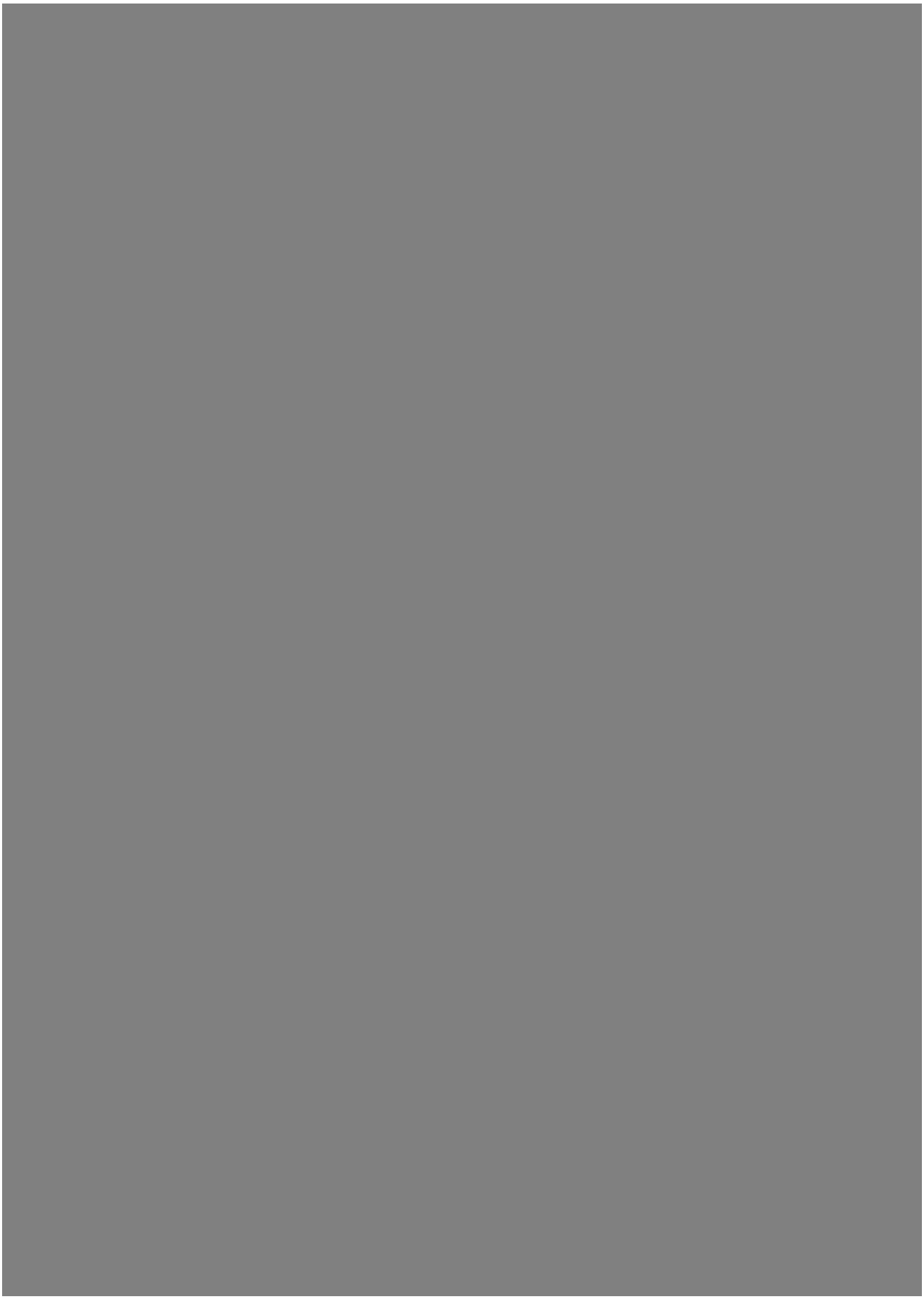


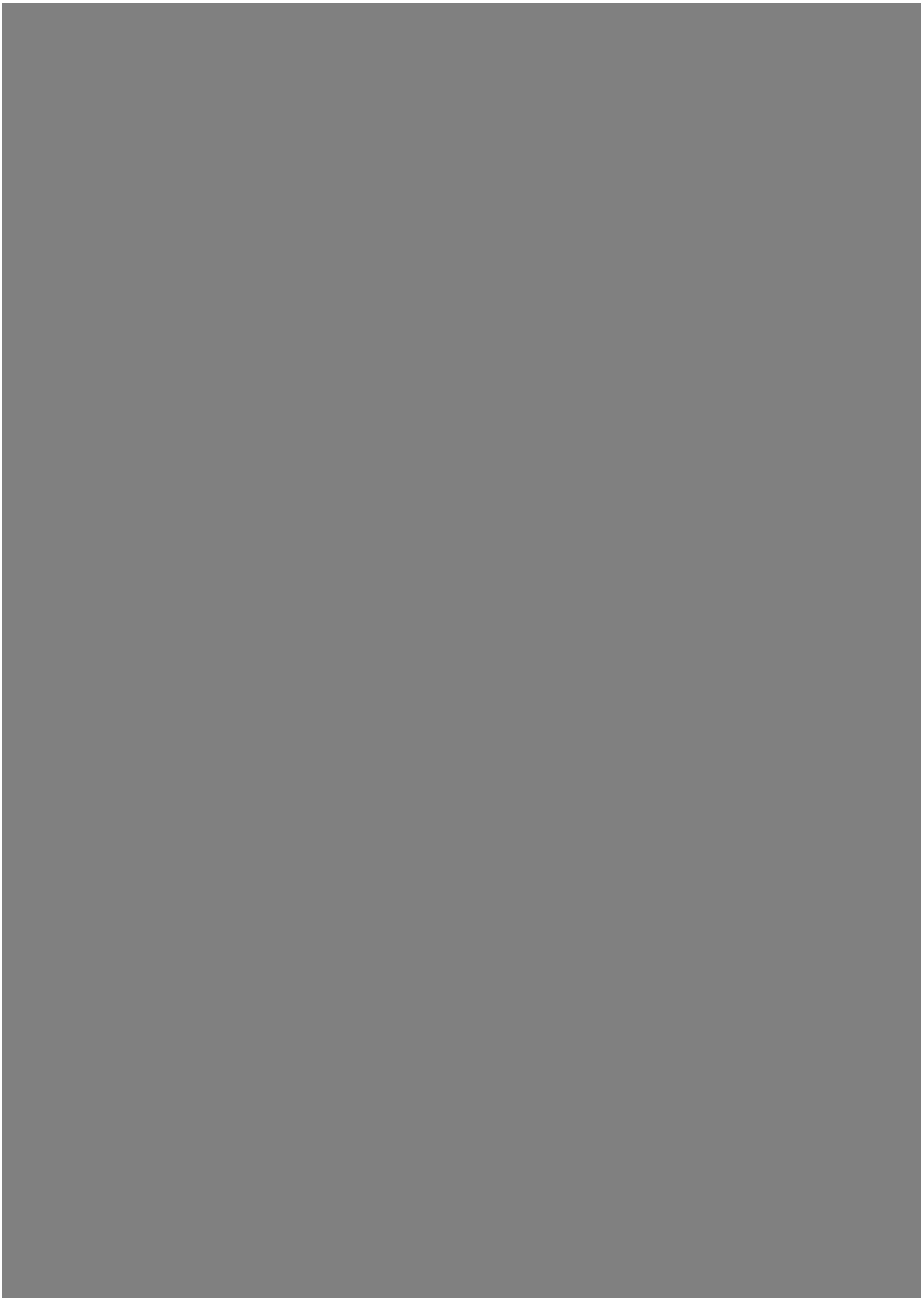


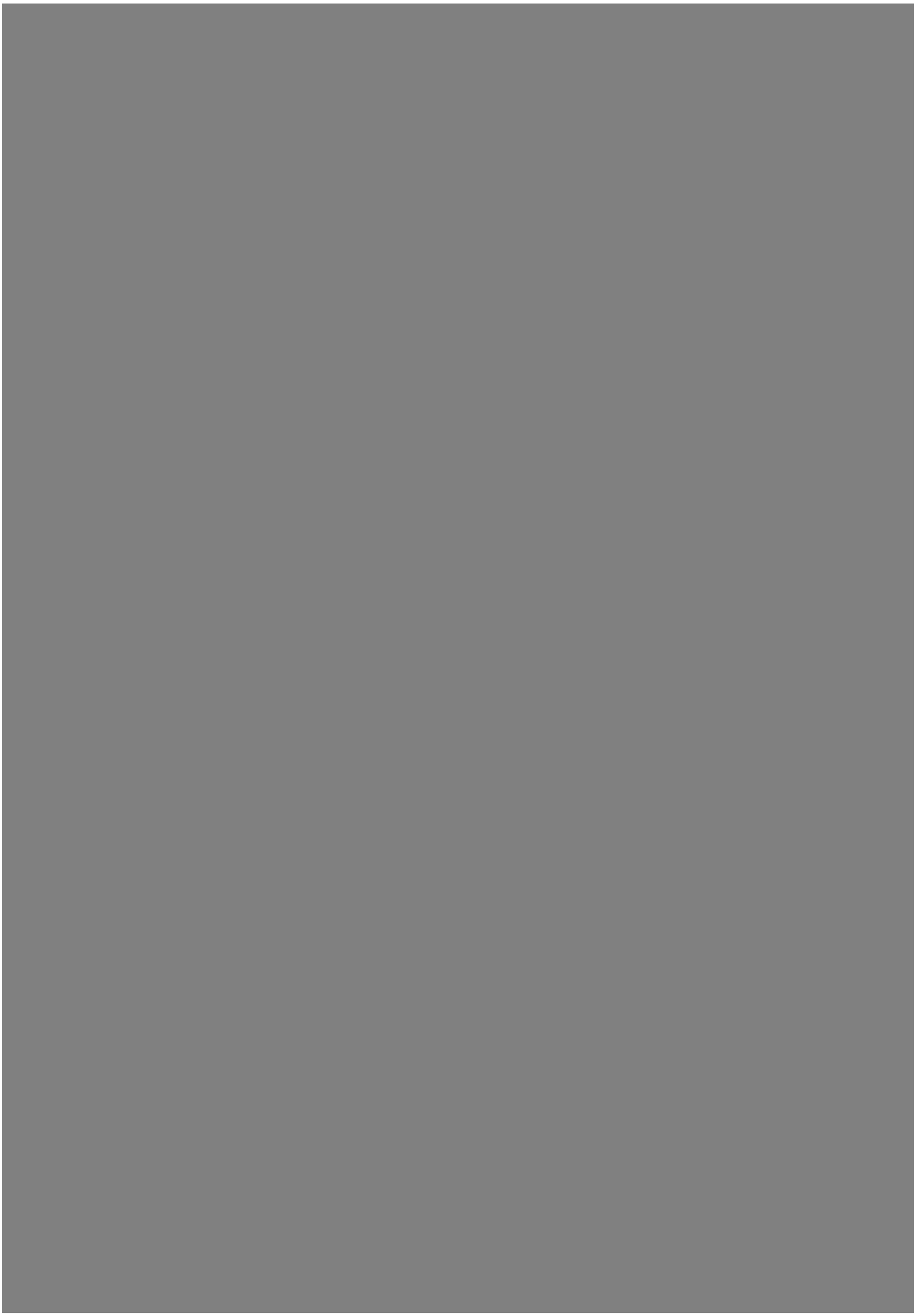






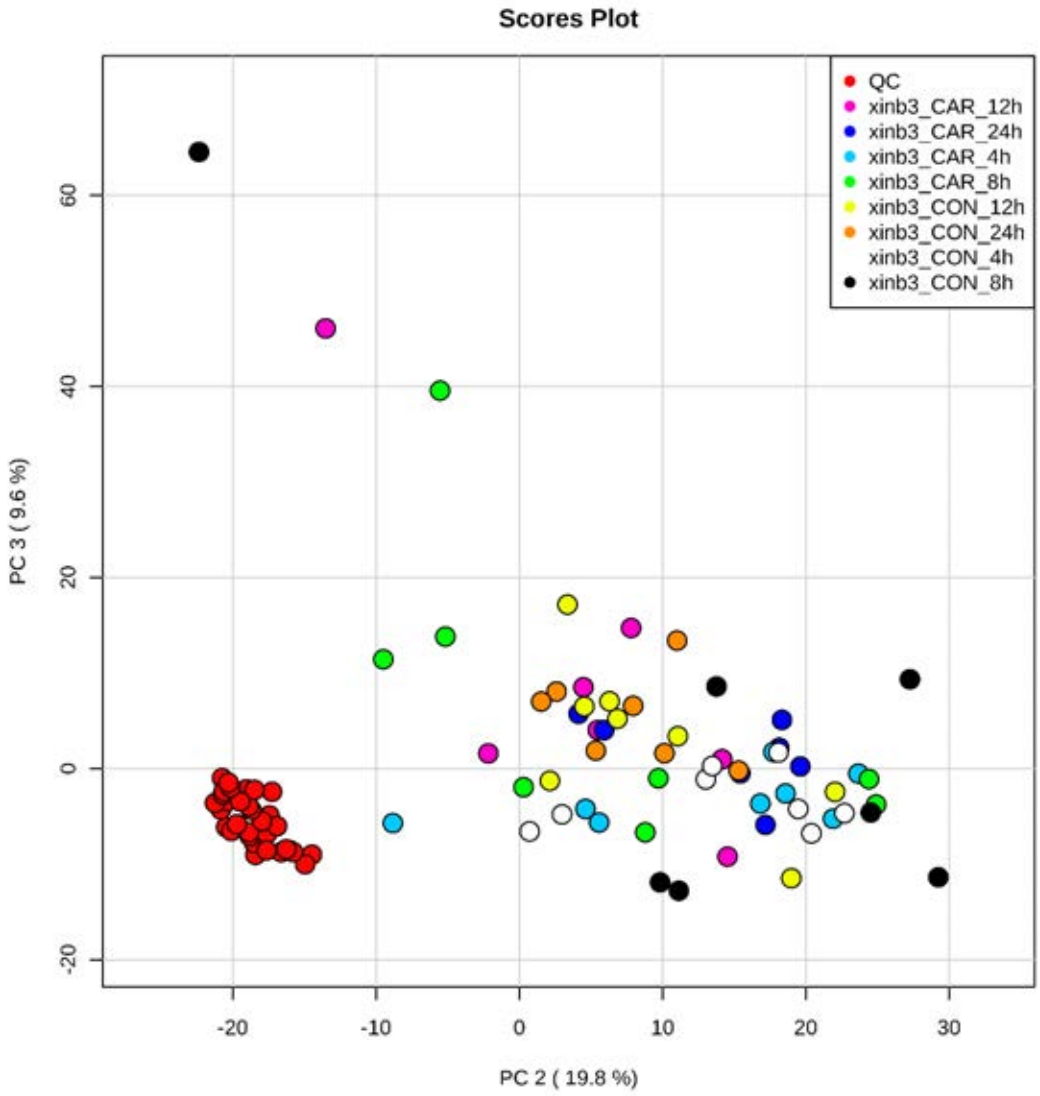




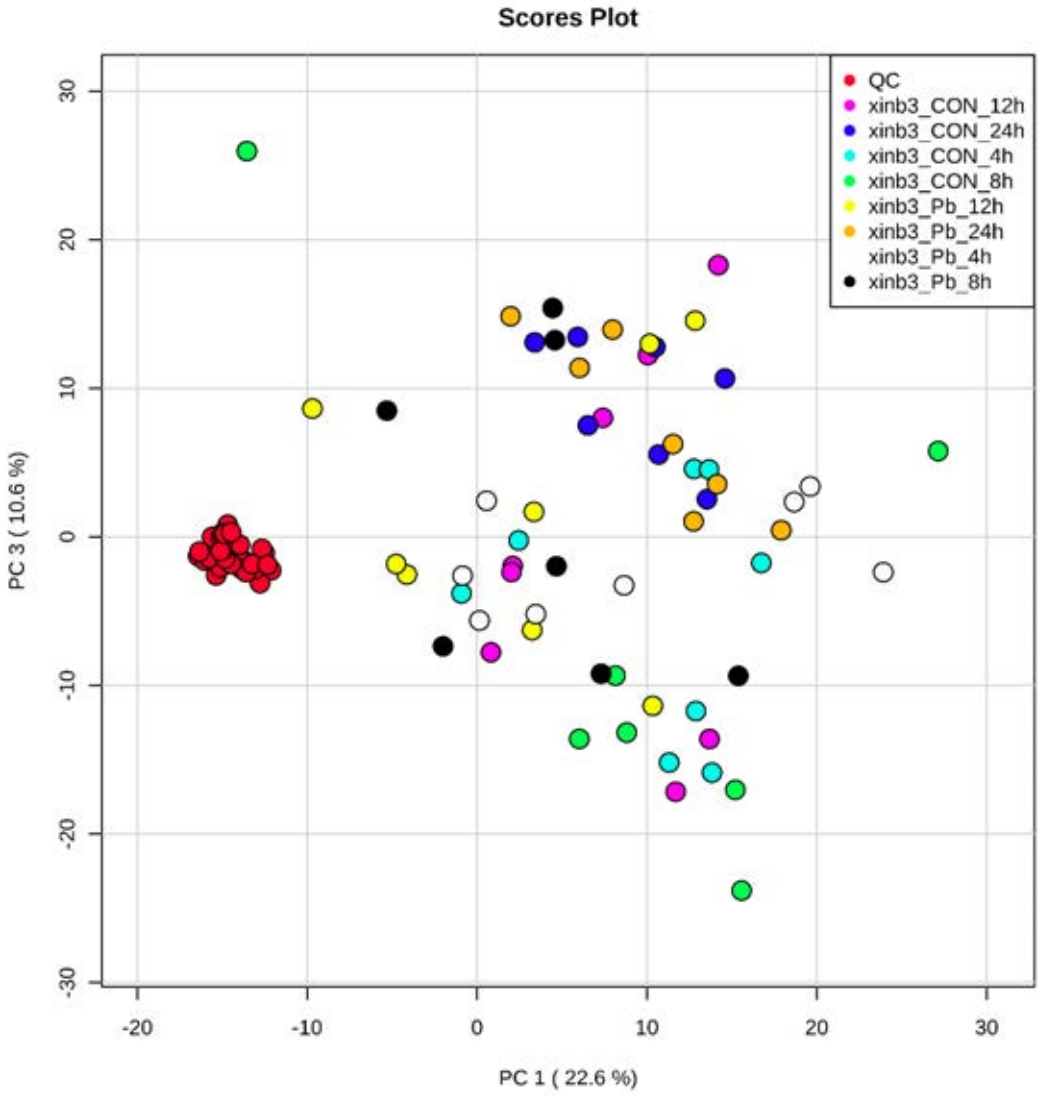


