A SYSTEMS BIOLOGY APPROACH TO THE PRODUCTION OF BIOTECHNOLOGICAL PRODUCTS THROUGH SYSTEMATIC IN SILICO STUDIES

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2012

Olusegun James Oshota

School of Chemical Engineering and Analytical Sciences

Contents

Word count 63871

List of Tables

List of Figures

The University of Manchester

Olusegun James Oshota
Doctor of Philosophy
A systems biology approach to the production of biotechnological products through systematic in silico studies
February 2, 2012

Background: Currently, the development of microbial strains for biotechnological production of chemicals and materials can be improved by using a rational metabolic engineering that may involve genetic engineering and/or systems biology techniques. Elementary flux mode analysis (EFM) and Flux balance analysis (FBA) are the two most commonly used methods for probing the microbial network system properties for metabolic engineering purposes. EFM can be used to identify all possible pathways. However, combinatorial explosion of the number of EFMs obtained during EFM analysis, especially for large reaction networks, hinders the use of EFM data for developing gene knockout strategies. The objective of this project was to identify interesting target products and design 'proof of principle' *Saccharomyces cerevisiae* strains capable of overproducing a target product; in this case lysine was chosen.

Methods: EFMs were calculated for a reaction network from *S. cerevisiae*. In order to make sense of the large EFM solution space, a novel approach based on computational reduction and clustering of EFM datasets into subsets was developed, which aided the prediction of knockouts for lysine production. A Pattern analysis method, based on regular expression matching, was also developed to interpret the EFM data. FBA frameworks, OptKnock and GDLS, were used to design *in silco* production strains based on genome-scale models of yeast. Double and triple *S. cerevisiae* lysine producing strains were constructed using a PCR-based deletion method. Absolute and relative metabolome measurements for lysine and other metabolites in the single and double mutants were achieved using GC-TOF-MS.

Results: The new computational and clustering methodology aided significantly the EFM-based *in silico* design of *S. cerevisiae* strains for enhanced yield of lysine and other value chemicals. Ethanol and lysine overproducing *in silico* strains were also developed by OptKnock and GDLS. Remarkably, the production strains with single deletions, lsc2 and glt1, excreted into the medium five times the amount of lysine than the control strain. Five *S. cerevisiae* double mutant strains were successfully constructed. Two-fold increase in flux towards lysine production was demonstrated by *S. cerevisiae* double mutant M1, while both *S. cerevisiae* double mutants M4 and M5 showed about four-fold increase in lysine production.

Conclusion: The general modelling and data reduction approaches developed here contributed in obviating the enormous problems associated with trying to obtain the EFMs from large reaction network models and interpreting the resulting of large number of EFMs. EFM analysis aided the development of single and double *S. cerevisiae* mutant strains, capable of increased yield of lysine. The computational method was validated by construction of strains that are able to produce several fold more lysine than the original strain.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the "Copyright") and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made **only** in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the "Intellectual Property Rights") and any reproductions of copyright works, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of the School of Chemical Engineering and Analytical Sciences.

Acknowledgements

First of all, I would like to thank Prof. Hans Westerhoff and the committee, who in 2007 (four years ago) decided to accept me onto the PhD programme at the Doctoral Training Centre (DTC) for Systems Biology, University of Manchester in Manchester. It was a rare privilege to be one of the students (2^{nd} cohort) of the training institution headed by this influential and foremost Systems Biologist. Apart from the training benefits that I received at the DTC, I carried out my PhD research at the Manchester Centre for Integrative Systems Biology (MCISB), which also had Prof. Westerhoff as its Director. I also benefited from discussions with and listening to Prof. Westerhoff at various Systems Biology meetings and seminars. I must not forget to thank Prof. Douglas Kell, another foremost Systems Biologist and also the past Director of MCISB. Prof. Kell was actually the one who came up with idea of the research for my thesis while he was the Director of MCISB. Upon contacting Prof Kell about the PhD research, and as he was leaving MCISB and on his way to assume the Directorship of BBSRC, he handed me over to his able former PhD student, Prof. Pedro Mendes (Chair in Computational Systems Biology, MCISB), who in his own right, is a Systems Biologist of note.

I gratefully acknowledge the immense contributions of my three Supervisors, Prof. Pedro Mendes, Dr. Naglis Malys and Dr. Evangelos Simeonidis to my PhD training and research work at MCISB.

My three years of PhD research under Prof. Mendes as my Principal Supervisor was a wonderful experience. In those three years I was fortunate to have Prof. Mendes,

whom I personally refer to as "a-must-have-Supervisor" for every PhD student. Not only did he benefit me by imparting to me his astounding and well-rounded knowledge of systems biology applied to metabolic engineering, I found highly rewarding the clarity and quickness of his understanding of problems and proffering of solutions to problems. He also guided and monitored my progress in other areas that were important to the completion of my PhD research.

Dr. Naglis Malys imparted to me a lot of expertise in the laboratory, and under the guidance of whom I was able to successfully create knockout strains for the validation of my in silico modelling. Dr Malys also helped me in numerous ways such as encouraging me to write better scientific reports of my experiments. Dr Evangelos Simeonidis was important to the computational part of my research work. He benefited me with his expertise in flux balance analysis and other general aspects of computational modelling. I also received immense encouragement and help from from Dr. Simeonidis regarding how to write a good scientific paper.

The first year training at the DTC was highly beneficial to my early understanding of the various aspects of systems biology. In this regard, I owe a big gratitude to Dr. Gerold Baier (Assistant Director, DTC), Dr. Mara Nardeli, Dr. Kieran Smallbone, Prof. Hans Westerhoff , Prof. Pedro Mendes and many others. Lynne Davies, Gemma Coleman and other administrative staff deserve to be thanked for the crucial administrative roles that facilitated my training and research work at DTC.

I'm greatly indebted to both Dr. Warwick Dunn and Dr. Cate Winder for carrying out the GC-MS analysis of my metabolomics samples and the pre-processing of data. I gratefully acknowledge and thank Dr. Samrina Rehman for giving me a good training on how to analyse the metabolomic data. My special thanks also goes to Dr. Hanan Messiha for ordering most of the reagents for my experiments.

My PhD work also gave me the opportunity to interact with many other experts in systems Biology who benefited my work in one way or another. They represented for me immense sources of scientific, intellectual, and even administrative help. They

also offered words of encouragement. In this regard, I thank Dr. Kieran Smallbone, Dr. Neil Swainston, Dr. Hanan Messiha, Dr. Juergen Pahle, Dr Joseph Dada, Dr. Malkey Varma, Dr. Daniel Jameson and Dr. Dieter Weichart .

During the course of my research I needed the help of others outside the University of Manchester. Evelyne Dubois,at Institut de Recherches du CERIA, Belgium was kind to donate lysine producing yeasts strains for controlling my experiments on lysine production. Stefan Hoops (virginia bioinformatics institute) helped to improve the features of elementary flux mode analysis (EFMs) in the COPASI software, which was useful to me for computing EFMs form large reaction networks quickly.

I would also like to thank my family for their support and words of encouragement throughout this PhD work.

Finally, I thank the BBSRC, EPSRC, DTC and MCISB for the financial supports towards my PhD training and research.

Abbreviations

EFM Elementary flux mode

FBA Flux balance analysis

GC-MS Gas chromatography-Mass spectrometry

MCA Metabolic control analysis

Wss Within sum of squares

About the author

My first qualification was in Chemical Pathology/Clinical Biochemistry (AIMLS diploma, 1988) from the University College Hospital, Ibadan, Nigeria. After this qualification, I was involved in clinical laboratory diagnostics in various hospitals. In 1993, I obtained a post-graduate diploma in Clinical Biochemistry (pass with distinction) from the University of Leeds (UK). Further post-graduate degrees obtained are in Molecular Pharmacology and Biotechnology (MSc, University of Leeds, UK, 2000), and Bioinformatics and Computational Biology (MRes, University of Leeds, UK, 2007: pass with distinction). I worked as a Research Scientist (in Molecular Biology, DNA microarray and Bioinformatics) at Health protection Agency, Porton Down, Salisbury, UK, for 6 years before the start of my MRes Bioinformatics course in 2006. Finally, I was on the PhD programme for 4 years (2007 - 2011).

Chapter 1

Introduction

This chapter is divided into four sections. The first section outlines the motivation and aims for this PhD work, and also the objectives for achieving the aims are described. The second section, "Thesis overview", outlines the structure and arrangements of the entire thesis. The last two sections,"Strategies employed in biotechnological production of products" and "Production of value chemicals and materials using metabolic engineering" are literature reviews covering the different aspects of the PhD work.

1.1 Motivation, Aims and Objectives

1.1.1 Motivation

Since antiquity, production of materials and products have formed an important and integral part of human activities. Notably, production of various products and materials have been carried out by processes based on fermentation. As an example, yeast (*Saccharomyces cerevisiae*) was used for thousands of years to ferment food and beverages. A few chemical products were made from microbial sources before the beginning of petrochemical industry in the 20th century. However, the extremely low prices of petrochemical resources, largely contributed to a shift in the microbial

production of many products using renewable resources to chemical synthesis of many products. The production of broad range of modern products, broadly categorised into fine chemicals, bulk chemicals, pharmaceutical products, plastics and fuels is strongly based on the chemical industry. Unfortunately, the reliance of the chemical industry on the petrochemical resources for production of chemicals and materials involves the consumption of large amounts of fossil resources and emission of large amounts of waste (Soetaert and Vandamme, 2006). Production of chemicals and products is highly dependent on feedstock cost and currently fossil carbon sources, such as oil and gas, will soon become too expensive. This, together with the increasing environmental concerns about pollution due to fossil fuels, depletion of oil reserves and advances in biotechnology are swaying policies towards sustainable raw materials as feedstocks for production of chemicals. The paradigm shift into an era whereby bio-renewable green products is beginning to serve as the feedstock is evidenced by policy papers by governments and institutions (Werpy and Petersen, 2004). The application of biotechnological principles and methods has made possible the production of many bio-based products and this trend is set to continue.

In general terms, biotechnology is defined as "the application of microorganisms/-cells, or components thereof (e.g., enzymes), for the production of useful goods and services" (Rogers et al., 2005). Currently, three different areas of modern biotechnology are recognised: red, white and green biotechnology. Red biotechnology describes pharmaceutical and medical processes such as the design of organisms to produce antibiotics, while white biotechnology refers to industrial processes such as the use of microorganisms to produce chemicals, materials and energy mainly based on the use of enzymes as catalysts, and green biotechnology focuses on genetically modified crops. White biotechnology is also known as industrial biotechnology and the "white" designation refers to the positive environmental benefits of industrial biotechnology. A number of advantages have been associated with the implementation of white biotechnology. White biotechnology contributes significantly to green chemistry involving the conversion of renewable resources, such as sugars or vegetable oils,

into a wide variety of chemical substances such as fine and bulk chemicals, pharmaceuticals, biocolourants, solvents, bioplastics, vitamins, food additives, and biofuels (Dale, 2003). Since the starting materials are from agricultural products instead of fossil fuels, white biotechnology has a beneficial effect on greenhouse gas emission, and also its cost-effective benefits may be significant. Other performance benefits are associated with white biotechnology such as increased conversion efficiency, higher product purity, lowered energy consumption and lower generation of chemical wastes (Soetaert and Vandamme, 2006).

A phenomenal growth rate (16 - 20%) is predicted for the industrial biotechnological production of fine chemicals which are currently produced through complex synthetic and combinatorial methods (Hirche, 2006). Industrial biotechnology is predicted to grow 1 - 20% in the areas of basic chemicals and commodities, speciality and added-value chemicals, and polymers respectively (Otero and Nielsen, 2010). However, a large proportion of the venture capital investment, about 88%, is raised by red biotechnology, while white and green biotechnology attract only 5% and 7% respectively (Hirche, 2006).

A very important issue in the implementation of white biotechnology is feedstock. It was estimated that up to 40% of all bulk chemicals will be based on global biomass (EuropaBio, 2004), as the new market for bio-renewable green products opens up for large volumes of diverse products and processing technologies, and hemicelluloses have shown to be important targets. Another vital issue for the successful implementation of red and white biotechnology is the efficient host organism. With the increasing shift towards a bio-based economy, there is a rising demand for developing efficient microbial strains that can produce products within the categories of fine chemicals, pharmaceuticals, food additives and supplements, flavour and aroma compounds, colourants, vitamins, pesticides, bio-plastics, solvents, bio-plastics, bulk chemicals and biofuels.

In the past two decades, metabolic engineering has been used extensively to achieve the improved production of commodity chemicals and materials. In addition, a wide

variety of compounds that are difficult to produce by classical chemical synthesis are treated as biotechnologically attractive targets. Increasingly, rational metabolic engineering is employing the use of advanced analytical tools for identification of appropriate targets for genetic modifications, including the use of mathematical models for design of optimised microbial strains (Burgard et al., 2003; Pharkya et al., 2003; Patil et al., 2005). The recent integration of systems biology, mathematical modelling and synthetic biology with metabolic engineering is also increasing the possibility for better optimised microbial strains for increased yields of chemicals and products. It is noteworthy that, despite the current trend in metabolic engineering, a large percentage of the successes in employing microbial cell factories have been achieved without detailed modelling.

Based upon the above premise, therefore, the application of rational metabolic engineering and systems biology to the production of biotechnologically attractive materials and products represents a credible alternative approach to solving a myriad of problems associated with reliance on fossils. The progress in the shift from fossils to renewable resources stands to benefit immensely, especially from the quantitative and theoretical concepts of systems biology. Hence, the motivation for this thesis, is based on the opportunity to apply cutting-edge *in silico* modelling and systems biology approaches in a rational and systematic manner to the development of optimised production strains capable of improved yields of various bio-products (e.g., bioethanol, amino acids, fumaric acid, and trehalose) in microoganisms, with renewable materials used as the feedstock.

1.1.2 Aims

The aims of this PhD project were:

1. To identify interesting biotechnological products and suitable host strains.

2. To develop ‘proof‘ of principle production strains for a product of interest.

3. To carry out *in silico* examination of the metabolic network of the organism and develop *in silico* strains for improved production of the product.

4. To validate *in silico* hypotheses and *in silico* designed strains by using genetic engineering to construct production strains for improved yield of target product.

1.1.3 Objectives

The following objectives (graphically represented in Figure 1.1) were employed to achieve the aims of the project:

1. Extensive literature searches were carried out to identify value products and suitable hosts.

2. *In silico* host design:

   (a) Characterisation of the network topology of host organism:

   Metabolic pathway analysis (EFM) was used to analyse a network of reactions obtained from genome-scale metabolic network reconstructions (Dobson et al., 2010) and the literature (Cakir et al., 2004). Constraint-based metabolic flux analysis (FBA) was also used to analyse the genome scale metabolic network reconstructions of yeast (Duarte et al., 2004; Mo et al., 2009; Dobson et al., 2010).

   (b) Gene deletion phenotype analysis:

   In silico gene deletion phenotype analysis was carried out to identify gene knockout strategies for improving the yield of a product in yeast (*S. cerevisiae*). The gene deletion strategy was to ensure deletion of competing pathways and hence a redirection of flux to the pathway of interest.

3. Model validation in the laboratory (synthesis stage):

   (a) Genetic engineering methods were used to knock out genes in the base yeast strains (single deletion strain library from EUROSCARF). In-house

made mutant strains were constructed using PCR generated strategies. Double deletion strains from the Boone Lab (Boone Lab, Donnelly Centre, University of Toronto,Toronto, Canada) were also involved.

(b) For maximization of products, the growth physiology of genetically modified yeast strains were assessed and optimal growth conditions established. Strain growth characteristics of the yeast mutant strains were determined.

4. Model validation in the laboratory (analysis stage):

   The changes in the level of the biological product of interest were monitored using a systems biology technique (metabolomics). Measurements of metabolite concentrations were carried out using GC-MS.

5. Optimisation of production strain:

   In an attempt to fine tune and improve performance of production strain, metabolic profiling of production strains were carried using GC-MS so as to discover bottlenecks to metabolic engineering of the production strains for improved yield of target product.

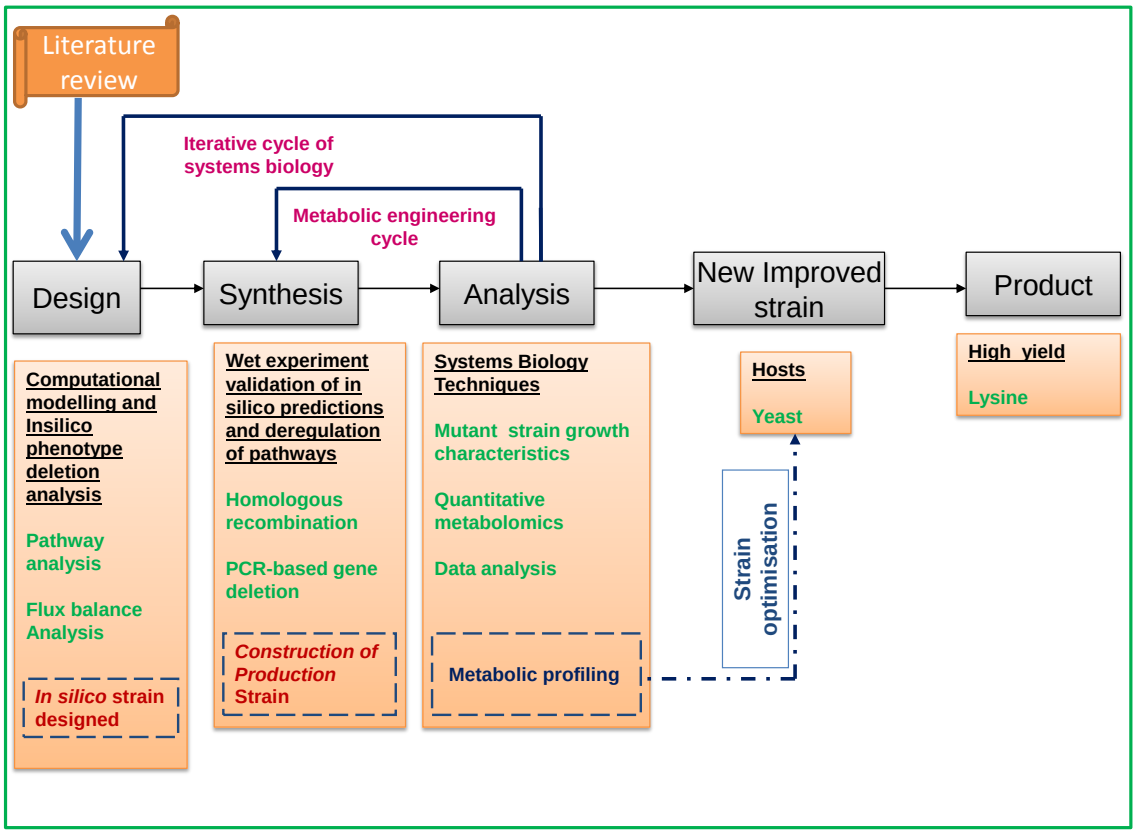

Figure 1.1: **Project pipeline for improved lysine production in yeast.** This schema summarises the entire project pipeline, starting from the identification of products from the literature to design stage, synthesis stage, and finally analysis stage for developing an improved yeast production strain for improved product yield (e.g., lysine). The diagram is further annotated to show different types of work carried for *in silico* modelling at the design stage, wet experiment validation of *in silco* model at synthesis stage, and quantification of product at analysis stage. Metabolic engineering cycle indicates how the results from the analyses of strains were fed back to the synthesis stage, while the systems biology cycle depicts the feedback from the analysis stage to the design stage for improved *in silico* strain optimisation, based on metabolic profiling of production strains.

1.2 Thesis overview

Chapter 1 is the introduction to the work carried out in this PhD research work. Descriptions of the theoretical, computational and experimental strategies currently employed in systems metabolic engineering are presented here. The chapter also covers a literature review of biotechnological production of value chemicals in microbial strains and their methods of production.

In Chapter 2, the description of selection of biotechnological products and host strain

for this PhD research is presented.

Chapter 3 covers the materials and methods for the all the computational modelling studies carried out. Materials and methods for modelling using elementary flux mode analysis for microbial strain development, methodologies for clustering analysis, pattern analysis using regular expression and constraint-based microbial strain development are presented.

In Chapter 4, materials and methods for all laboratory experiments are described here. This chapter contains the description of the synthetic media used for the cultivation of *S. cerevisiae* control and mutant strains. Materials and methods of all genetic manipulations carried out, methods for confirmation of successful gene deletions are covered; the extraction, derivatisation and GC-MS analysis are also presented.

Chapter 5 is a report of the results of constraint-based flux analysis methods for the development of *S. cerevisiae* strains. Results for characterising the genome scale networks of *S. cerevisiae* using FBA and the results for knockout strategies based on OptKnock and GDLS for enhanced production of several target metabolites are included here.

Chapter 6 presents methodologies for deciphering the elementary flux modes (EFMs), including data pre-processing and classification. The chapter also reports how the EFM data was subsequently utilised for *in silico* phenotype gene deletion studies for enhanced lysine production in *S. cerevisiae*.

The construction and the experimental validation of *S. cerevisiae* double mutant strains using genetic engineering is reported in Chapter 7. This report describes the cultivation of mutant strains and the results of gene deletions. The results of experimental validation for 6 single *S. cerevisiae* strains, 5 double mutant *S. cerevisiae* strains and 1 triple mutant *S. cerevisiae* strain are also presented in this chapter.

In Chapter 8, studies carried out for comparing the endometabolome of each of the

constructed *S. cerevisiae* double mutant strains are presented in this chapter. In addition, the discussions covers the differences and similarities in the metabolic profiles of the mutants compared with the control strain.

1.3 Strategies employed in biotechnological production of products

1.3.1 Rational metabolic engineering and systems biology

Metabolic engineering has been defined as the "directed improvement of product formation or cellular properties through the modification of specific biochemical reactions or introduction of new ones with the use of recombinant DNA technology" (Stephanopoulos, 1999). In a slightly different way, expanding the realm of genetic engineering tools, Lee and Papoutsakis (1999) defined metabolic engineering as "directed modification of cellular metabolism and properties through the introduction, deletion, and/or modification of metabolic pathways by using recombinant DNA and other molecular biological techniques". Rational metabolic engineering refers to the engineering of the cellular metabolism based on available information about the pathways, enzymes, and their regulation.

In essence, the main goals of metabolic engineering can be categorised as follows (Kern et al., 2007): (1) Improvement in yield, productivity and phenotype, (2) extension of the substrate range, (3) deletion or reduction of by-product formation and (4) introduction of pathways leading to new products.

Metabolic engineering of the cellular systems of various microbial strains for improved production of products have evolved over the decades. The traditional approach involves metabolic pathway manipulation by way of classical breeding and random mutation followed by selection (Parekh et al., 2000). The advent of recombinant DNA technology made possible a replacement approach of metabolic engineering involving

the introduction of targeted genetic changes whereby gene deletion and overexpression of genes are employed to develop strains capable of higher yield of metabolites of interest. However, the inherent limitations of this method of metabolic engineering also became apparent. There is now a recognition that rational metabolic engineering requires a global approach incorporating integrated pathways and networks. The application of mathematical and computational methods becomes desirable owing to the intricate interconnectedness of enzyme-catalysed reactions, pathways and integrated networks of pathways (Torres and Voit, 2002).