13. Appendix E – Statistical plots

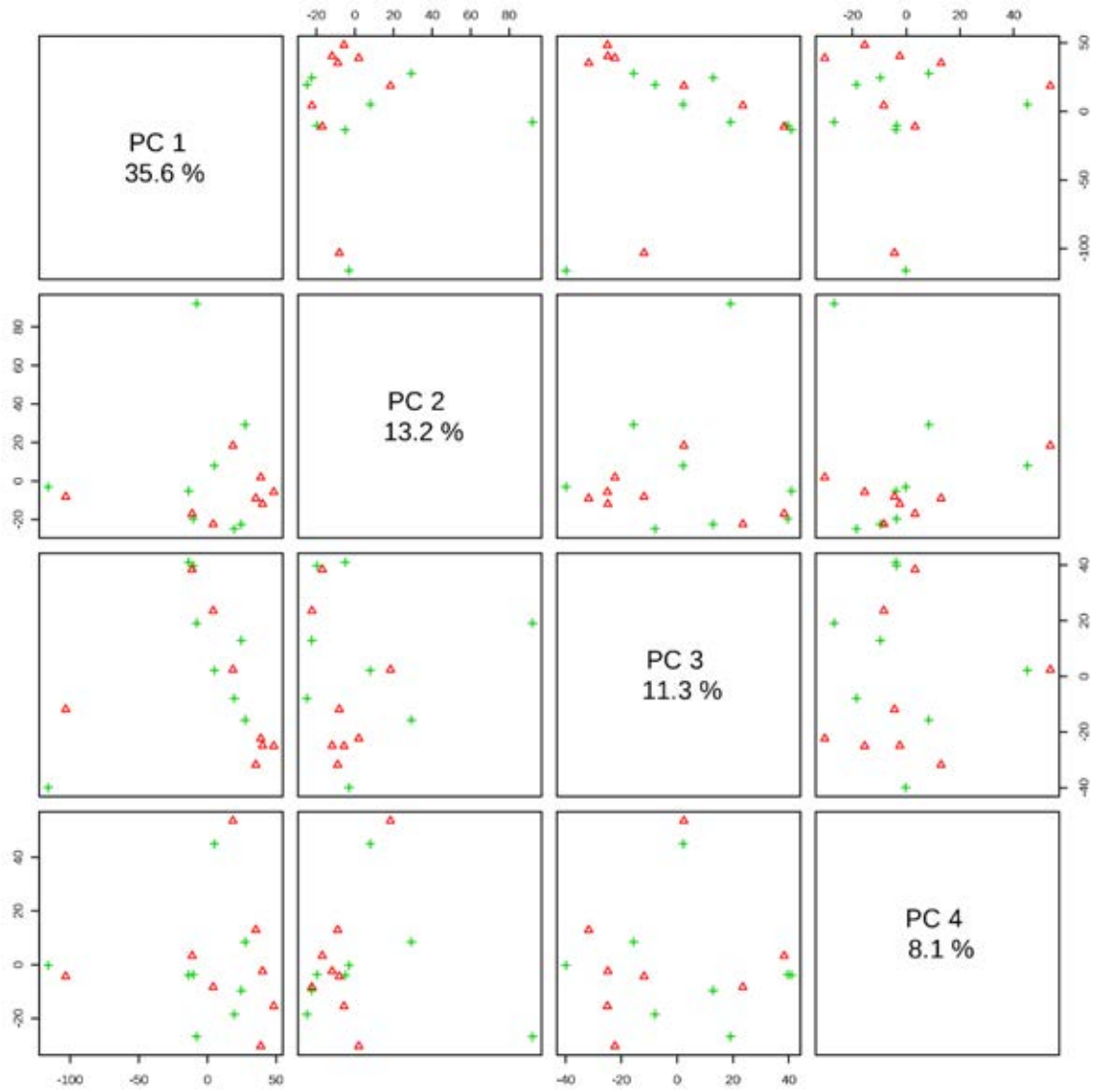
13.1. Carbaryl treatment PCA plots



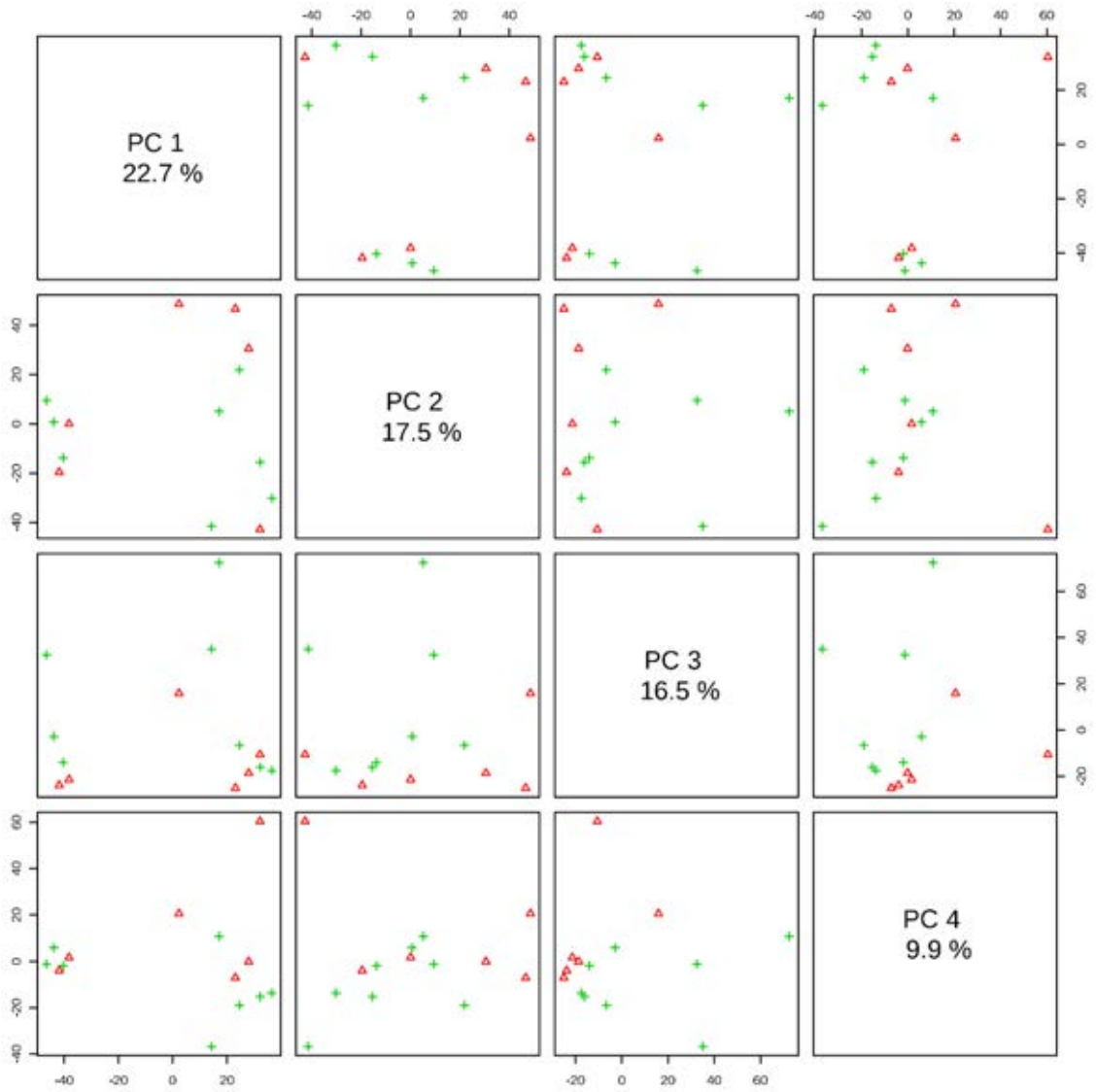
PC2 vs PC3 PCA scores plot of all Carbaryl and Control groups.