Currently, rational metabolic engineering exploits an integrated, systems-level approach for optimizing a desired cellular property or phenotype (Tyo et al., 2007), and this new trend brings the application of systems biology to metabolic engineering into focus. Systems biology aims at understanding a global picture of the various networks in biological systems by integrating information generated from high-throughput experiments and computational modelling and simulation (Barrett et al., 2006). This approach provides a means of bridging the gap between molecules and physiology by elucidating how dynamic interactions give rise to function (Bruggeman and Westerhoff, 2007). The approaches being exploited by systems biologists such as large-scale functional genomics methods, genetic analysis of model organisms, bioinformatics, mathematical and computer modelling have become important in the toolbox for metabolic engineering of microorganisms.

Rational metabolic engineering usually involves three stages (Figure 1.2). The first step is the identification of targets for genetic modification based on metabolic networks. The strategy at this phase usually involves the cataloguing of biochemical reactions of the host organism from the literature (KEGG, SGD, textbook, metabolic network reconstructions etc); *in silico* modelling and simulations (e.g., metabolic flux analysis, flux balance analysis, elementary flux mode analysis, kinetic modelling, metabolic control analysis) are then carried out to either knock out competing pathways or extend metabolic pathways with the introduction of new enzymes. In the second stage, synthesis involves carrying out genetic modification with the aid of

plasmids or genomic alterations in order to effect in the host strain the genetic alteration(s) suggested by the in silico modelling and simulations. At synthesis stage, gene deletions, increases in enzyme activity and novel pathways may be incorporated into the strain. This stage also represents the interface of systems metabolic engineering with the other disciplines. Examples are the systematic pathway design using synthetic biology, the design of new catalytic function using protein engineering, and selection of a desired phenotype in strains using evolutionary engineering. In the third stage, analysis, whereby the recombinant strain is probed for physiological characteristics (e.g., growth in media) and product yield (metabolomics). Transcriptomics and flux analysis may be carried out for further characterisation of the developed strain and the results are fed back to the analysis stage for further genetic improvement of the strain, and this cycle is repeated until the construction of an optimal strain is complete.

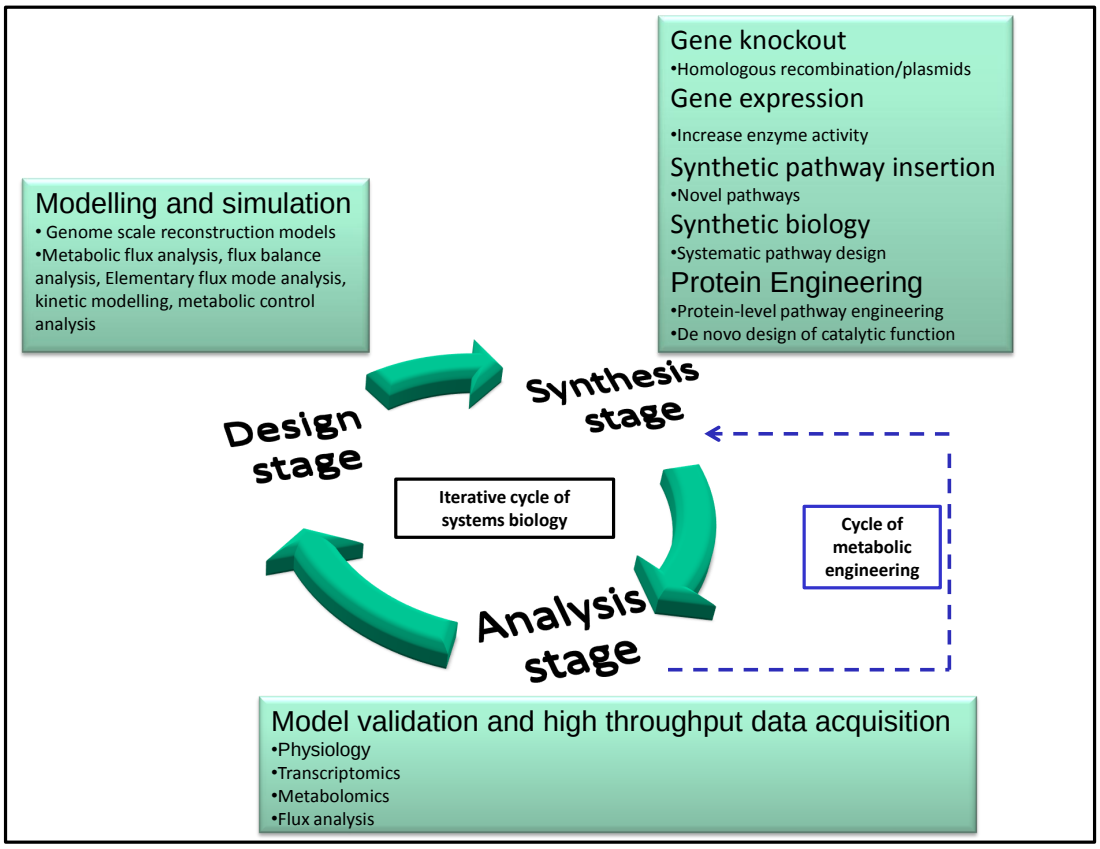

Figure 1.2: **Cycles of metabolic engineering.** Figure depicts the fusion of systems biology with metabolic engineering for biotechnological production of products, including the various methods and techniques currently available. Systems biology cycle in metabolic engineering iterates between three stages: design, synthesis and analysis, while the metabolic engineering cycle iterates between the synthesis stage and the analysis stage. Identification of genetic targets occurs at the design stage, and genetic changes are incorporated into host strain at the synthesis stage, and finally the recombinant strains are analysed for improved yield of product of interest.

1.3.2 Systems metabolic engineering modelling approaches

As indicated above, it is now a common practice to carry out the first step of target identification of targets (Figure 1.2 - Design stage) in rational metabolic engineering of microbial strains by employing *in silico* design strategies such as the characterisation of the metabolic network space using a quantitative method. A quantitative method can be used to probe the network topology analysis to identify gene knockout strategy for developing strains capable of overproducing value chemicals and materials. Further strain optimisation can benefit from coupling of experimental data from the analysis stage (Figure 1.2) to these modelling strategies. Non-linear optimization

of biochemical pathways (Mendes and Kell, 1998), metabolic flux analysis (Iwatani et al., 2008; Maertens and Vanrolleghem, 2010), flux Balance Analysis (Savinell and Palsson, 1992; Varma and Palsson, 1993) and metabolic pathway analysis (Schuster et al., 1999) are the commonest modelling approaches for the characterisation of strains for metabolic engineering purposes.

1.3.2.1 Theory of metabolic network analysis

Metabolism is the chemical engine that drives the living system. A metabolic network can be described as a collection of enzyme-catalysed reactions and transport processes whereby substrate metabolites are consumed and final metabolites are generated (Schilling et al., 2001). The reactions involved in the transformation of one metabolite to the other are referred to as internal reactions, while those involved in the transport of metabolites in and out of the systems are referred to as exchange reactions. It is possible to carry out a quantitative description of metabolic networks describing the transient behaviour of metabolite concentrations, based on the dynamic mass balances of each metabolite to generate a system of ordinary differential equations as follows:

$$\frac{dX_i}{dt} = \sum_j S_{ij} v_j \qquad (1)$$

where v_j is the j_{th} metabolic flux, $[X_i]$ corresponds to the concentration of the metabolite, and the stoichiometric coefficient, S_{ij}, represents the number of moles of metabolite i formed (or consumed) in one cycle of reaction j. The steady state in a chemical network is defined by a constant concentration of metabolites. Mathematically, it is expressed by a set of homogeneous linear algebraic equations in matrix form as follows (Schilling et al., 2001):

$$\mathbf{Sv} = 0 \qquad (2)$$

The stoichiometric matrix $\mathbf{S}$ is an m x n matrix, where m is the number of metabolites and n is the number of reactions or fluxes within the network, while $\mathbf{v}$ is a

vector with the activity of each flux. Furthermore, the null space of **S** represents the capabilities of a given metabolic phenotype, including the building blocks that can manufactured, the efficiency of the energy extraction and conversion of carbohydrates into biomolecules for a given substrate, and also where the critical links are in the network (Varma and Palsson, 1994). Decomposing every reversible reaction into a forward and a backward reaction leaves all reactions with either a positive or zero activity. Therefore there is an additional constraint added for each reaction that their flux must be non-negative:

$$\mathbf{v} \geq 0 \quad (3)$$

The solution space of Equation 2, considered with the inequality constraint, takes the shape of a convex polyhedral cone with a finite number of edges (Figure 1.3) where the edges represent the nonnegative linear combination of the generating vectors of the cone (Schilling et al., 1999).

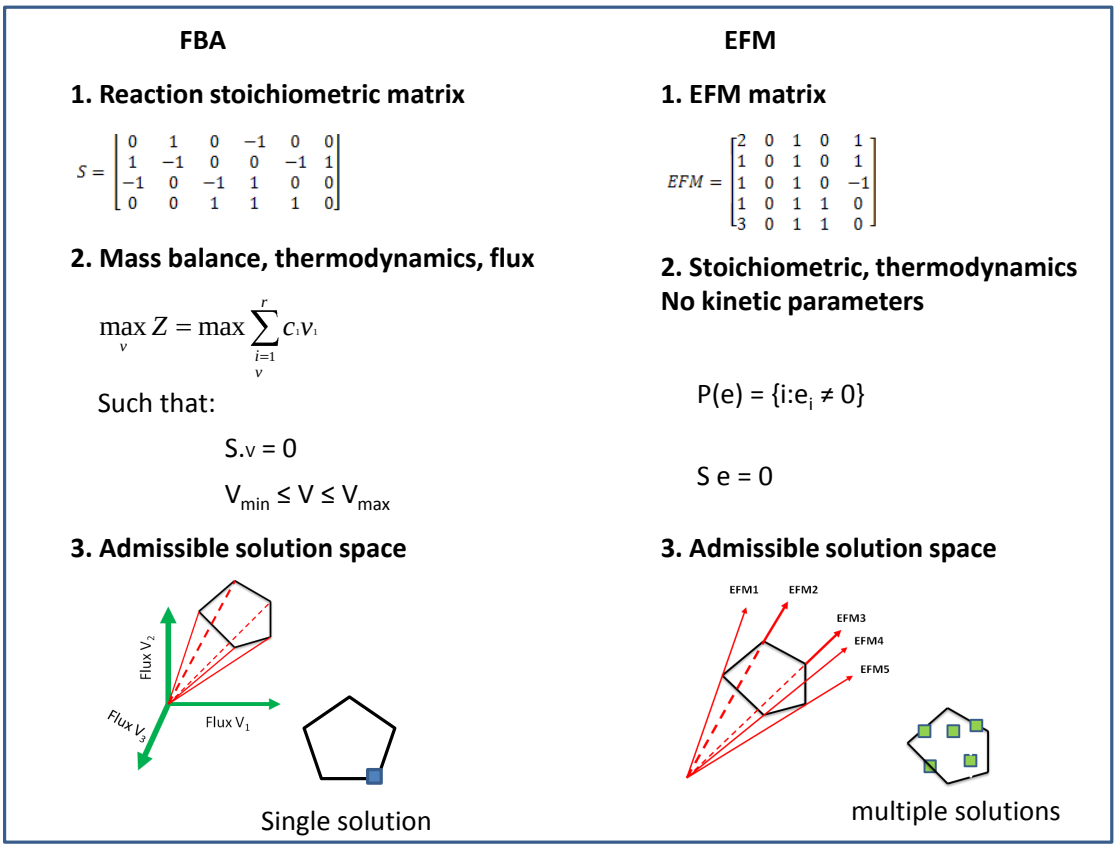

Figure 1.3: **Components of FBA and EFM.** Components of FBA and EFM in terms of matrix, problem statement and convex flux cone of admissible solution space. The edges of the cone contain all permissible fluxes.

For cellular metabolism, Equation 2 is typically an underdetermined system where the number of independent metabolites is fewer than the number of reactions (Trinh et al., 2009). The number of balance equations in Equation 2 depend on the number of metabolites, while the number of reactions in Equation 2 defines the number of unknowns. Given an invariant structure of stoichiometric matrix **S**, coupled with experimentally determined fluxes, the linear Equation 2 together with the inequality constraints (Equation 3) for a metabolic flux vector **v**, can be solved. Metabolic pathway analysis, flux balance analysis and metabolic flux analysis are three methods that can be used for this purpose.

1.3.2.2 Metabolic flux analysis (MFA)

Metabolic fluxes can be used to characterise the phenotype of the cell (Kim et al., 2008). Metabolic flux analysis (MFA) is a very powerful technique for measuring intracellular fluxes. The conventional metabolic flux analysis involves calculation of metabolic fluxes in a metabolic network based on determination of measurable fluxes under a steady-state condition for intracellular metabolites. However, there are shortcomings inherent in this approach in that it does not yield information about parallel or bidirectional metabolic segments, which is important (Iwatani et al., 2008). Furthermore, certain cycle fluxes are not observable and energy metabolites must be balanced in detail (Wiechert, 2002). Hence, an extension of the measurements is carried out with ^{13}C-labelled compounds. In ^{13}C studies, cells are grown in ^{13}C-labelled tracer substrates, until the ^{13}C label has distributed throughout the metabolic network, and formed metabolic products. The ^{13}C labelling pattern identified in specific compounds depends on a particular flux distribution, which can be used for calculation of fluxes (Kohlstedt et al., 2010). Either mass spectrometry (MS) or nuclear magnetic resonance (NMR) spectroscopy is then used to quantify the labelling pattern. Using the ^{13}C-based flux analysis, reliable *in vivo* fluxes can be calculated based on isotope-labelled such as ^{13}C-labelled glucose (Nielsen, 2003; Sauer, 2006), which may be useful for analysing metabolic change or state as a result of genetic engineering (Iwatani et al., 2008). An example of application of ^{13}C-MFA to improve the yield of amino acids in microorganisms include increased lysine production (Ohnishi et al., 2005).

1.3.2.3 Constraint-based flux analysis using Flux Balance Analysis (FBA)

Computational tools have been developed to probe the system properties of genome scale models. Flux balance analysis is a widely used approach for studying biochemical networks (Duarte et al., 2007; Feist et al., 2007; Feist and Palsson, 2008; Oberhardt et al., 2009; Dobson et al., 2010). FBA is based on convex analysis and

employs a linear programming (LP) optimisation method to determine the metabolic flux vector optimal for a defined objective subject to a set of underlying physicochemical (e.g., substrate uptake rates, and/or product secretion rates) and thermodynamic constraints. In a metabolic network, with reaction flux column vector $\mathbf{v}$ and metabolites (row) $\mathbf{x}$ in matrix $\mathbf{S}$, the system of mass balance equations at steady state (dx/dt = 0) is represented as $\mathbf{Sv} = \mathbf{0}$. The constraints due to stoichiometry and the reaction bounds (upper and lower) define the allowable solution space. In practical terms, FBA uses linear programming to solve the equation $\mathbf{Sv} = 0$, given a set of upper and lower bounds on $\mathbf{v}$ and a linear combination of fluxes as an objective function, to give an output of a flux distribution, $\mathbf{v}$, which maximizes or minimizes the objective function (Orth et al., 2010). For example, in order to use FBA to predict maximum growth rate, objective functions take on the form (Figure 1.3):

Maximise Z = $\mathbf{c}^T\mathbf{v}$ (a linear combination of fluxes)

subject to:

$\mathbf{Sv} = \mathbf{0}$ (mass balance equation)

$V_{min} \leq \mathbf{v} \leq V_{max}$ (reaction bounds)

where $\mathbf{c}$ denotes a row vector of coefficients (weights) that multiply into the column vector.

In general terms, the standard form of linear programming (LP) is formulated as follows:

$max_{\mathbf{v}}\ \mathbf{c}^T\mathbf{v}$

Subject to: $\mathbf{Sv}$= b

$\mathbf{v} \geq = 0$

where

$\mathbf{v}$ = vector of variable to be determined (decision variables)

$\mathbf{S}$ = matrix of known coefficients

$\mathbf{b}$ = vector of known coefficients

$\mathbf{c}$ = vector of weights

$\mathbf{c}^T\mathbf{v}$ = scalar objective function, that is linear combination of the decision variables. The components of every LP are:

1. An m x n matrix $\mathbf{S}$, where $n > m$ and typically n is much greater than m (i.e, an underdetermined system - columns more than rows for an undetermined problem)

2. $\mathbf{S}$ vector $\mathbf{b} \in R^m$

3. $\mathbf{S}$ vector $\mathbf{c} \in R^n$

$\mathbf{cv}$ is the inner product of the two vectors $\mathbf{c}$ and $\mathbf{v}$, and $\mathbf{Sv}$ is the product of the matrix S and vector v. Hence, it is possible to find the maximum value that the inner product cv can attain as v runs through all feasible vectors $\mathbf{v} \in R^n$ with nonnegative components ($\mathbf{v} > 0$) satisfying the additional, and important restriction $\mathbf{Sv} = \mathbf{b}$.

Flux balance analysis identifies only one optimal solution in the presence of co-existing alternative optimal solutions or suboptimal solutions (Trinh et al., 2009). The optimal path, in this case, represents only a pathway that lies in the vertex of the admissible flux cone (Figure 1.3), satisfying the defined objective function. Notable applications of FBA are in the field of metabolic engineering. In this project, FBA is used for predicting the flux distributions for various gene (and reaction) knockout or knockdown conditions, enabling the identification of specific changes that may facilitate optimal yield of a particular product, which may then lead to experimental design of the desired phenotype (Gianchandani et al., 2010). Examples of the use of flux methods to engineer *Escherichia coli* strains include overproduction of threonine (Lee et al., 2007), valine (Park et al., 2007), lactic acid (Fong et al., 2005) and succinic acid (Lee et al., 2005a). Becker *et al.* (2007) developed a constraint-based reconstruction and

analysis tool box (Cobra Toolbox) based on constraint-based analysis of genome scale models for quantitative prediction of both steady-state and dynamic optimal growth behaviour, the effects of gene deletion of cellular behaviour and sampling range of possible metabolic states. Other similar tools are MetaFluxNet (Lee et al., 2005b) and CellNetAnalyzer (Klamt et al., 2007).

A number of computational extensions based on FBA framework have been developed. An FBA formulation termed MOMA (minimization of metabolic adjustment), using a quadratic equation to find flux states of mutants, was successfully used to predict viability and quantitative flux distribution in *E. coli* after simulated gene knockouts (Lee et al., 2006a). MOMA was also used in gene knockout simulations to refine candidate mutations for increasing production of lycopene (Alper et al., 2005). ROOM (regulatory on/off minimization) formulation minimizes the number of significant flux changes from the wild-type flux distribution in order to identify metabolic flux state of mutants (Shlomi et al., 2005), and the resulting optimization problem was solved by mixed integer linear programming (MILP). With the introduction of a further extension of bi-level formulation of FBA called OptKnock, it has become more practicable to predict gene deletion strategies for the overproduction of a desired metabolite (Burgard et al., 2003). OptKnock solves for the optimal flux distribution that simultaneously optimizes two objective functions, biomass growth and secretion of the desired metabolite using a bi-level optimization, and have been experimentally validated in *E. coli* for the production of lactate (Fong et al., 2005). OptGene (Patil et al., 2005), an improved OptKnock algorithm using genetic algorithm to reduce computational time, allows identification of target genes to be knocked out, while OptReg (Pharkya and Maranas, 2006) extends OptKnock to allow for up- and/or downregulation and knockout strategy for improving product yield. OptOrf (Kim and Reed, 2010) Identifies metabolic engineering strategies for overproduction of desired metabolite based on gene deletion and overexpression. GDLS (Lun et al., 2009) is a heuristic algorithm search for knockout strategy which combines the strengths of sequential approach with those of bi-level FBA (Burgard et al., 2003; Pharkya

et al., 2004; Pharkya and Maranas, 2006). Optstrain (Pharkya et al., 2004) implements a combinatorial optimization on a multi-species database and then suggests the non-native functionality and the gene deletions for maximising product yield in the production strain.

1.3.2.4 Practical considerations for FBA

In FBA, a metabolic network of reactions is represented mathematically as a stoichiometric matrix (**S**) of size m X n, that is m rows of metabolites and n columns of reactions. The entries in each column are the stoichiometric coefficients of the metabolites participating in a metabolic network of reactions. The coefficient is negative for each metabolite consumed, positive for every metabolite produced, and zero for every metabolite not participating in a particular reaction. Since most biochemical reactions involve only a few different metabolites, **S** is a sparse matrix (Orth et al., 2010).

Vector **v** with length n describes the flux through all of the reactions in a network, while vector **x** with length m defines the concentrations of all metabolites. At steady state, the system of mass balance equations is as defined in section 1.3.2.1:

$dX/dt = 0$

$\mathbf{Sv} = 0$

Since the system is characterised by $n > m$, there are more unknown variables than equations, so there is no unique solution to this system of equations. The stoichiometries impose constraints on the flow of metabolites through the network. Constraints can either be in form an equation describing balanced reaction inputs and outputs and as inequalities that impose bounds on the system. Even though a range of solutions is possible in the solution space, it is possible to use FBA with linear optimization to find the optimal solution at the edge of the polyhedral cone. Apart from the constraint imposed by the stoichiometry, ensuring $\mathbf{Sv} = 0$ at steady state in the system, upper and lower bounds can also be applied to define the maximum and minimum

allowable fluxes of the reactions

The goal of FBA is to minimise or maximise an objective function $Z = \mathbf{c}^T\mathbf{v}$, which can be any linear combination of fluxes, where $\mathbf{c}$ is a vector of weights indicating how much each reaction contributes to the objective function (Orth et al., 2010). Equation $\mathbf{Sv} = 0$ is then solved using Linear programming, subject to a set of upper and lower bounds on $\mathbf{v}$ and linear combination of fluxes as an objective function. The output of FBA is a set of flux distribution, v, which maximises or minimises the objective function. $\mathbf{c}$ is a vector of weights indicating how much each reaction ($\mathbf{v}$) contributes to the objective, when minimising or maximising for a reaction. In the case of minimizing or maximizing the biomass reaction for biomass production, $\mathbf{c}$ is a vector of zeros with a value of 1 at the position for biomass.

It is possible to identify alternate optimal solutions either by using Flux variability analysis (FVA), a method that uses FBA by minimizing or maximizing every reaction in a network pathways (Mahadevan and Schilling, 2003) or by implementing a mixed-integer linear programming-based algorithm (Lee et al., 2000). Robustness analysis permits more detailed phenotypic analysis involving studying the effect of varying one of the objective functions of interest and when two fluxes are varied simultaneously, a phenotypic phase plane (Edwards et al., 2002) can be formed.

1.3.2.5 Pathway analysis

Pathway analysis is a powerful theoretical method applicable to rational metabolic engineering, as it allows for the consideration of the metabolic network topology. Two of the most prominent concepts applicable to pathway analysis are based on elementary flux mode and extreme pathways. Elementary flux mode analysis, as one of the metabolic pathway analysis tools (Schuster et al., 1999) allows for the calculation of solution space containing all possible steady-state flux distributions of a network (Kromer et al., 2006). All metabolic capabilities in steady states are composed of elementary flux modes (EFMs). EFMs are unique for a given metabolic

network and are minimal sets of reactions, each of which can generate a valid steady state (Schuster et al., 1999). In a metabolic network represented by stoichiometric matrix **S** of m rows and q columns, EFM is defined by flux vector **e** composed of q elements (e_1,e_2,...e_q), each describing the net rate of the corresponding reaction. The pathway represented by **e** can be identified by the utilised reactions, denoted as follows (Klamt and Stelling, 2006):

$$\mathrm{P}(\mathbf{e}) = \{i{:}e_i \neq 0\}$$

Hence pathway P(**e**)satisfies all reactions that participate in EFM **e**. The concept of EFM is associated with three fundamental conditions, namely a steady-state condition, a feasibility condition and a non-decomposability condition (Klamt and Stelling, 2003). In steady state, none of the metabolites is consumed or produced according to Equation 2 (section 1.3.2.1). Hence, EFM **e** is in the null-space of S, thereby fulfilling $S\mathbf{e} = 0$.