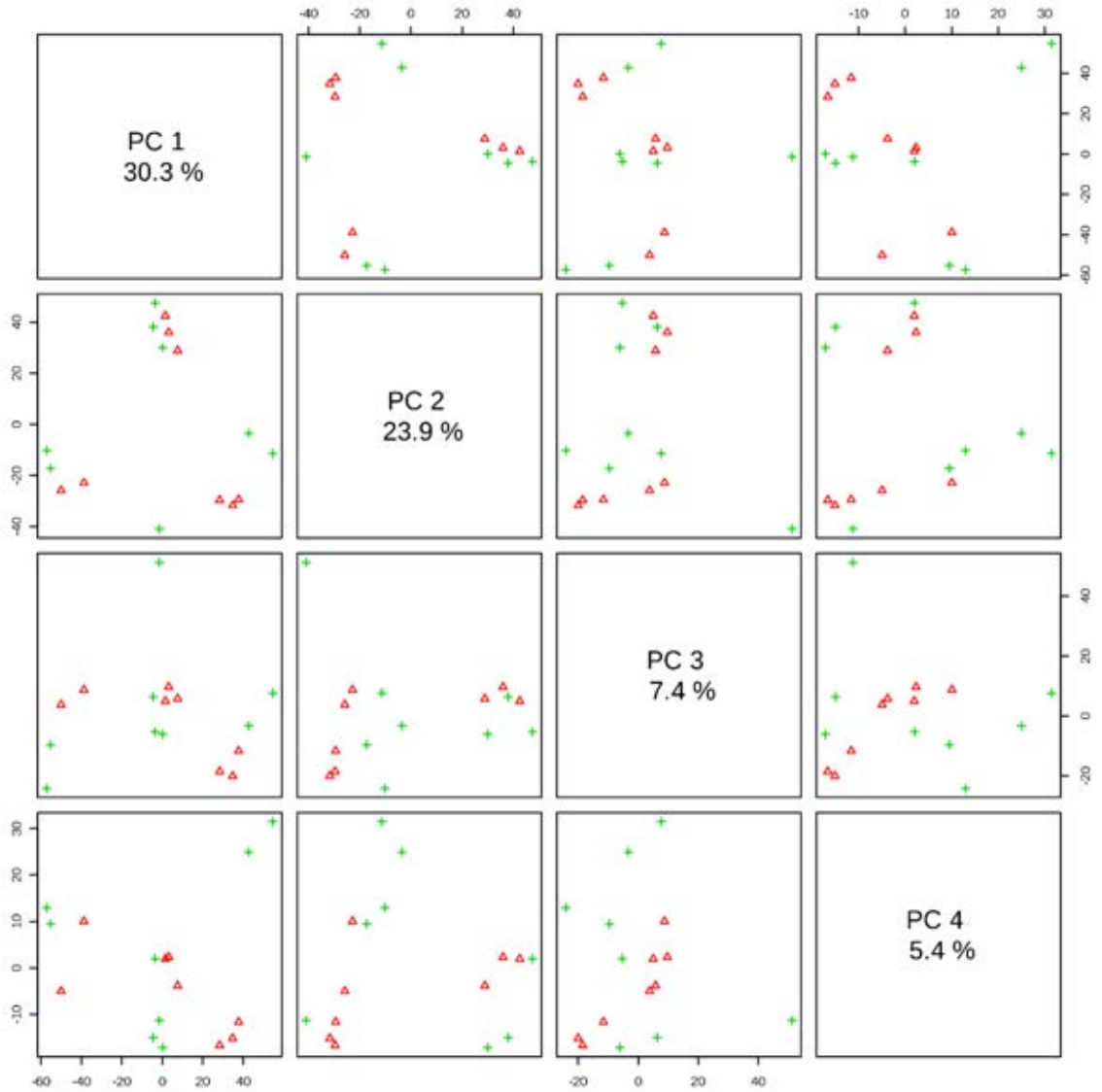
PC1 vs PC3 PCA scores plot of all Carbaryl and Control groups.



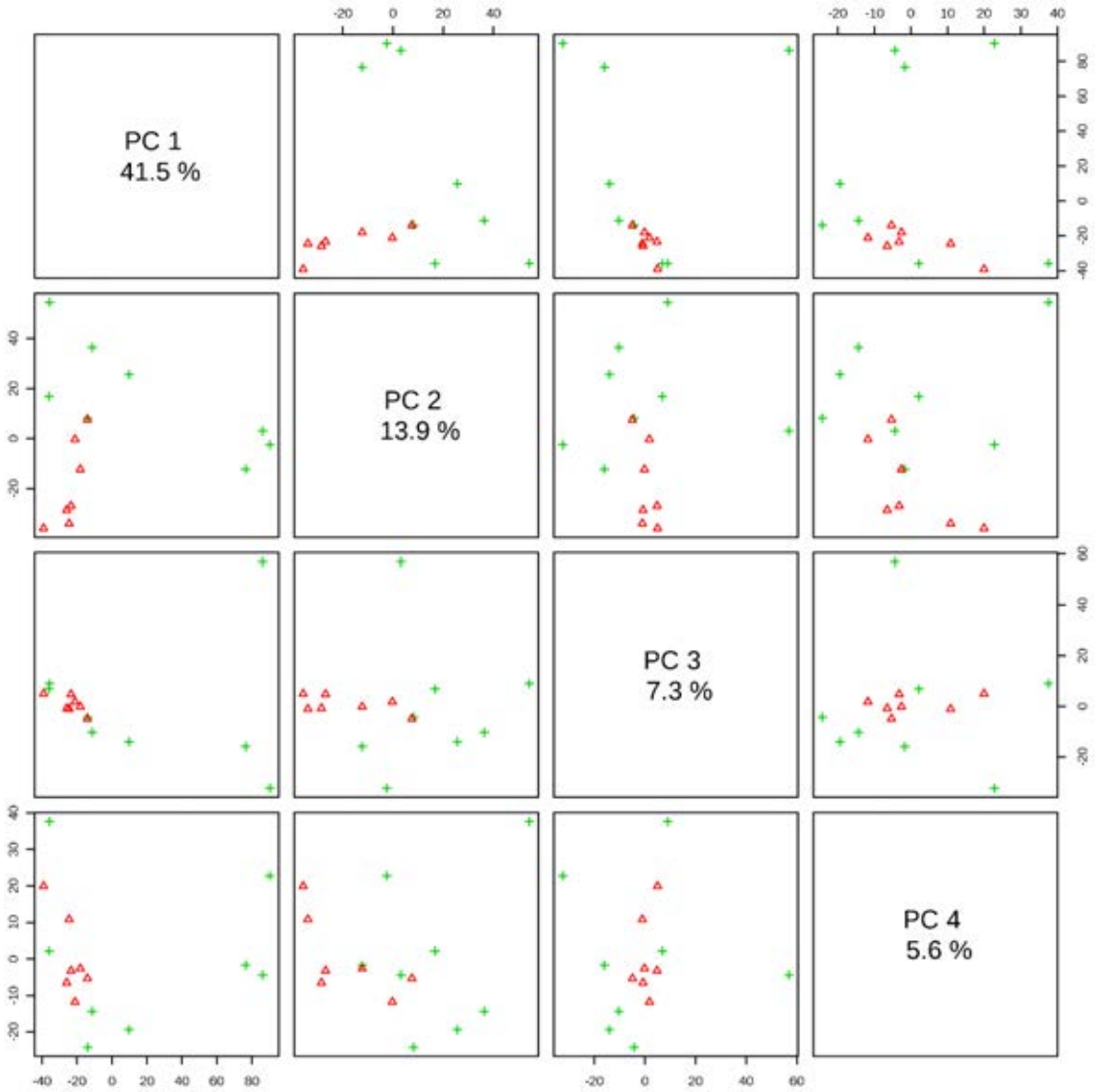
PCA plots of Control 4h and Carbaryl 4h groups. The red points are the Control 4h group and the green points are the Carbaryl 4h group.



PCA plots of Control 8h and Carbaryl 8h groups. The red points are the Control 8h group and the green points are the Carbaryl 8h group.



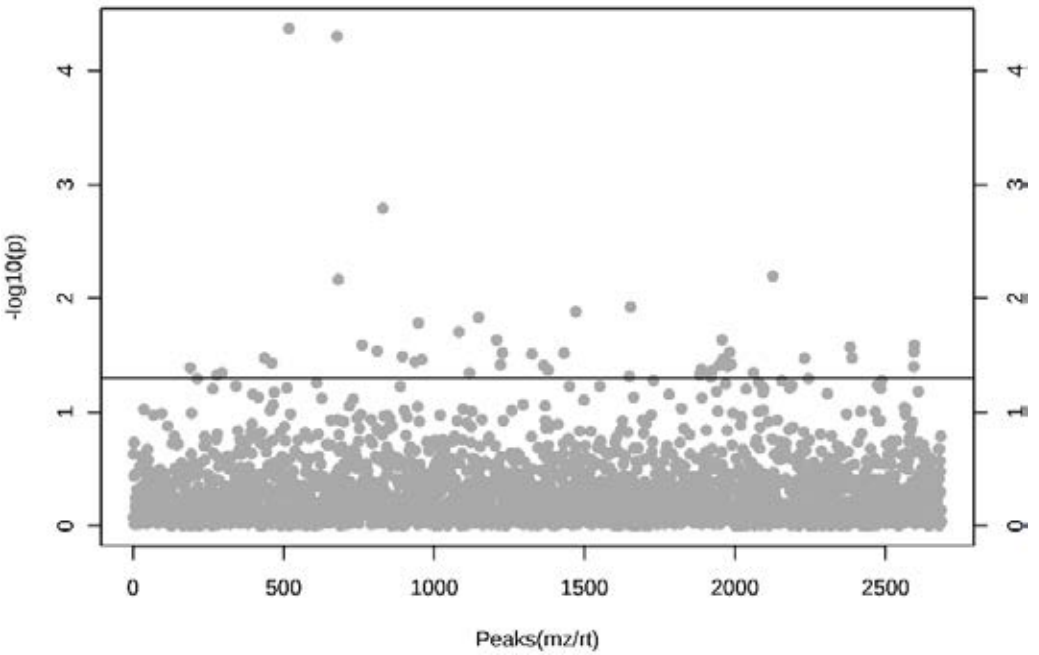
PCA plots of Control 12h and Carbaryl 12h groups. The red points are the Control 12h group and the green points are the Carbaryl 12h group.



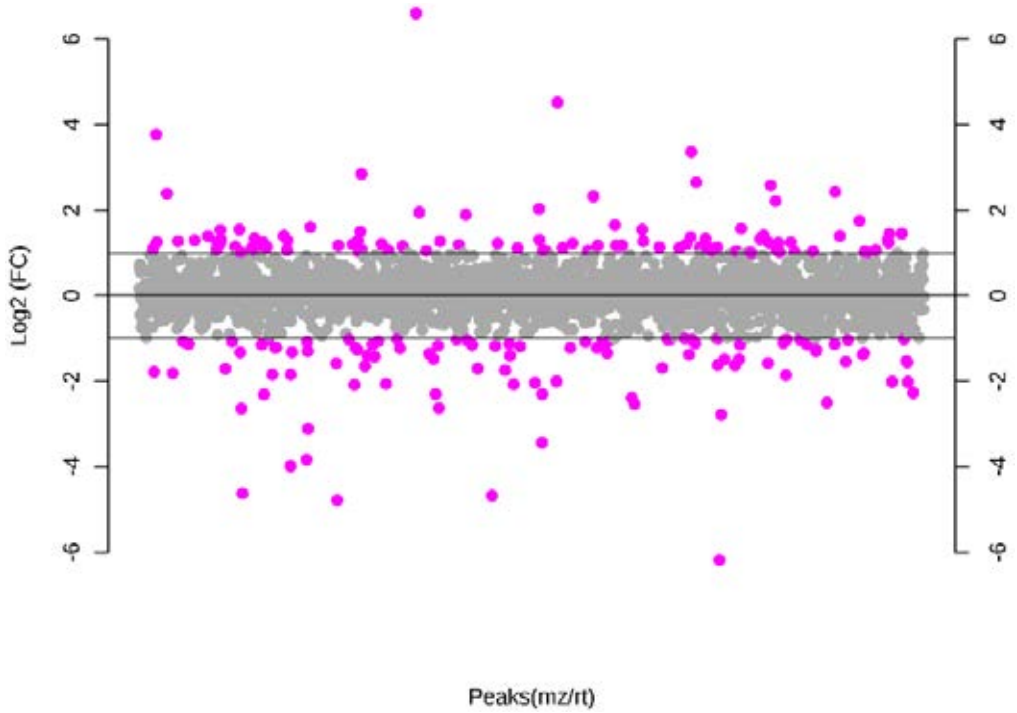
PCA plots of Control 24h and Carbaryl 24h groups. The red points are the Control 24h group and the green points are the Carbaryl 24h group.

13.2. Carbaryl treatment univariate plots

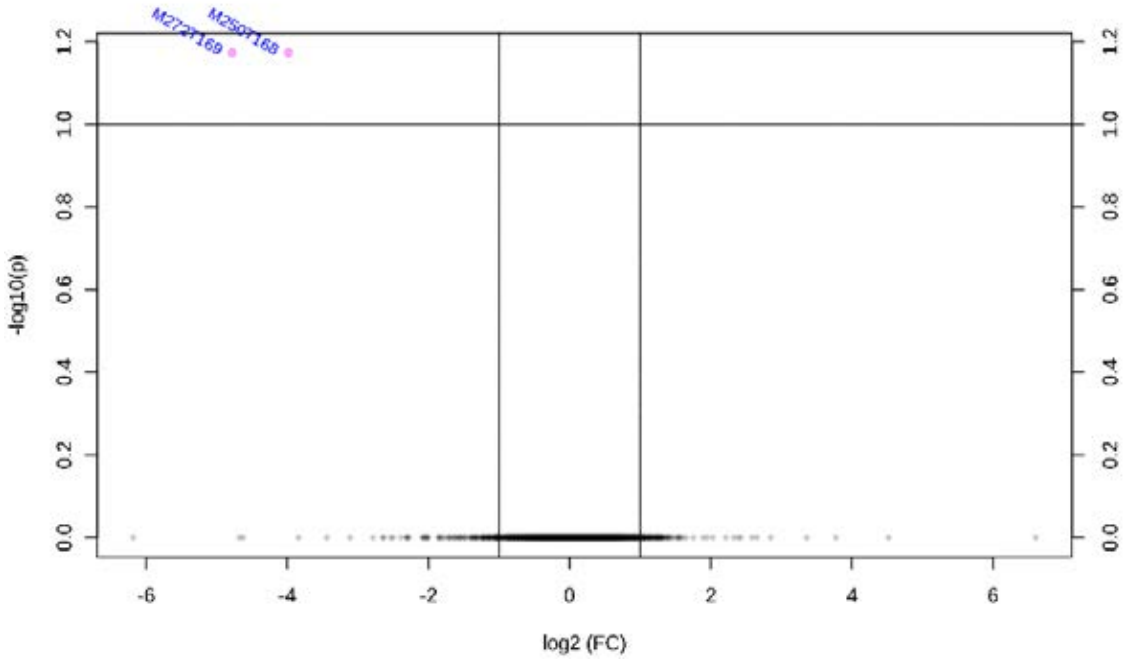
13.2.1. Four-hour time point



T-test plot for the Control vs Carbaryl 4h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

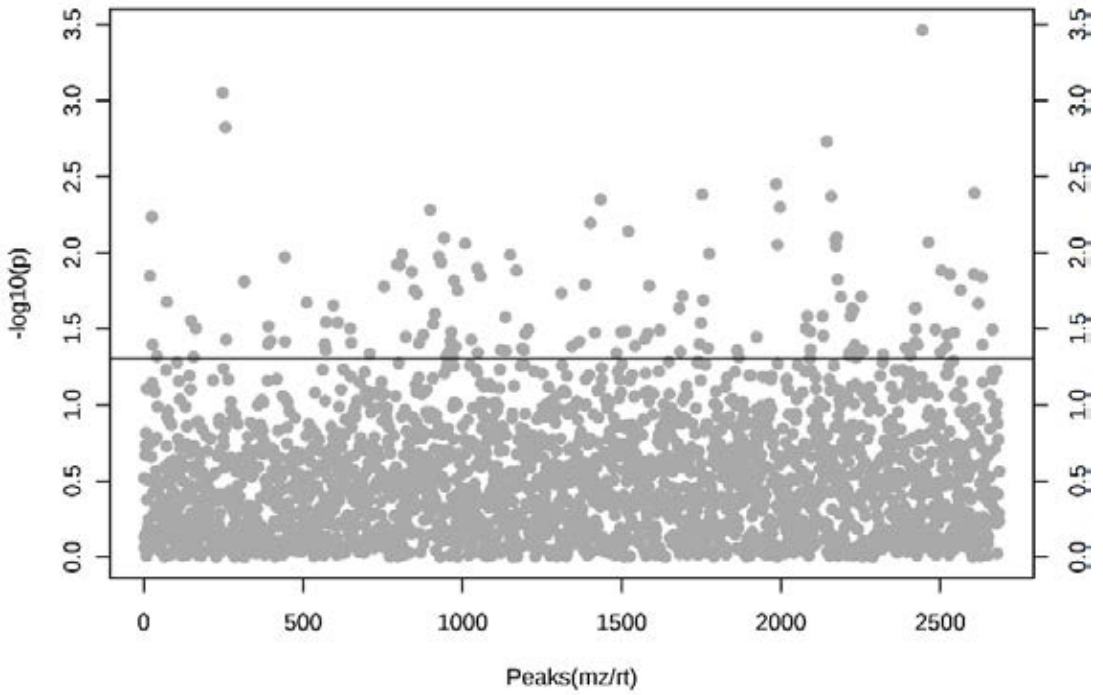


Fold change plot for the Control vs Carbaryl 4h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 199 peaks.

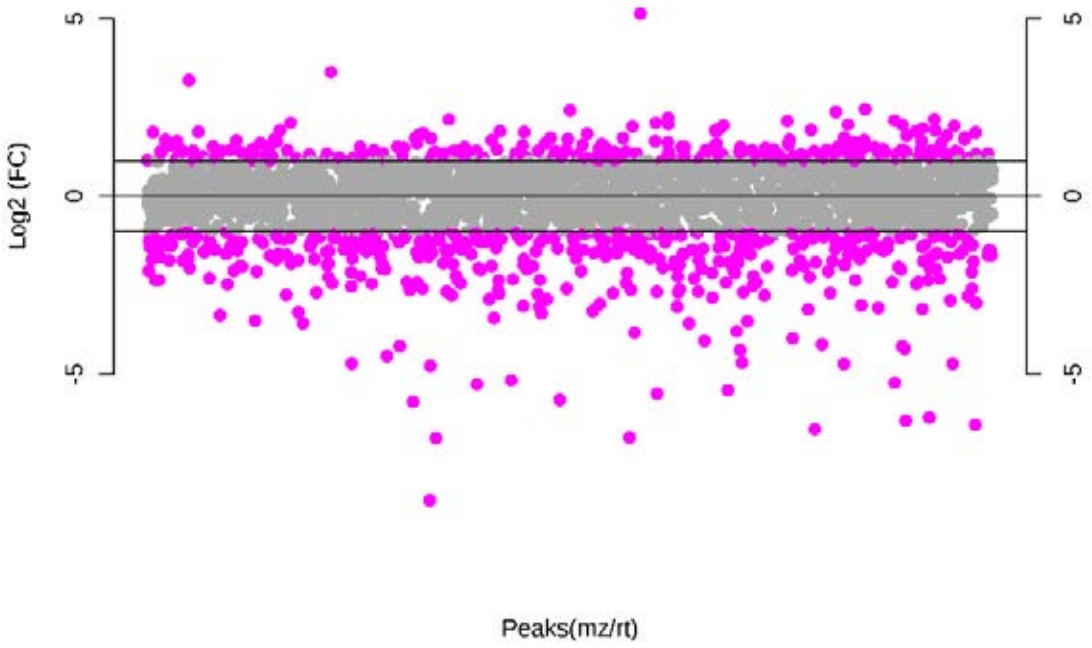


Volcano plot for the Control vs Carbaryl 4h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 2 peaks.

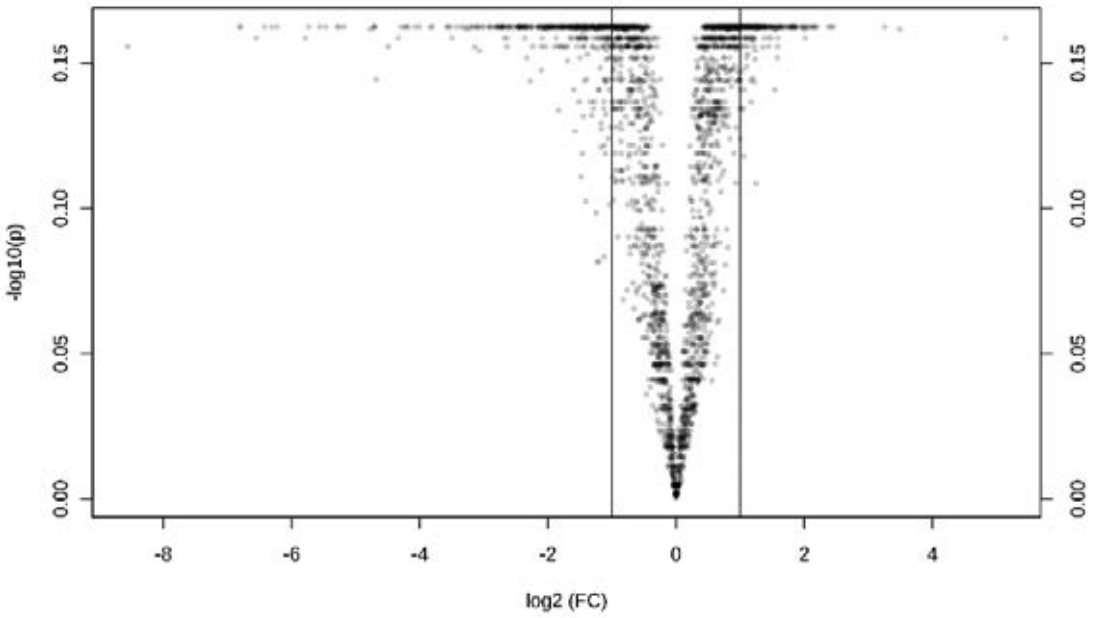
13.2.2.Eight-hour time point



T-test plot for the Control vs Carbaryl 8h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