EFM analysis does not require the knowledge of measured fluxes for the identification of existing metabolic flux vectors in metabolism. It is possible to use EFM analysis to carry out a rigorous examination of network topology in terms of the identification of all possible pathways (Trinh et al., 2006). EFM analysis identifies all unique solutions in the admissible flux space (Figure 1.3). Hence, the analysis of the flux vectors of individual pathways allows the identification of the most efficient pathway for the production of a metabolite of interest. The first reported application of EFMs for biotechnological yield of products was in the optimisation of production of 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) in *E. coli* (Patnaik and Liao, 1994) and the optimal production of L-methionine in *E. coli* and *Corynebacterium glutamicum* (Kromer et al., 2006). There are also reports of the use of elementary flux mode analysis to identify knockout strategies, for example in the development of *E. coli* strains producing ethanol from hexoses (Trinh et al., 2008), ethanol from glucose (Trinh and Srienc, 2009) and carotenoids from glucose (Unrean et al., 2010). PySCeS (Olivier et al., 2005), YANA (Schwarz et al., 2005), COPASI (Hoops et al.,

2006), Metatool (von Kamp and Schuster, 2006) and CellNetAnalyzer (Klamt et al., 2007) are among the available software tools for computing EFMs from a network of reactions. Furthermore, (Terzer and Stelling, 2008) introduced a program written in Java (EFMtool) and implemented in MATLAB with the capability to compute large-scale EFMs based on the concept of bit pattern trees.

The analysis of EFM data has several issues: (1) calculation of the EFMs in large metabolic networks is computationally expensive (Boghigian et al., 2010); (2) EFM computation is characterised by a combinatorial explosion of the number of EFMs (Klamt and Stelling, 2002); (3) analysis of EFM data is a difficult task (Rocha et al., 2010), in that a lot of effort is required to process and interpret the large number of EFMs in order to obtain useful information for metabolic engineering purposes. A number of strategies have already been proposed to solve the intractable problem of the large number of EFMs in order to obviate their limitations for use in metabolic modelling and analysis (Schmidt et al., 2003; Schwartz and Kanehisa, 2005; Schuster et al., 2002; Song and Ramkrishna, 2009). Aggregation around common motif (AcoM) is a method developed for classifying elementary flux EFMs into subsets based on substructures, which aids biological interpretation (Peres et al., 2006; Peres et al., 2011). However, clustering analysis has been used to classify Petri net t invariants (subnetworks) into biologically meaningful groups (Grafahrend-Belau et al., 2008). Even though these methodologies exist to help break the large original EFM dataset into smaller subsets, they lead to EFM subsets that either are not sufficiently reduced for metabolic modelling or do not yield enough useful information that would enable the biotechnologist to make the best decisions for the design of an optimal microbial strain. Hence, in order reduce the problems encountered with the use of EFM for biotechnological production of fuels and value chemicals, there is a need to develop computational algorithms and methods for the processing and classification of EFM data.

1.3.2.6 Kinetic modelling and metabolic control analysis

The concept of "rate-limiting step" dominated the traditional pathway engineering approaches for changing the flux or the concentration of a particular metabolite in a metabolic pathway (Moreno-Sanchez et al., 2008). Finding the correct targets for modification in microorganisms for biotechnological improvement of product yield with the aid of either intuitive or qualitative approaches is usually complicated by compartmentation, cofactor coupling, allosteric effects etc. Unfortunately, one of the shortcomings of stoichiometric network analysis is that it does not account for kinetics and the regulation of the enzymatic reactions. Kinetic models, unlike FBA and EFM, enable predictions based on how changes in kinetic properties, enzyme and metabolite concentrations affect fluxes through metabolic pathways. Hence, analysis of kinetic models allow for identification of changes necessary for improving cellular phenotypes.

A set of differential equations can be used to describe the time dependence of the metabolite concentrations. In kinetic modelling, kinetic equations are incorporated in metabolic models based on mass balances of extracellular and intracellular metabolites, expressed in the general form as follows (Maertens and Vanrolleghem, 2010):

$$\frac{dx_{S_i}}{dt} = D(x^0_{S_i} - \frac{x_X}{\rho_X}) \sum_j S_{M_{ij}} r_j \tag{1.1}$$

$$\frac{dx_{M_i}}{dt} = \sum_j S_{M_{ij}} r_j - \mu x_{M_i} \tag{1.2}$$

where:

x_{M_i} = the concentration of intracellular metabolite M_i

x_{S_i} = the concentration of extracellular metabolite S_i

$S_{M_{ij}}$ = the stoichiometric coefficient of metabolite M_i in reaction j

r_j = rate of reaction j

ρ_X = specific weight of biomass

x_X = biomass concentration

D = dilution rate

$x^0_{S_i}$ = the concentration of extracellular metabolite S in the feed

μ = specific growth rate

μx_{M_i} = dilution effect due to growth

In mechanistic dynamic modelling of metabolic models, complex mechanistic equations describing rate equations (r_j) are used, such as Michaelis-Menten kinetics.

Alternative approaches employing non-mechanistic kinetics for approximate modelling include power law approximation (Savageau, 1976), the loglinear approximation (Hatzimanikatis and Bailey, 1996; Hatzimanikatis and Bailey, 1997), the linlog approximation (Visser and Heijnen, 2003; Visser et al., 2004) and convenience kinetics (Liebermeister and Klipp, 2006).

Metabolic control analysis (MCA) is a mathematical framework for describing biochemical networks. MCA quantifies how steady-state fluxes and concentrations change in response to changes in network parameters (e.g., changed enzyme level) (Visser and Heijnen, 2002). The MCA framework was developed by Kacser and Burns (1973) and by Heinrich and Rapport (1974). In MCA, different coefficients, global and local, are defined, and they are the quantitative indices of the effects of perturbations on fluxes, concentrations and rates. Different coefficients can be distinguished - control coefficients,kinetic response coefficients and elasticities. Control coefficients are a measure of how a change of an enzyme level affects a flux or concentration. The flux control coefficient for a change in the amount of enzyme *xase* on a flux J_{ydh} measured through step *ydh* is defined by (Fell, 1992):

$$C_{xase}^{J_{ydh}} = \frac{\partial J_{ydh}}{\partial E_{xase}} . \frac{E_{xase}}{J_{ydh}} = \frac{\partial ln J_{ydh}}{\partial ln E_{xase}} \tag{1.3}$$

In the case of concentration control coefficients, the variable affected by the chosen parameter such as enzyme *xase*, is a metabolite concentration, S. Concentration control coefficients are given by (Fell, 1992):

$$C_{xase}^{S} = \frac{\partial S}{\partial E_{xase}} . \frac{E_{xase}}{S} = \frac{\partial ln S}{\partial ln E_{xase}} \tag{1.4}$$

The change in the rate of reaction as a result of change in metabolite level, while everything else is constant, is described by a local property called the elasticity. Response coefficients are used to describe the effect of a change of an external parameter, such as the concentration of an extracellular component, on fluxes and concentrations. Summation theorems provide relations between the various control coefficients of a network such that they add up to one (flux-control) or zero (concentration-control). This results in that the control coefficients depend on one another, thus are the system properties (global). The relationships between elasticities and control coefficients are best described by the connectivity theorems.

Kinetic modelling in combination with MCA was used for increased flux in the metabolic engineering of *Lactococcus lactis* for improved lactic acid production (Hoefnagel et al., 2002). Yang *et al* (1999) developed a model of *C. glutamicum* for intracellular lysine synthesis in batch fermentation and used MCA to determine the control on the overall lysine synthesis flux exerted by individual enzymatic reactions.

1.3.3 Experimental approaches to Systems metabolic engineering

Traditionally, multiple rounds of random mutation and selection have been used to develop production strains for various products. However, targeted metabolic engineering employs genetic manipulation tools, including gene deletion, gene expression

tuning, protein engineering and evolution. Gene deletion is used to redirect carbon flux to the pathway of target product by deleting genes in competing pathways. Homologous recombination is the most common gene deletion strategy. Similarly, gene expression tuning, is a commonly used genetic manipulation technique for metabolic pathway re-design. Protein engineering has been used to increase enzyme activity.

Genetic engineering techniques have been used to achieve the following rational metabolic engineering strategies:

1. Removal or inhibition of enzymes to ensure blockage of competing pathways (Shimada et al., 1998).

2. Enhanced carbon flow to the primary metabolic pathways from central metabolism (Stephanopoulos and Sinskey, 1993).

3. Amplification of gene(s) to improve the synthesis of existing products (Pines et al., 1997).

4. Heterologous expression of enzyme(s) to extend the substrate range (Panke et al., 1998) or to produce a novel product (Misawa and Shimada, 1997).

5. Modification of secondary metabolic pathways as necessary to enhance energy metabolism and availability of required enzymatic cofactors.

6. A combination of modifications required to achieve the goal.

It is noteworthy that recent advances in transcriptomics, proteomics and metabolomics technologies and computational systems biology have provided metabolic engineering with a more complete understanding of the cell. In addition, advances in synthetic biology enable the implementation of novel gene networks that are guided by predictive models (Tyo et al., 2010). The combined applications of synthetic biology and systems biology to metabolic engineering is making a new trend whereby the iterative engineering is moving to linear design-based engineering of microbial strains (Tyo et al., 2010). Strain properties have also been improved due to additional tools from protein engineering.

Gene expression profiling enables the analysis of cell physiology and system-wide global regulation at transcript level (Lee and Park, 2010). This can be used to analyse the cell physiology and to select target genes to be engineered as well. Comparative genome analysis can be used to identify the genes to be modified in order to obtain a metabolic phenotype of interest (Lee and Park, 2010). Ohnishi *et al.* (2005) employed this concept to develop an improved lysine producer, APG-4 strain, by introducing a $Ser - 361 - > Phe$ mutation in the *gnd* gene, as a result of information obtained from comparing the genomes of wild-type *C. glutamicum* and an L-lysine producing strain. The final strain, capable of producing 95 g/L of L-lysine after the introduction of *mqo* mutation, was also based on genome comparison.

Proteomics takes a broad, comprehensive and systematic approach to the investigation of protein levels in the biological systems (Lee and Park, 2010). Proteomics attempts to quantify the level of all proteins in the cell using methods such as MALDI-TOF for peptide mass fingerprinting and electrospray (ESI), fourier transform ion cyclotron resonance (FT-ICR) coupling with tandem mass spectrometry (MS/MS) for peptide identification. Comparative proteome analysis of *Mannheimia succiniciproducens* was used to obtain gene targets for manipulation for enhanced production of succinic acid in the same organism (Lee et al., 2006b).

Metabolomics focuses on the identification of metabolites and the measurement of metabolite levels. It can be used to obtain a comprehensive picture of the biological system (Oldiges et al., 2007). Metabololomics provides a platform for experimental validation of *in silico* designed strains and for further optimisation of engineered microbial strains for biotechnological production of bioproducts. Metabolite profiling allows comparative analysis of physiological states under different conditions, such as a wild-type versus its mutants, under different culture conditions (Lee et al., 2006b). Metabolite profiling is an analytical method for relative quantification of selected metabolites either from specific pathways or compound classes in biological samples (Fiehn, 2002). Metabolite profiling is focused on (semi)-quantification of a group or groups of chemical functionalities after minimal sample preparation.

It is also characterised by lower analytical precision than targeted analysis. Quantification in metabolite profiling is based on comparing samples against a reference sample. Metabolite profiling involves techniques such as gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) or capillary electrophoresis-mass spectrometry (CE-MS) which permit for separation of individual components by one technique and their identification by the other (Griffin, 2004). These techniques provide detailed chromatographic profiles of the sample, and either relative or absolute quantifications of components in the sample. Production strain improvement by metabolic engineering can be based on targeted approaches of metabolic profiling or target analysis.