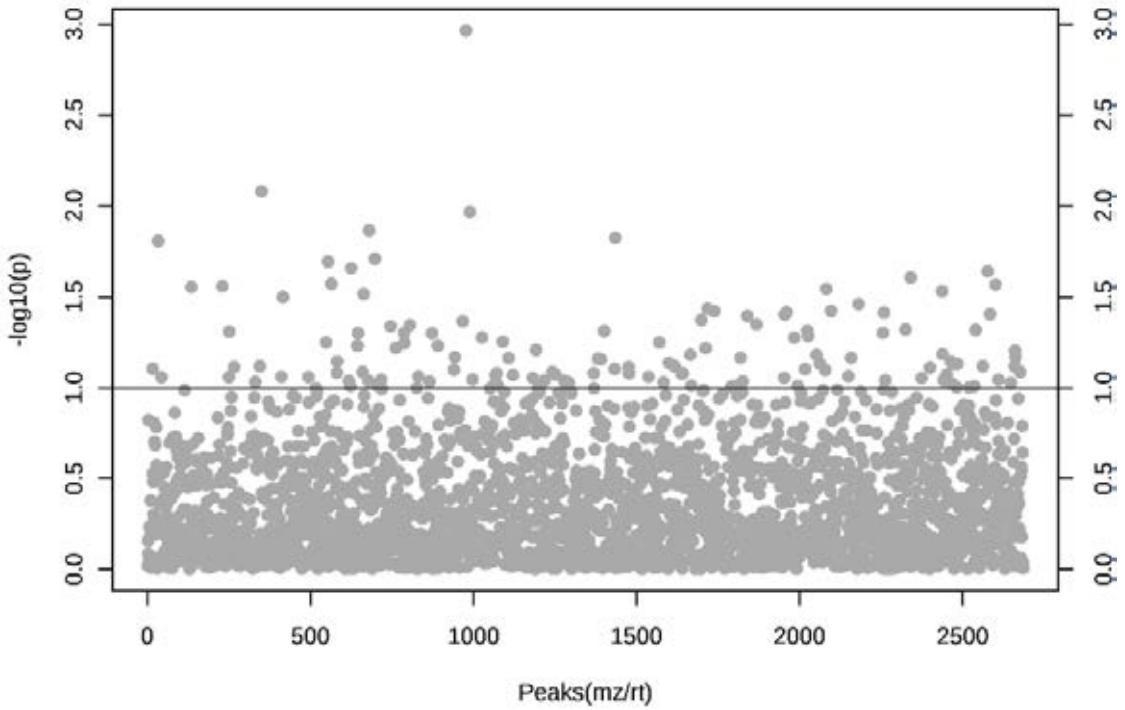


Fold change plot for the Control vs Carbaryl 8h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 681 peaks.

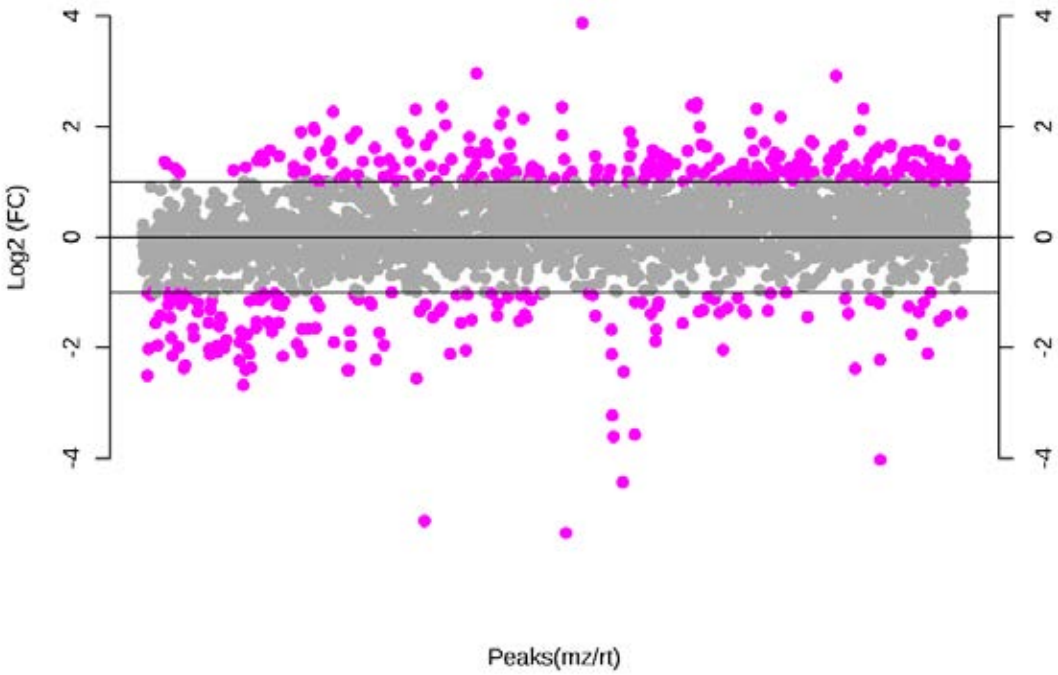


Volcano plot for the Control vs Carbaryl 8h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.

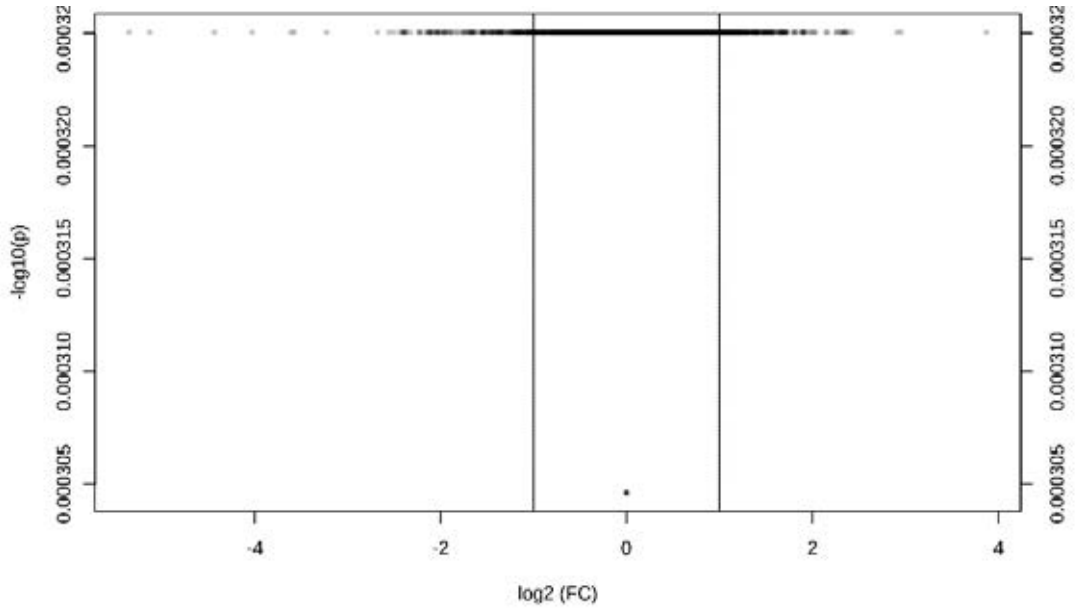
13.2.3. Twelve-hour time point



T-test plot for the Control vs Carbaryl 12h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

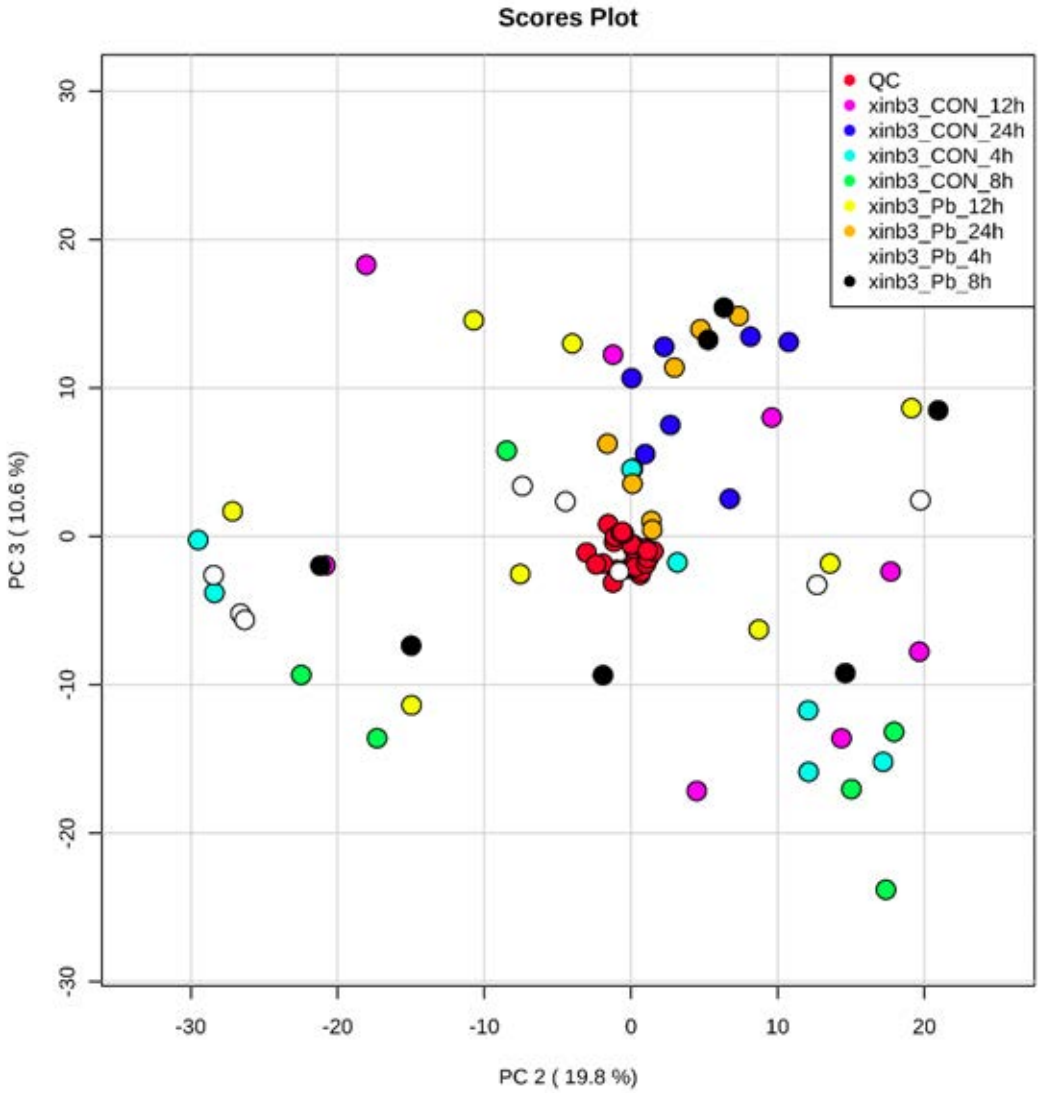


Fold change plot for the Control vs Carbaryl 12h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 416 peaks.

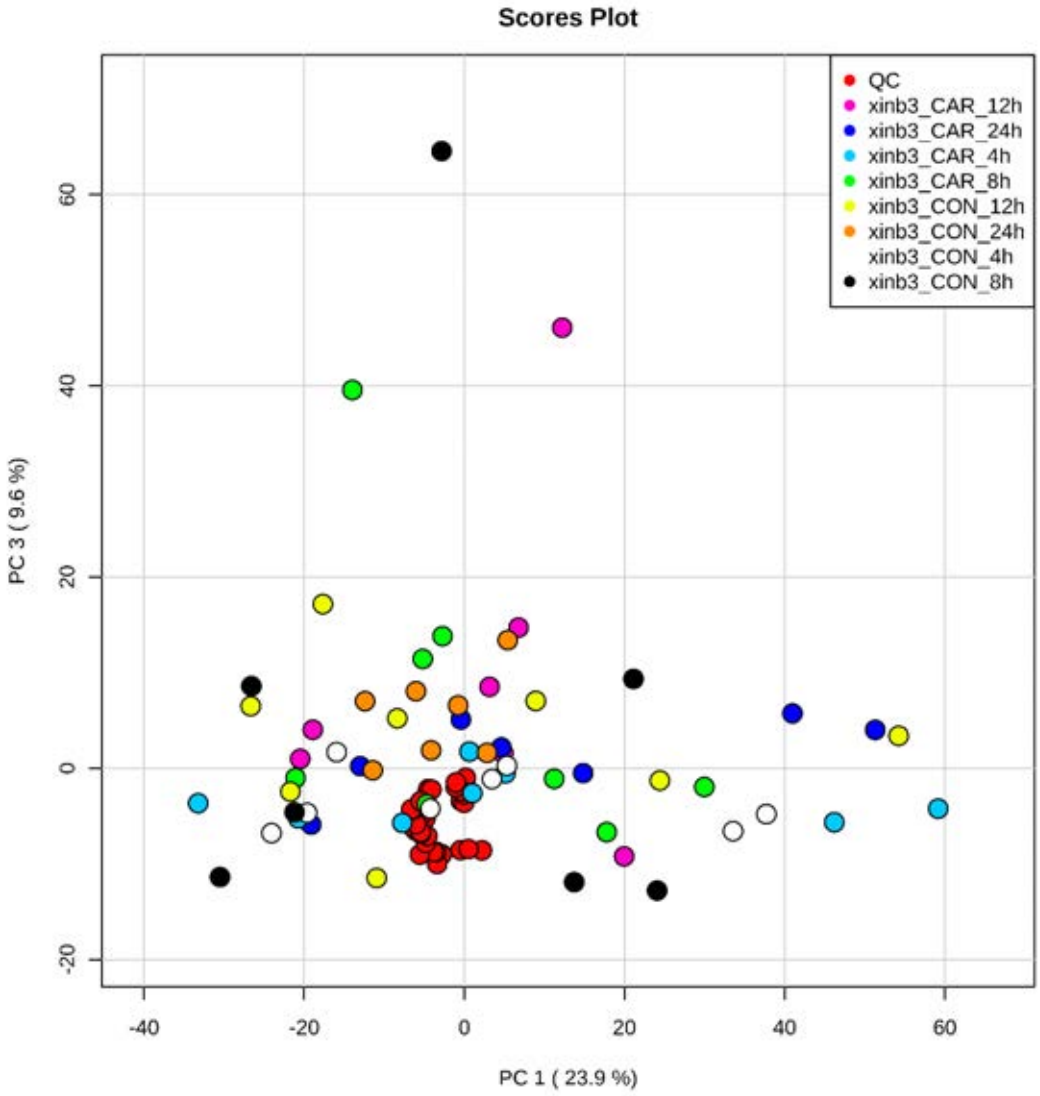


Volcano plot for the Control vs Carbaryl 12h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.

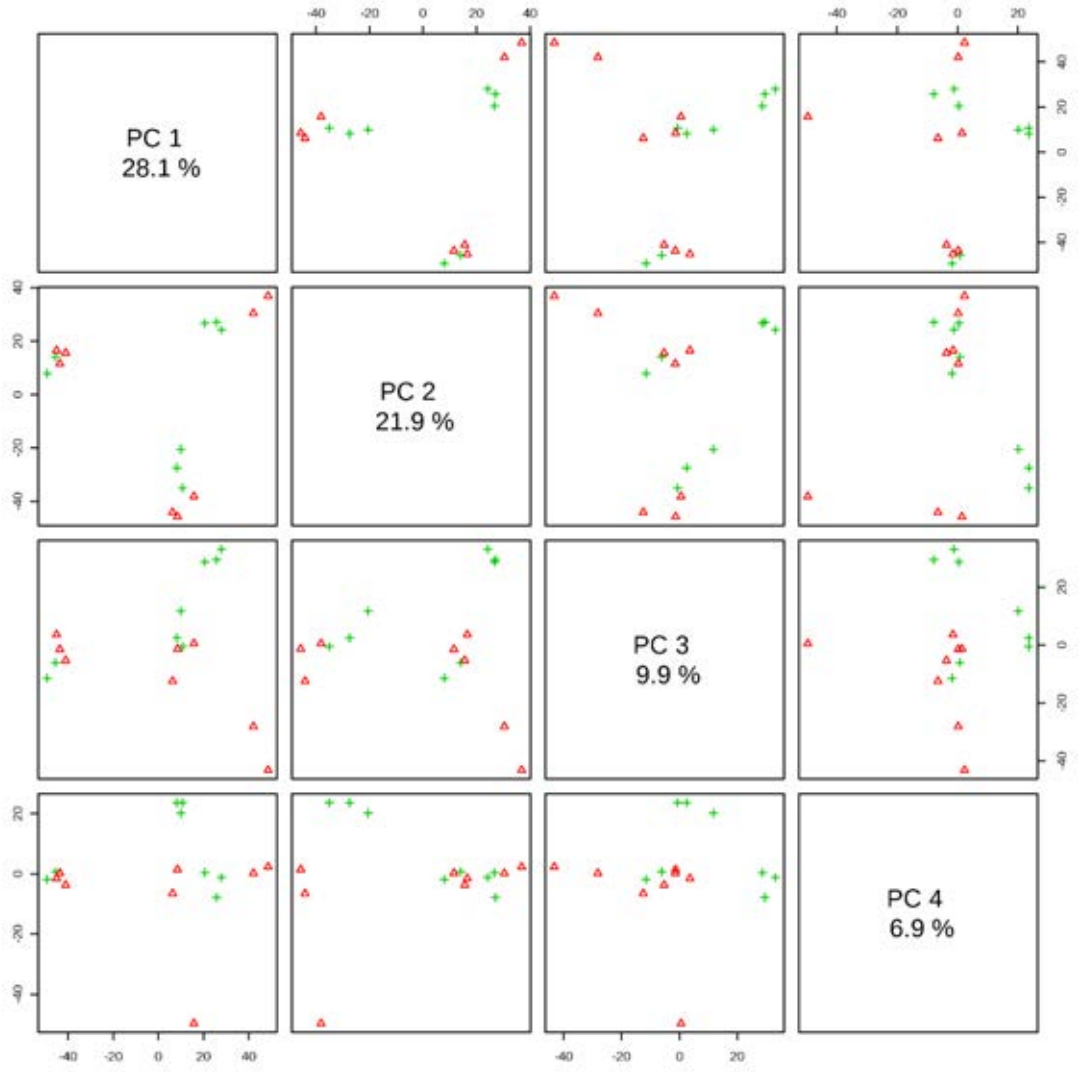
13.3. Lead treatment PCA plots



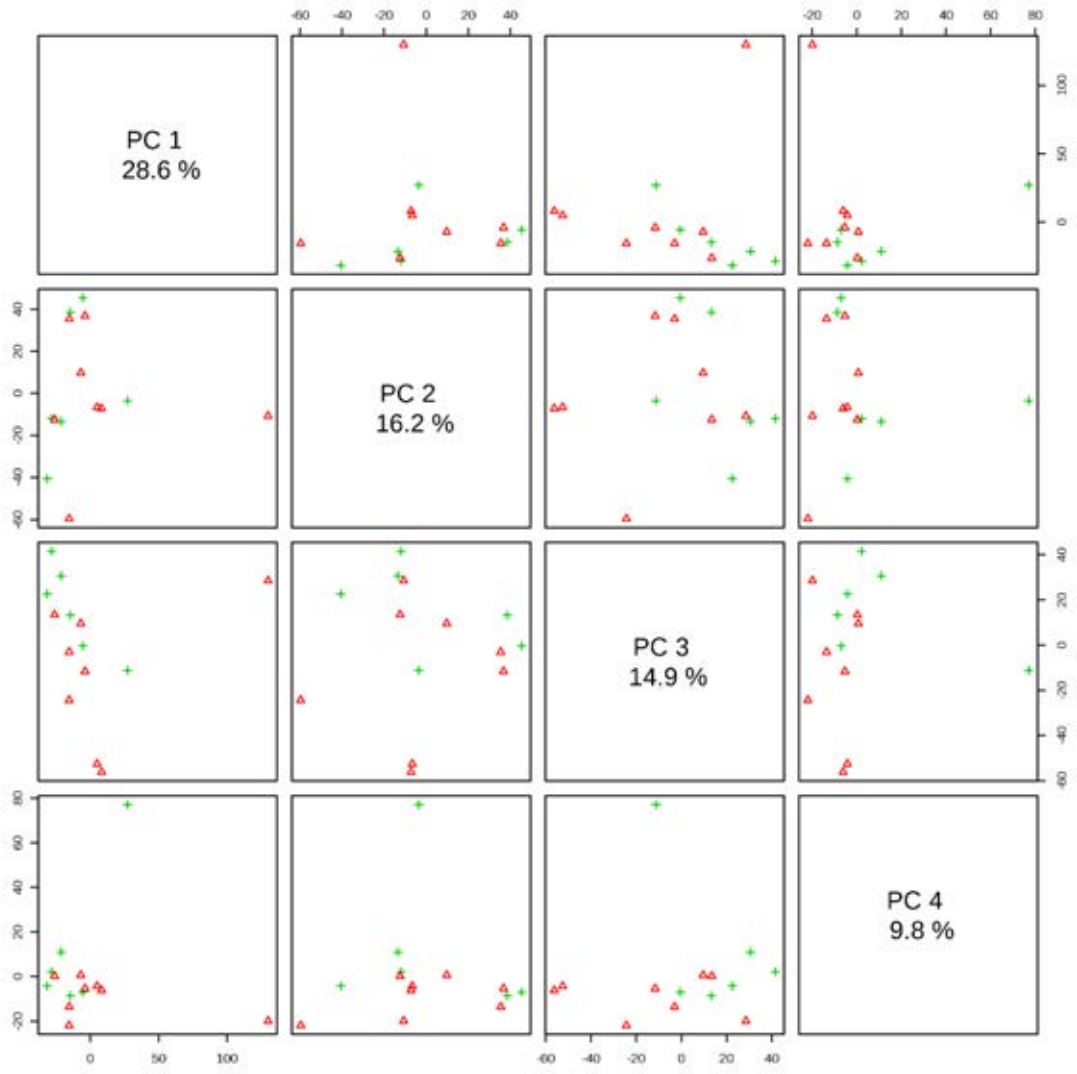
PC2 vs PC3 PCA scores plot of all Lead and Control groups.



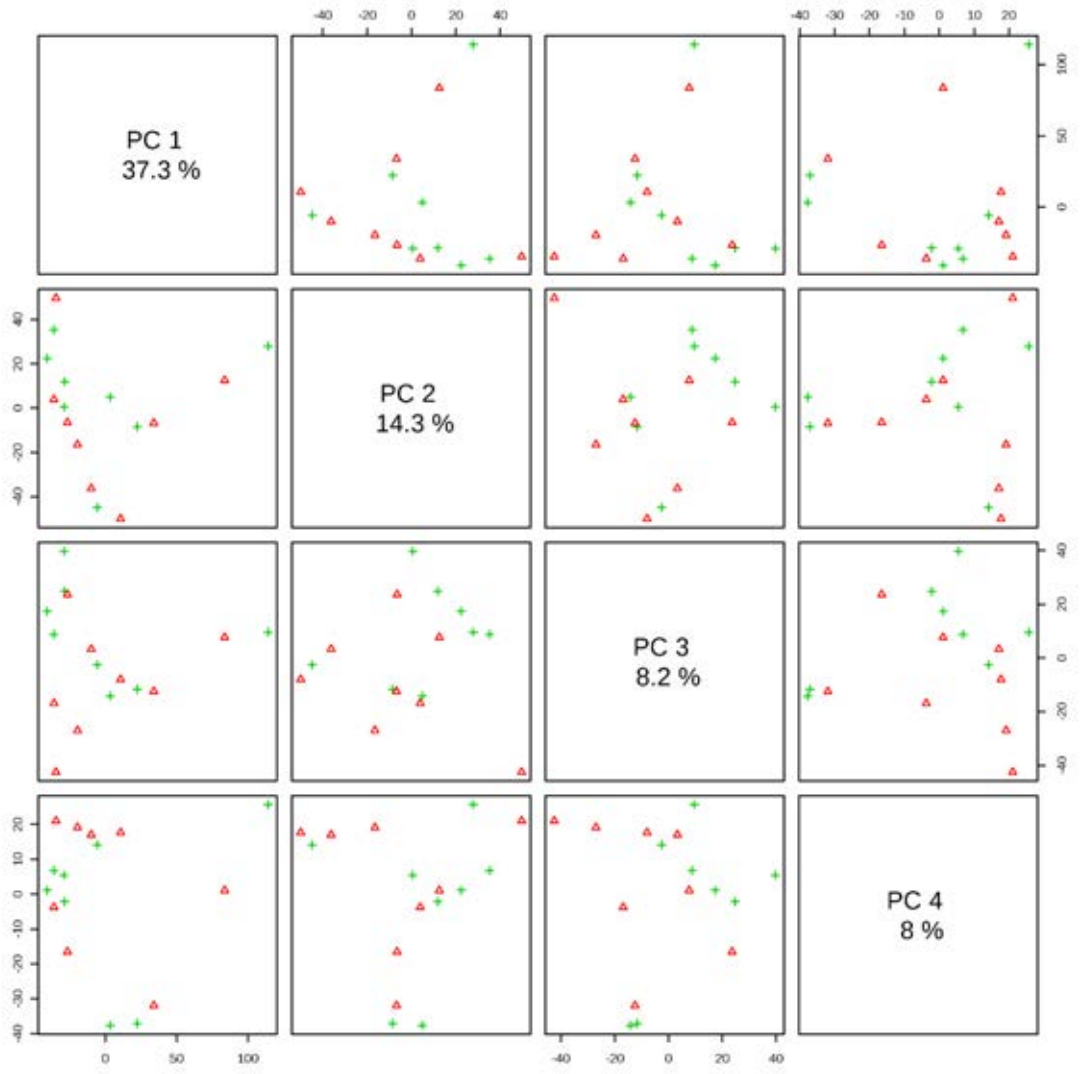
PC1 vs PC3 PCA scores plot of all Lead and Control groups.