The complex components of biological samples are first separated into simpler components based on the interactions with the stationary (analytical column) and mobile phases (carrier gas) of GC-MS. Samples injected into GC-MS become vaporised and then migrate through the GC column allowing individual components to interact with the stationary phase to varying degrees. The separated components are eluted from the analytical column and then ionized into molecular ions using an electron source before being introduced into the mass spectrometer. Mass spectrometer generates ion fragments from metabolites which are subsequently separated according to their mass to charge ratio (m/z) and detected to generate a mass spectrum of the eluent peaks. Each eluted peak in the sample yields a fragmentation profile, characteristic of the molecular mass and structure of the metabolite. Finally, compounds are identified based on ratios of isotopes and their distribution and also functional groups characterizing the fragment ions. Comparison of fragmentation ions with similar information in mass spectral databases aid identification of compounds.

Determination of the fluxome, the entire set of metabolic fluxes, can provide the understanding the cellular metabolic capacities/activities under various conditions (Lee et al., 2006b). Wang *et al.* (2006) combined constraint-based flux analysis and MFA to identify target knockout, leading to the construction of a new strain capable of high yield of 1.29 mol succinate/mol glucose. ^{13}C tracer experiments based

on different substrate labellings were combined with proteomics for determining the flux distributions in lysine-producing *C. glutamicum* ATCC 21526 (Wittmann et al., 2004); their findings suggested the reduction of secretion of dihydroxyacetone and glycerol as metabolic engineering targets for optimising the strain for improved L-Lysine yield.

There are various examples of combined omics analysis for enhance product yield in mircoorganisms (Askenazi et al., 2003). Lee *et al* (2005a), combined comparative genome analysis of mixed-acid-fermenting *E. coli* and succinic acid-overproducing *M. succiniciproducens* with *in silico* metabolic analysis for the development of optimised strains for enhanced succinic acid yield. The results of the study indicated that combinatorial disruption of five genes, *ptsG*, *pykF*, *sdhA*, *mqo*, and *aceBA* identified by genome comparative analysis alone did not redirect flux to benefit increased succinic acid production as expected. However, further fine-tuning with *in silico* metabolic analysis based on linear programming indicated the triple knockout combination of *ptsG*, *pykF*, and *pykA* for enhanced succinic acid production. The triple-knockout strain was experimentally verified to show over seven-fold increase in succinic acid production.

1.3.4 Host organisms for metabolic engineering

The most industrially important host organisms (biocatalysts) for metabolic engineering are yeast and bacteria. Much has been carried out in both academic laboratories and industry on bacteria and yeast. The development of these microorganisms has been successful partly because they combine the advantages of rapid growth and ease of genetic manipulation.

The yeast *S. cerevisiae* is extremely well suited for biotechnological production of value chemicals, biofuels, pharmaceuticals, materials and food ingredients because it is one of the most intensely studied eukaryotic model organisms, offering large

amounts of data detailing its genetics, biochemistry, physiology, and large-scale fermentation performance (Nielsen and Jewett, 2008). In addition, yeasts are eukaryotes and have ability for protein folding, assembly and post translational modifications, and lack oncogenic and viral DNA. Yeasts are a phylogenetically diverse group of eukaryotic microorganisms (Kurtzman, 1994). Apart from the yeast *S. cerevisiae*, there are alternative yeast whole-cell biocatalysts such as Candida sp.,*Cryptococcus sp.*, *Geotrichum sp.*, *Kluyveromyces sp.*, *Pichia sp.*, alternative *Saccharomyces sp.*, *Schizosaccharomyces pombe*, *Torulopsis sp.*, *Trichosporon sp.*, *Trigonopsis variabilis*, *Yarrowia lipolytica* and *Zygosaccharomyces rouxii* (Pscheidt and Glieder, 2008). However, *S. cerevisiae* is of paramount importance as a biocatalyst because of its GRAS (generally regarded as safe) status (American Food and Drug Administration) and also the fact that abundant data on physiology and genetics of this organism are available. The use of *S. cereviasiae* has beome more popularised by the availability of various advanced genetic techniques, such as highly efficient transformation methods (Gietz and Woods, 2001), high efficiency of homologous recombination, many specialised expression vectors, selectable markers (Guldener et al., 1996; Gueldener et al., 2002; Janke et al., 2004) and immunotags.

On the other hand, bacteria are dominant in the production of metabolites, heterologous proteins, aromatic compounds, amino acids etc, due to the wealth of knowledge about their physiology, genetics, molecular biology, biochemistry and fermentation (Branduardi et al., 2007). *E. coli* possesses a clear genetic background and good growth properties with low nutrient requirements (Yu et al., 2011), and it is becoming increasingly important in biofuel production. Gram-positive bacteria, in particular, *C. glutamicum* have a long industrial history in the production of amino acids.

1.4 Production of value chemicals and materials

The broad range of value products and materials that are produced using biotechnological methods, especially pathway engineering, attests to the ever-growing strengths

of this approach over several decades. In commercial terms, the products range from very cheap bulk chemicals (e.g., ethanol: 38 million ton/year at 400 Euros/ton) to extremely expensive fine chemicals (e.g., vitamin B12: a few ton/year at 25 000 Euros/kg) (Soetaert and Vandamme, 2006). Just as there is increasing concern for the environment and an ever increasing demand for these products, there is also increase in the drive to improve the production strategies and technologies by utilising safer and cost-effective feedstock. Industrial biotechnology is being used as a platform for either replacing a single step in a chemical synthesis or replacing an entire sequence of chemical synthesis steps with one single fermentation or biocatalysis step. While industrial biotechnology is already well established in the production of fine chemicals and pharmaceuticals, there is an increasing trend in the production of bulk chemicals, biofuels and bio-plastics. The choice of microorganism to be used as a biocatalyst for the production of a particular product usually depends on the wealth of available knowledge about the organism, productivity capacity for a desired product by the organism and ease of handling. Consideration for safety is also given a high premium, such as using yeast instead of other organisms, because of its GRAS status, in the biotechnological production of products for human consumption. However, heterologous expression of genes (and an entire pathway) and the use of synthetic biology in the modern era industrial biotechnology renders the consideration for the genetic or physiological suitability of a particular microorganism for producing a product of lesser importance. Demonstrative examples of the successes of metabolic engineering for improved yield of various red and white products are reported in this section. Tables 1.1 and 1.2 show examples of primary and secondary metabolites respectively, indicating the metabolic engineering (with or without systems biology) strategies, the microorganims employed, and the yields of bio-products.

1.4.1 Primary metabolites - Fine and Bulk chemicals

Fine chemicals (e.g., amino acids, vitamins and pharmaceutical products) are commercially important and pure chemical substances, while the bulk chemicals (e.g.,

biofuels) are produced in massive quantities.

1.4.1.1 Amino acids, amino acid intermediate products and derivatives

The amino acids for feed and food application can be divided, based on their method of synthesis, into four different categories, namely: (1) those synthesized by chemical methods, (2) those synthesised through enzymatic synthesis, (3) Glutamate produced by *C. glutamicum* and (4) the other amino acids such as lysine, threonine, phenylalanine and the rest produced by fermentation (Kramer, 2005).

The ever-increasing demand for L-glutamic acid, L-lysine, L-threonine and others as sources of nutrition has stimulated successful development of bacteria such as *C. glutamicum* and *E. coli* as biocatalysts for the production of these amino acids (Sprenger, 2007). Glutamate is the most widely used amino acid and mainly produced in *C. glutamicum* with annual production estimated to be in excess of 1.5 million tons per year (Schultz et al., 2007). *C. glutamicum*, *E. coli* and other microorganisms are able to excrete glutamate.

L-Phenylalanine is produced by chemical, enzymatic or microbial processes (Sprenger, 2007), and it is used in the synthesis of the low calorie sweetener, aspartame, a methyl ester of the dipeptide L-aspartyl-L-phenylalanine). Commercial interest in L-tryptophan is also increasing for various uses. Low cost and the possibility to obtain pure L-phenylalanine from microbial production have encouraged metabolic engineering of *C. glutamicum* (Katsumata and Ikeda, 1993; Ikeda and Katsumata, 1999; Liu et al., 2004) and *E. coli* (Berry, 1996; Rueffer et al., 2004; Tatarko and Romeo, 2001; Yakandawala et al., 2008). The best published L-phenylalanine yield is 50 g/l from a genetically engineered *E. coli* strain with feedback-resistant *pheA* (*pheAfb*r) and feedback-resistant *aroF* (*aroFfbr*) genes (Backman et al., 1990).

C. glutamicum (Katsumata and Ikeda, 1993) and *E. coli* (Berry, 1996) are mainly used in the microbial fermentation of L-tryptophan. The engineering of efficient tryptophan organisms entails similar alterations of precursor pathways and of the

biosynthetic pathway to the approach for phenylalanine overproduction. In both *C. glutamicum* and *E. coli*, there are several strictly controlled steps towards the production of L-tryptophan, and hence improved yield of this amino acid requires removal of all the metabolic controls in the common pathways as well as in the L-tryptophan branch. (Azuma et al., 1993) obtained 50 g/l of L-tryptophan after 91 hours in a genetically engineered *E. coli* strain (with plasmid containing trp operon) in fermentation culture containing glucose and anthranilic acid. Using a rational engineering approach, increase in the production of L-tryptophan up to 50 g/l was achieved in a L-tryptophan producing *C. glutamicum* after the first enzyme (3-deoxy-D-arabino-heptulosonate 7-phosphate synthase) in the common pathway towards chorsimate was amplified, followed by sequential removal of intermediates (Katsumata and Ikeda, 1993; Ikeda et al., 1994). A further improvement was obtained by engineering the central metabolism to increase the availability of PEP and E4P as shown in other studies (Ikeda and Katsumata, 1999; Patnaik and Liao, 1994).

L-arginine is a metabolically versatile amino produced by bacterial fermentation. It is a widely used amino acid with applications in the food flavouring and pharmaceutical industries. Industrial production of L-arginine makes use of mutant strains of *Corynebacterium* in microbial fermentations (Ikeda, 2003; Utagawa, 2004).

A L-lysine producing *C. glutamicum* strain, previously obtained by random mutagenesis, was improved for L-isoleucine production by targeted metabolic engineering efforts involving amplification of the feedback-resistant threonine dehydratase and homoserine dehydratase, resulting in a final strain producing 18.1 g/l L-isoleucine (Morbach et al., 1995; Morbach et al., 1996). A pyruvate dehydrogenase-deficient *C. glutamicum* ($\Delta aceE\ \Delta pqo\ \Delta pgi$) overexpressing *ilvBNCE* (Blombach et al., 2007) was improved for valine biosynthesis (Blombach et al., 2008) by deleting *pgo* and *pgi* genes leading to increase in pyruvate and NADPH available respectively, and a final strain of *C. glutamicum*($\Delta aceE\ \Delta pqo\ \Delta pgi$) strain capable of producing 48 g/L of L-valine. The the first reported systems metabolic engineering of *E. coli* for amino acid

production by genome engineering combined transcriptome analysis and gene knockout simulation of the genome-scale model (Park et al., 2007). Based on an *E. coli*, feedback inhibition and transcriptional attenuation controls on valine biosynthesis were removed by site-specific genome engineering, genes of the competing pathways were deleted, and the *ikvBN operon* (involved in first reaction of valine biosynthesis) was amplified; moreover, *ikvCED*, *lrp* and *ygaZH* genes (encode the L-valine biosynthesis genes, a global regulator leucine responsive protein and an L-valine exporter respectively) identified by transcriptome analysis were overexpressed, *lrp* (prevents uptake of L-valine). Further improvements on the strain include a triple knockout ($\Delta aceF$, Δmdh, $\Delta pfkA$) from *in silico* prediction, resulting in a final strain producing 0.378 g L-valine/ g glucose. L-threonine overproducing *E. coli* strain was constructed by a similar systems metabolic engineering approach.