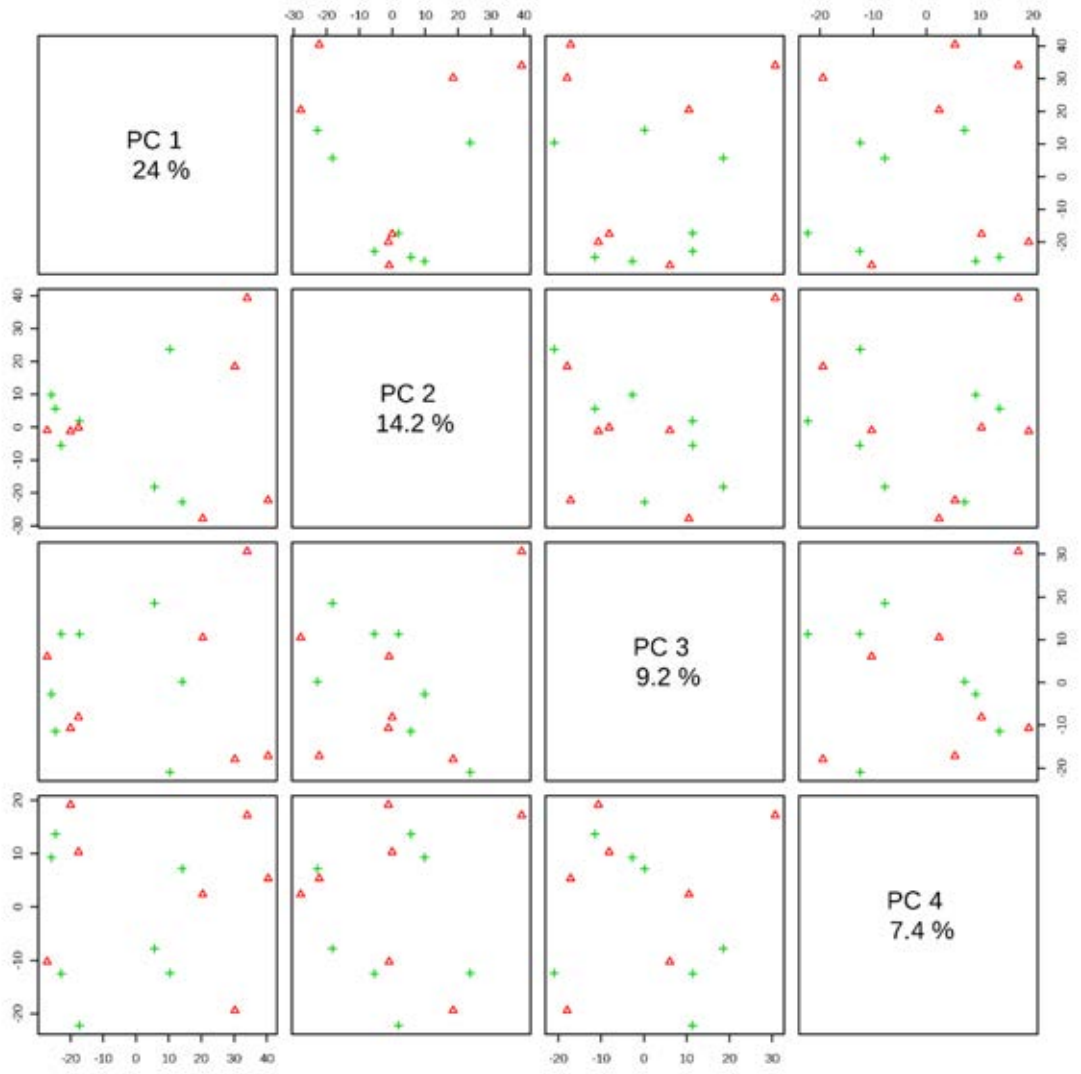
PCA plots of Control 4h and Lead 4h groups. The red points are the Control 4h group and the green points are the Lead 4h group.



PCA plots of Control 8h and Lead 8h groups. The red points are the Control 8h group and the green points are the Lead 8h group.



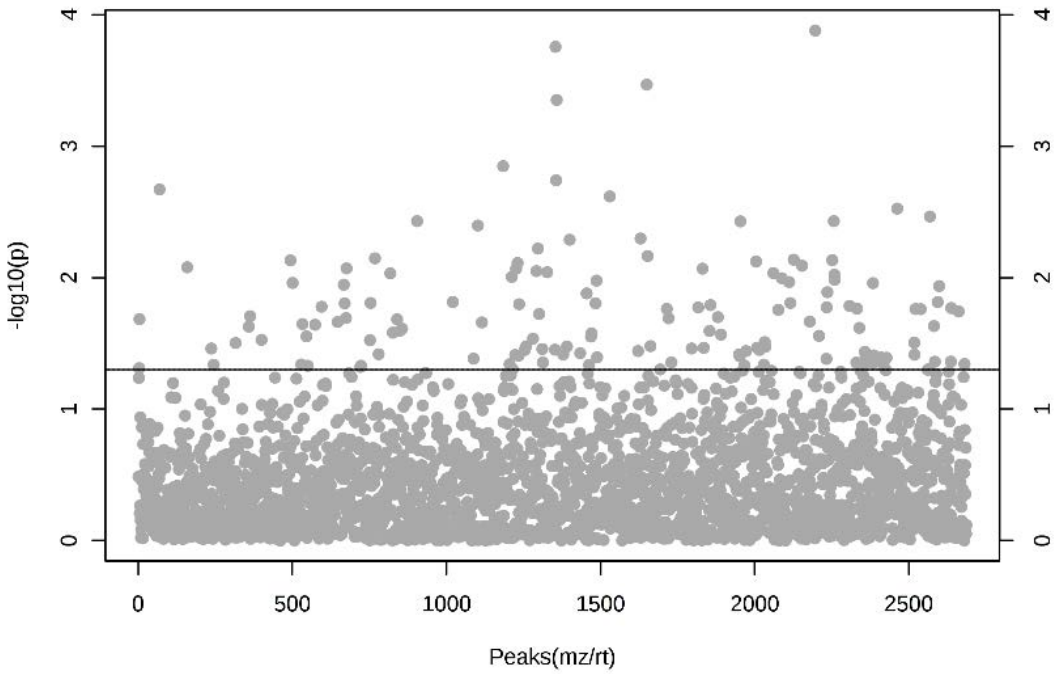
PCA plots of Control 12h and Lead 12h groups. The red points are the Control 12h group and the green points are the Lead 12h group.



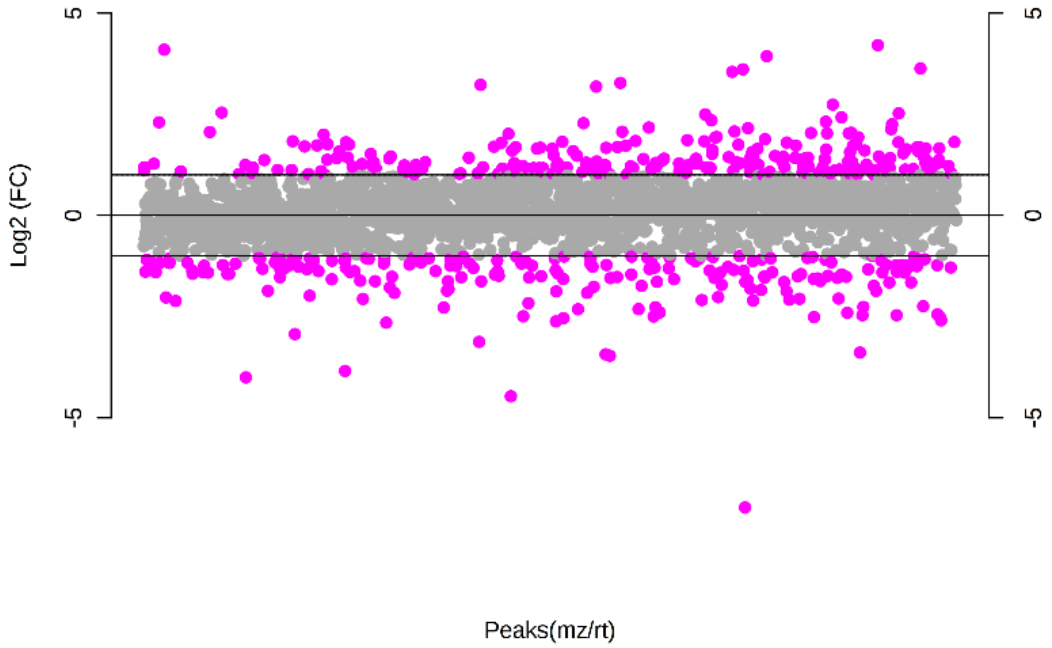
PCA plots of Control 24h and Lead 24h groups. The red points are the Control 24h group and the green points are the Lead 24h group.

13.4. Lead treatment univariate plots

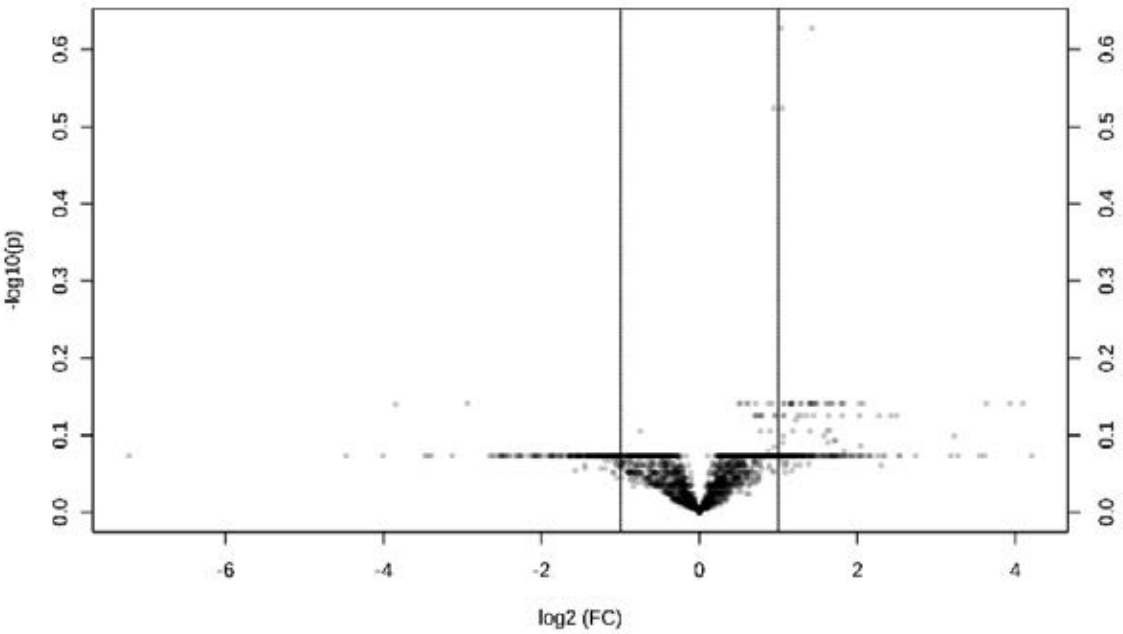
13.4.1. Four-hour time point



T-test plot for the Control vs Lead 4h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

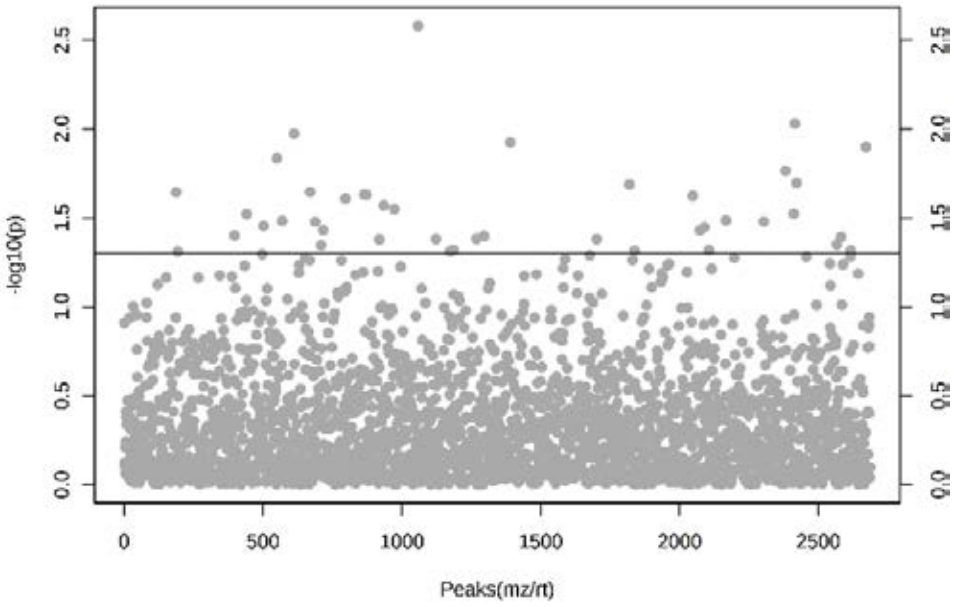


Fold change plot for the Control vs Lead 4h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 448 peaks.