The construction of an L-serine-producing strain of *C. glutamicum* ($\Delta pabABC\,\Delta sdaA$) producing 32.8g/L L-serine was based on targeted metabolic engineering, by overexpressing the feedback-resistant 3-phosphoglycerate dehydrogenase, phosphoserine phosphatase and reduced folate supply (Peters-Wendisch et al., 2005; Stolz et al., 2007).

Transcriptome profiling has been applied to the mapping of metabolic characteristics and identification of genes for metabolic engineering as exemplified by the improvement of L-lysine production by the introduction of mutations identified through transcriptome analysis to define L-lysine producer (Park and Lee, 2008). Transcriptome analysis of a classical L-lysine producer *C. glutamicum B-6* revealed the amino acid biosynthetic genes that were induced (Mitsuhashi et al., 2006) leading to increased lysine production.

Systems metabolic engineering approach was used to develop a genetically defined L-threonine overproducing *E. coli* strain (Lee et al., 2007). Lee *at al* (2007) carried out deletions of *thrA* and *lysC* gene to remove the feedback inhibitions of *aspartokinase I* and *III* respectively, deletions of *metA* and *lysA* genes to ensure precursor availability for threonine biosynthesis, removal of *thrL* transcriptional attenuation regulations,

and deletion of *tdh* in combination with mutation of *ilvA* were carried out to prevent degradation of threonine. In addition to these steps, overexpression of *ppc* gene (encodes phosphoenolpyruvate carboxylase), threonine transporter genes *rhtA*, *rhtB* and *rhtC* (involved in the export of threonine) and *acs* gene (encodes Acetyl-CoA synthetase to reduce acetate accumulation). It was also necessary to carry out the deletion of *iclR* genes (to encourage carbon flux through the glyoxylate shunt), *tdcC* gene encoding a *Thr* transporter (uptake of threonine into the cell)predicted through the use of transcriptome profiling and in silico flux simulations, in order to obtain the final recombinant *E. coli* strain capable of 82.4 g/L of threonine in 50 hours of fed batch culture.

Leinfelder and Winterhalter (1999) successfully engineered a cysteine overproducing *E. coli* strain. L-ornithine is an intermediate in the biosynthesis of arginine that finds application in the synthesis of pharmaceuticals for treatment of liver disorders in humans. Commercial production of L-ornithine is from a citrulline-requiring mutant of a coryneform bacterium (Choi et al., 1996).

D-phenylglycine (D-Phg) has an important application as a side chain building block for semi-synthetic penicillins and cephalosporins (Muller et al., 2006). The current approach for producing D-Phg involves a two-step chemo-enzymatic synthesis (Wegman et al., 2001). Muller *et al.* (2006) reported the first completely fermentative production of D-phg by introducing a heterologous pathway into an L-phenylalanine producing *E. coli* strain involving genes from *Amycolatopsis orientalis*, *Streptomyces coelicolor*, and *Pseudomonas putida*, which together catalyses the conversion of phenylpyruvate via mandelate and phenylglyoxylate to D-phg.

Natural end products of the aromatic amino acid pathway are tryptophan, tyrosine and phenylalanine; these pathways can be extended to yield products such as melanin and indigo (Ensley et al., 1983). Catechol (Draths and Frost, 1991), adipic acid (Draths and Frost, 1994) and quinic acid (Draths and Frost, 1992) are examples of other biosynthetic products that can be generated from intermediates in the aromatic amino pathway incorporating foreign genes from other microorganisms

(Chotani et al., 2000). Hence, the aromatic amino acid pathways are of interest since they present multiple product opportunities contributing to reduced technical and commercial development costs (Chotani et al., 2000).

1.4.1.2 Organic acids

This part of the review covers the metabolic engineering strategies for microbial production of the following organic acids: Fumaric acid, citric acid, succinic acid, malic acid, acetate and pyruvate.

The U.S Department of Energy listed succinic acid, fumaric acid and malic acid in the top 12 of the most interesting chemical building blocks derivable from biomass (Werpy and Petersen, 2004). Three intermediates of the oxidative citric acid cycle (TCA), Fumaric acid, L-malic and citric acids, are synthesized and secreted to high level of concentrations in *Aspergillus* sp. and *Rhizopus* sp. (Goldberg, 2006). Being a naturally occurring organic acid in the TCA cycle, many of the microorganisms are able to produce small quantities of fumaric acid (Roa Engel et al., 2008). Fumaric acid is highly valuable in the food industry as intermediates in the production of L-malic acid and L-aspartic acid (Goldberg, 2006); potential and interesting applications of fumaric acid derives from its non-toxic nature and could be as a better option for the polymer industry over other carboxylic acids (Roa Engel et al., 2008), and also a highly pure fumaric acid could be used to treat psoriasis (Altmeyer et al., 1994). Fumaric acid is currently produced chemically from a petroleum derivative, maleic anhydride; however, the increasing prices of petroleum is encouraging renewed interest in the fermentative fumaric acid production. The organic acid-producing ability of fungi has been exploited in the production of fumaric acid by fermentation (Goldberg, 2006). Genetic modification of microorganisms for production of fumaric acid is sparse, and *R. oryzae* is the micoorganism with the highest productivity and yield of fumaric acid by fermentation. The market size for citric acid is large and it continues to increase, mainly due to is use in the expanding food and beverage and also in the use of health-related products (Goldberg, 2006). Citric acid is the

most important food acidulant, and further specific uses are in the manufacture of pharmaceuticals, wines, ciders, candies, jellies, jams, soft drinks, vegetable juices, toiletries and cosmetics. Global citric acid production is estimated to be 1.4 million tonnes per year, with an annual increase of 3.5 - 4%. Currently, strain improvement for citric acid production involves mutagenesis and selection.

Succinic acid is one of the most important green chemicals and has a broad spectrum of application, including its use as surfactant, ion chelators, food additives, pharmaceutical supplements and antibiotics (Hong, 2007). It is also used in the chemical industry as a precursor to important chemicals, such as 1,4-butanediol, a monomer for various aliphatic polyesters (Willke and Vorlop, 2004; Song et al., 2006; McKinlay et al., 2007). Bacterial strains growing under anaerobic conditions usually produce succinic acid as the main product (Zeikus et al., 1999). Currently, most of the world's succinic acid production is based on a chemical process utilizing fossil materials as feedstock. However,the most efficient succinic acid bacterial strains are *Anaerobiospirillum succiniciproducens* and *Actinobacillus succinogenes*. Lee *et al* (2008) engineered a *M. succniciproducens* strain, LPK7, capable of producing succinate at 1.42-fold higher than the wild-type strain by knocking out the *IdhA*, *pflB*, *pta* and *acKA* genes, albeit causing retarded cell growth, to prevent formation of lactic and formic acids as by-products.

Jantama (2008) used a combination of gene deletions and metabolic evolution to develop derivatives of *E. coli* capable of producing either succinate or malate in simple fermentation consisting of mineral salts. The study reported that the best performing succinate biocatalysts, strains KJ060 (*IdhA*, *adhE*, *ackA*, *focA* and *pflB* deleted) and KJ073 (*IdhA*, *adhE*, *ackA*, *focA*, *pflB*, *mgsA* and *poxB* deleted), produced 622-733 mM of succinate, while strain KJ071 (IdhA, adhE, ackA, pflB and mgsA deleted), the best malate producer, produced 516 mM malate.

Malic acid is a four carbon dicarboxylic acid, widely used in the polymer, food and pharmaceutical industry, and its other non-food applications include metal cleaning and finishing and electroless plating (Goldberg, 2006). It is produced by two different

methods: Maleic anhydride from the petrochemical source, serves as the precursor for the production of racemic maleic acid, while enantiomerically pure L-malic acid is derived from fumarate by an enzymatic process (Chibata et al., 1983). However, one-step fermentation strategy for the production of malic acid with glucose as the feedstock has been established. A shift from the earlier *Aspergillus flavus* fermentation for malic acid to other microorganisms, including yeast, has been stimulated by concern regarding the potential aflatoxin production of this organism (Zelle et al., 2008). Malic acid, with an estimated annual production of around 140,000 tons per year (Mecking, 2004), has a potential to move from its present position as an intermediate-volume chemical to a very large volume, commodity-chemical intermediate derived from renewable sources. Metabolic flux analysis was used to predict increasing production of malate in *E. coli* by amplification of phosphoenolpyruvate (PEP) carboxylation flux, resulting in the development of a WGS-10 strain expressing PEP carboxykinase from *M. succiniproducens*, which produces 9.25 g/l of malic acid in 12 hours of aerobic fermentation (Moon et al., 2008). Zelle *et al.* (2008) engineered a malate producing yeast strain capable of producing 59 g/l at a yield of 0.42 mol (mol glucose). This new strain reflects the introduction of three genetic modifications into a C2-independent pyruvate decarboxylase-negative *S. cerevisiae strain*, namely overexpression of the native pyruvate decarboxylase, high-level of expression of the *MDH3* gene allele for the cytosol-retargeted malate dehydrogenase and the heterologous expression of *Schizosaccharomyces pombe* malate transporter gene.

1.4.1.3 Biofuel

Interest in renewable energy sources have intensified recently as a result of high cost of oil and supply instability, and also because of increasing environmental concerns. Ethanol is one of the candidates for replacing dependence on fossil-fuel, and this enthusiasm for bioethanol productions has been been driven in part by mandates in many countries to use bioethanol and by an abundant supply of corn in the United States (Keasling and Chou, 2008). Due to inputs from advances from genomics

and technology, metabolic engineering has emerged as the leading tool for deriving renewable energy. Diverse biomass resources are available as sources of feedstocks for microbial conversion into useful products and materials in form of agricultural lignocellulosic residues, edible and non-edible crops and waste.

The most commonly used feedstocks for the industrial production of biofuels are starches and simple sugars derived from sources such as sugar cane (sucrose) and corn (starch). However, the current production of starch-based ethanol is unsustainable due to competition from food and animal feed industry. Hence, other cost-effective and sustainable feedstocks, such as lignocellulosic sugars and fatty acids are being developed to replace sugar cane and corn as a feedstock for biofuel production. Lignoceric feedstocks from agriculture and forestry are composed of cellulose, hemicellulose and lignin; cellulose is a homopolymer of glucose, while hemicellulose is composed of hexose sugars (glucose, mannose and galactose) and pentose sugars such as xylose and arabinose (Hayn et al., 1993). The conversion of lignocellulosic sugars into fermentable materials is a challenging task than with either sugar cane or corn. This is due to problems in digesting lignin, a highly recalcitrant network polymer of aromatic alcohols which accounts for 25% of most common sources of cellulosic biomass (Aristidou and Penttila, 2000). A pretreatment step is usually required. In the the process of pretreatment of lignoceric feedstocks, cellulosic and hemicellulosic portions are partially hydrolysed into products that are digestible by celluloses, with a concomitant generation of fermentation inhibitors. The presence of inhibitors is a big challenge for industrial production of bioethanol in yeast.

The metabolic engineering strategies that have been applied to laboratory strains of *S. cerevisiae* to improve xylose fermentation involved a number of different modification categorised as follows (Hahn-Hagerdal et al., 2007): modifications of transport (xylose transport), Xylose utilisation pathway (e.g., xylose reductase - XR/ xylitol dehydrogenase - XDH), arabinose utilisation pathways (e.g., Fungal and *E. coli* arabinose pathways), xylose and arabinose combined (e.g., XR/XDH/XK pathways), reducing xylitol formation (Aldolase reductase Gre3 deletion), improving the efficiency

of metabolism (e.g., overexpression of TAL and the non-oxidative PPP) and anaerobic growth on xylose (e.g., evolutionary engineering). Industrial production yeast hosts for ethanol production have been developed based on the xylose and arabinose utilisation pathways, complemented with random mutagenesis (Wahlbom et al., 2003) and evolutionary engineering (Sonderegger et al., 2004). Xylose-fermenting industrial yeast strains for ethanol include TMB 3400 and TMB 3006, engineered with heterologous XR and XDH genes.