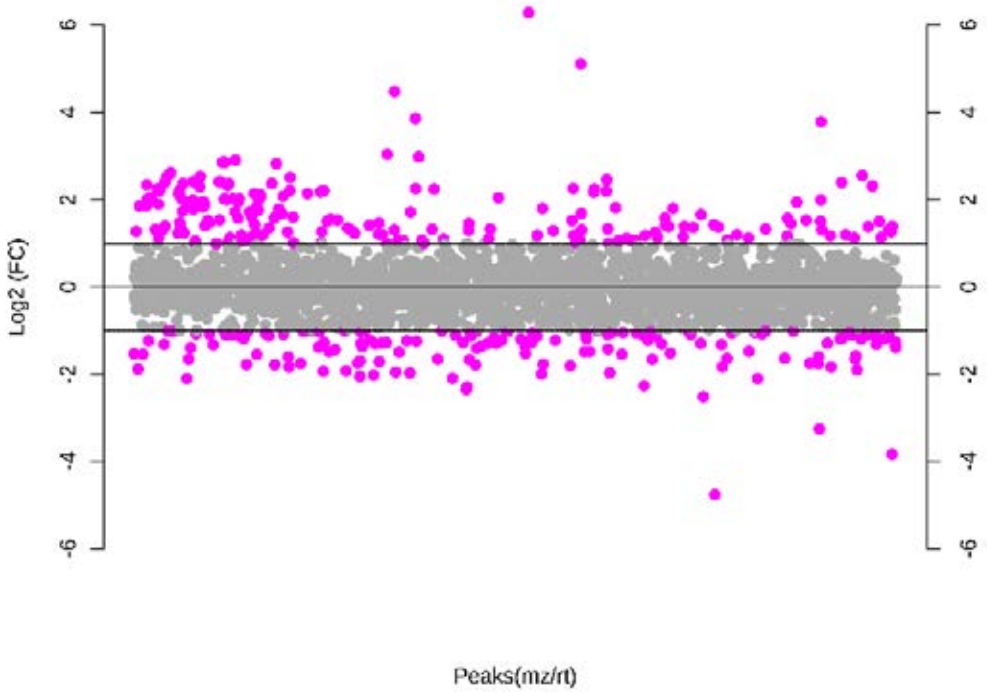


Volcano plot for the Control vs Lead 4h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.

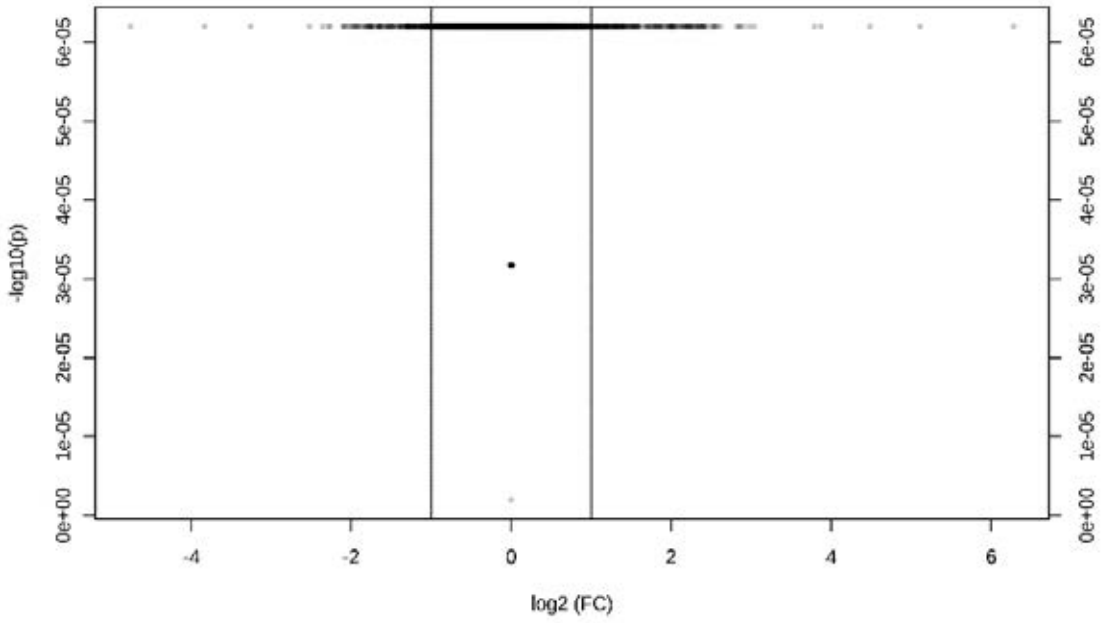
13.4.2.Twelve-hour time point



T-test plot for the Control vs Lead 12h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05

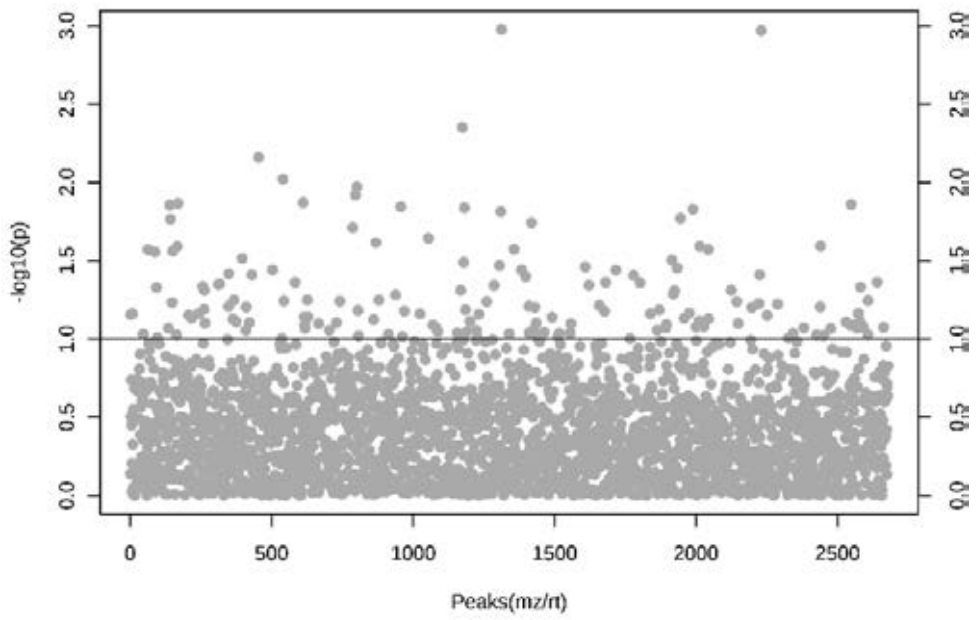


Fold change plot for the Control vs Lead 12h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 336 peaks.

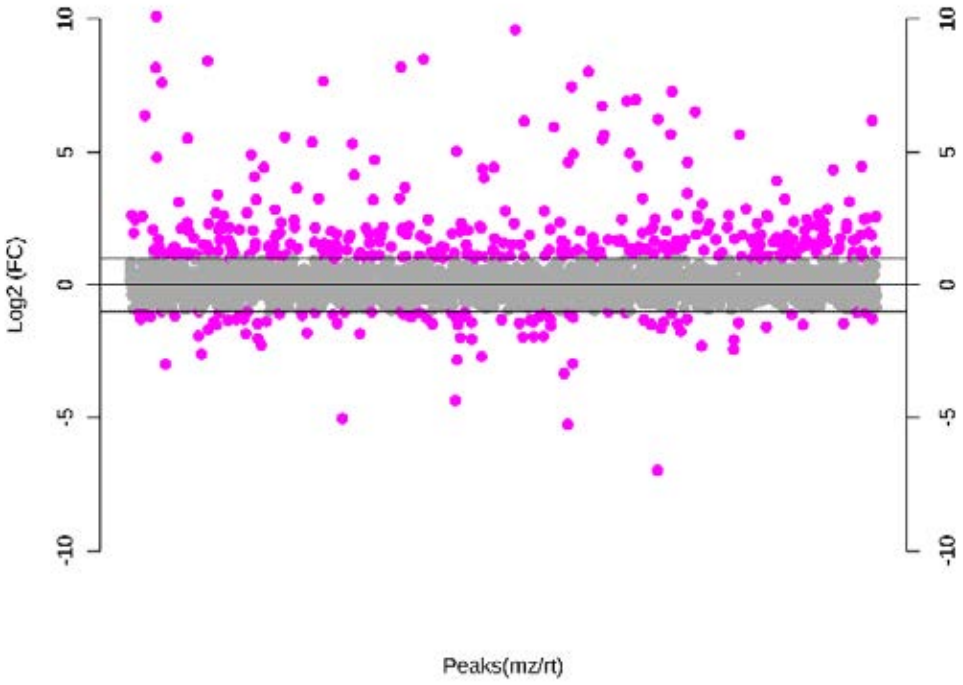


Volcano plot for the Control vs Lead 12h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.

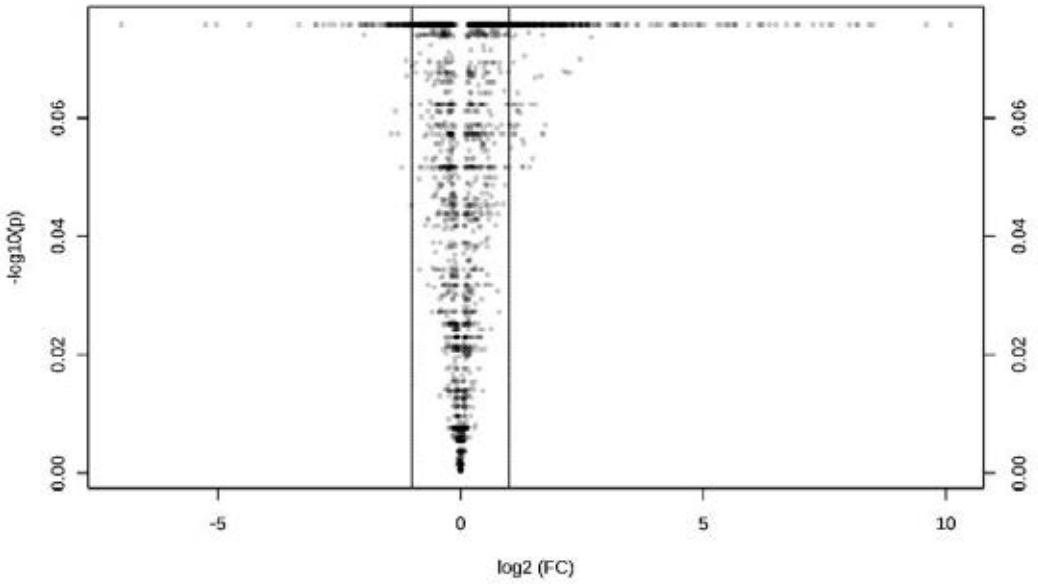
13.4.3.Twenty Four-hour time point



T-test plot for the Control vs Lead 24h time point sample groups. -Log10(p) values are shown on the y-axis. Points are coloured pink if the FDR corrected p-value is less than 0.05. 0 peaks have an FDR corrected p-value of less than 0.05



Fold change plot for the Control vs Lead 24h time point sample groups. Log2 fold change values are shown. Points are coloured pink if the raw fold change value is at least ± 2.0, of which there are 446 peaks.



Volcano plot for the Control vs Lead 24h time point. The x-axis shows log2 fold change values, the y-axis shows -log10 FDR corrected p-values. Points are coloured pink if the FDR corrected p-value is less than 0.1 and the raw fold change value is at least ± 2.0, of which there are 0 peaks.