Despite many efforts in the genetic engineering of yeast and bacteria for the fermentation of xylose and arabinose, bioconversion of pentoses to ethanol remains a huge challenge. The expression levels of the native *S. cerevisiae* genes for xylose utilization (Deng and Ho, 1990; Kuhn et al., 1995; Richard et al., 1999; Toivari et al., 2004) are not high enough to support growth on xylose. *S. cerevisiae* possess the genes for xylose assimilation, but their low expression prevents significant sugar assimilation (Jeffries and Jin, 2004).

The *Pichia stipitis* genes *XYL1* and *XYL2* encoding XR and XDH, respectively, when introduced in *S. cerevisiae* (Kotter and Ciriacy, 1993; Tantirungkij et al., 1993), resulted in growth on xylose, and when combined with overexpression of the endogenous XKS1 gene encoding xylulokinase (XK) allowed for xylose fermentation (Ho et al., 1998; Eliasson et al., 2000; Toivari et al., 2001).

Although several yeasts and fungi can utilize L-arabinose as a carbon and energy source, most of them are unable to ferment it into ethanol. Overexpressed of all the structural genes of the fungal L-arabinose pathway (XYL1, lad1, lxr1, XYL2, and XKS1) in *S. cerevisiae*, resulted in slow rate of L-arabinose into ethanol (0.35 mg of ethanol g^{-1} h^{-1}) under anaerobic condition (Richard et al., 2003). In a different study, Wisselink *et al.* (2007) combined the expression of the structural genes for the L-arabinose utilization pathway of *Lactobacilllus plantarum*, the overexpression of the *S. cerevisiae* genes encoding the enzymes of the nonoxidative pentose phosphate pathway and extensive evolutionary engineering (Sauer, 2001) to develop a production *S. cerevisiae* strain with a high fermentation rates of arabinose consumption (0.70 g /h

[dry weight]$^{-1}$) and ethanol production (0.29 g h^{-1} g [dry weight]$^{-1}$) and a high ethanol yield (0.43 g g^{-1}) during anaerobic growth on L-arabinose. A high ethanol yield for a recombinant *Saccharomyces* was obtained in a study involving fermentation of 53 g glucose l^{-1} and 56 g xylose l^{-1} mixture by *Saccharomyces* sp. 1400 (pLNH33) mixture to achieve an ethanol concentration of 50 g l^{-1} in 36 h (Krishnan et al., 1999).

In the bacterial pathway, the enzymes L-arabinose isomerase (AraA), L-ribulokinase (AraB), and L-ribulose-5- phosphate 4-epimerase (AraD) are involved in converting L-arabinose into L-ribulose, L-ribulose-5-P, and D-xylulose-5-P, respectively. An ethanologenic *E. coli* strain was engineered with heterologous genes for pyruvate decarboxylase (pdc) and alcohol dehydrogenase (adhB) genes from *Zymomonas mobilis*. The resultant recombinant ferments hexoses and pentoses to ethanol at a high rate and yield (Ingram et al., 1999).

Although ethanol is the major biofuel in the transport sector, it suffers from the disadvantage that it has low energy content compared with petroleum-derived fuels and also it is incompatible with the transportation infrastructure. These drawbacks in the use of ethanol is paving the way for research into advanced biofuel, especially those that can supplement or replace gasoline and biodiesel, such as short-chain alcohols (or alkanes) and biodiesel (or cyclic isoprenoids) respectively (Lee et al., 2008), (Peralta-Yahya and Keasling, 2010).

Several potential advanced biofuels have been successfully produced in microbes, either by fine tuning discrete steps or redirecting flux to the desired production pathway. Despite these successes, the majority of metabolic engineering efforts in this direction are yet to benefit from the application of iterative rounds of systems analysis and metabolic engineering to ensure higher yields of advanced biofuels (Peralta-Yahya and Keasling, 2010). Application of functional genomics for profiling engineered microorganism will provide system-level information about metabolism and fuel toxicity, while metabolic flux analysis may help reveal bottlenecks in the engineered pathways. However, combined gene expression and transcription network connectivity data, genetic knockouts, and network component analysis (NCA) was used to study

isobutanol response network of *E. coli* under aerobic conditions (Brynildsen and Liao, 2009).

The application of molecular, systems and synthetic biology to the engineering of the microbial isoprenoid and fatty acid pathways enhances the feasibility for microbial biosynthetic production of advanced biofuel candidates such as alcohols, esters, alkanes and alkenes from these pathways (Peralta-Yahya and Keasling, 2010). Naturally, various *Clostridium* species produce isopropanol. Isopropanol is currently used as gasoline and diesel additive (http://www.epa.gov/otaq/regs/fuels/additive/webdies.htm, http://www.epa.gov/otaq/regs/fuels/additive/web-gas.htm), and the highest reported isopropanol is from *Clostridium acetobutylicum* is 1.8 g/l, while the highest reported yield for butanol in *Clostrdium beijerinckii*is 19.6 g/l. The highest isopropanol production level of 4.9 g/l 4.9 g/L was achieved by reconstructing the *Clostrdium* isopropanol pathway in *E. coli*, through the combined expression of *C. acetobutylicum thl* and *adc*, *E. coli atoAD*, and *C. Beijerinckii adh* (Hanai et al., 2007). In a reconstructed *Clostridium* isopropanol pathway in *E. coli* (Jojima et al., 2008), *C. Acetobutylicum thl*, *ctfAb*, *adc*, and the *C. beijerinckii* adh were expressed in *E. coli* to produce 13.6 g/l, corresponding to 51% of the maximum theoretical yield in 36 h of cultivation.

There are also success stories about re-routing of biosynthetic pathways for the production of medium-chain alcohols. Re-routing of the amino acid biosynthetic pathways in yeast through the "Ehrlich pathway" provides the opportunity to overproduce higher alcohols, such as propanol and 2-methyl-1-butanol (2MB) from isoleucine via two biosynthetic intermediates, 2-ketobutyrate and 2-keto-3-methylvalerate (KMV), respectively (Peralta-Yahya and Keasling, 2010).

1.4.1.4 Oligosaccharides and derivatives

Microbial synthesis of oligosaccharides and polysaccharides is challenging as it is a carbon- and energy-intensive process with the precursor sugar nucleotides involving multiple interacting pathways. Despite these challenges, a number of successful metabolic engineering efforts to synthesize oligosaccharides of diverse structures have been carried out using engineered *E. coli*, *Pichia pastoris*, *Corneybacterium ammoniagenes* and *C. glutamicum* (Ruffing and Chen, 2006). A microbial coupling approach successfully produced 188 g/l of oligosaccharides (Koizumi et al., 1998).

Trehalose is a stable, odour-free, non-reducing disaccharide composed of two molecules of glucose linked by a alpha-1,10-glycosidic bond. The protein-stabilising properties of trehalose have ensured its wide range of applications, from cosmetics to the food industry. Trehalose is believed to have important metabolic roles as reserve carbohydrate and protective functions against adverse growth conditions such as heat shock, osmotic shock or starvation in *S. cerevisiae*. The view that trehalose may contribute to yeast viability is of great interest to the wine and brewing industries. Trehalose is currently obtained by an imobilized enzyme method based on maltodextrins; however, a few microbiological alternative approaches based on heterologous expression of genes in *C. glutamicum* have been reported. Padilla *et al.* (2004) reported the heterologous expression of the otsBA operon from *E. coli* in a *C. glutamicum* recombinant lacking trehalose-maltose-isomerizing activity. The metabolic engineering strategy resulted in a five- to six- fold increase in the flux of OtsAB pathway and about four-fold increase in the trehalose excretion rate during the exponential growth phase. In a different study, improved trehalose yield was achieved in *C. glutamicum* by employing a simultaneous overexpression of *C. glutamicum* treY and trZ genes in the TrYZ pathway and the *E. coli* galV gene in *C. glutamicum* (Carpinelli et al., 2006).

Glycerol, a sugar alcohol, is a commodity chemical with a wide range of applications

including cosmetics, food, pharmaceuticals, lubricants, antifreeze solutions and tobacco (Chotani et al., 2000). Most of the glycerol is produced biochemically, but glycerol can also be synthesized through chemical means by way of propylene. Glycerol production has been reported in certain species of bacteria, algae, protozoa and yeast (Ben-Amotz and Avron, 1979; Steinbuchel and Muller, 1986; Albertyn et al., 1994). Since the biochemical production of glycerol is inherently problematic and isolation of glycerol from animal fat and other sources are laborious and inefficient, efforts have been placed towards the biotechnological production of glycerol. Glycerol is produced from the glycolytic intermediate, dihydroxyacetone-3-phosphate as a result of catalytic conversions from two enzymes, dihydroxyacetone-3-phosphate dehydrogenase and glycerol-3-phosphatase. The level of glycerol achieved in *Saccharomyeces* during alcoholic fermentation can be increased by osmotic stress. Osmotic stress, which induces production of glycerol as a result of transcriptional activation of GPD1 (Albertyn et al., 1994). Geertman *et al.* (2006) generated the highest glycerol yield on glucose in *S. cerevisiae*, up to 1.08 mol/mol, by co-feeding of formate to aerobic, glucose-limited chemostat cultures of *S. cerevisiae* lacking pyruvate decarboxylase, external NADH dehydrogenase and the respiratory chain-linked glycerol-3-phosphate dehydrogenase. The production of 0.469 glycerol (g glucose) in aerated batch culture of *S. cerevisiae* was achieved by the overexpression of GPD1 in a tpi1D mutant defective in triose phosphate isomerase (Cordier et al., 2007).

1.4.1.5 Bioplastic and precursors for fibres

Synthetic polymers, such as polyethylene, polypropylene, polystyrene and polyvinyl chloride rely almost entirely on fossil fuel as feedstock. Annual global production of ethylene is the highest of all organic compounds (McCoy et al., 2006). Limited amounts of ethylene are also produced biologically in plants (De Paepe and Van der Straeten, 2005; Ecker, 1995), and also some microorganisms are able to produce it naturally (Fukuda et al., 1993). Pirkov *et al* (2008) developed a genetically engineered strain of *S. cerevisiae* expressing *Psedomonas syringae* (plant pathogenic bacterium)

ethylene forming enzyme (EFE) (Fukuda et al., 1992; Goto et al., 1985; Nagahama et al., 1994) which catalyses the formation of ethylene from 2-oxoglutarate, arginine and oxygen (Fukuda et al., 1992). In this study, the highest productivity was achieved during the respiro-fermentative growth on glucose and in addition, when glutamate was used as the source, ethylene production was three times higher than when ammonia was the source of nitrogen.

Poly lactic acid (PLA) is a polymer of lactic acid isomers currently regarded as potential replacement for conventional petroleum-based plastics as renewable product, and hence the urgent need for large-production of lactic acid. However, the degree of optical purity of lactic acid affects certain important physical characteristics of PLA, such as thermostability. Ishida *et al* (2006) metabolically engineered a strain of yeast capable of producing 122 g/l of 99.9% optically pure L-lactic acid by deleting the coding region of pyruvate decarboxylase and then inserted six copies of the bovine L-lactate dehydrogenase genes into the genome under the control of pyruvate decarboxylase promoter of a wild type yeast.

1,3-Propanediol is a monomer with the potential for use in manufacture of polyester fibres, polyurethanes and cyclic compounds (Chotani et al., 2000). It finds an important application in polymers prepared from 1,3-propanediol and terephthalic acid, which has an estimated market value of one to two billion pounds per year in 10 years (Nakamura and Whited, 2003). 1,3-Propanediol is produced naturally from glycerol by fermentation (Zeng and Biebl, 2002), albeit anaerobically in microorganisms such as *Cirobacter*, *Clostridium*, *Enterobacter*, *Klebsiella* and *Lactobacillus species* (Nakamura and Whited, 2003). The three currently known chemical production of 1,3-Propanediol are capital intensive and also a large amount of pollutants are generated by these processes. Biological efforts towards the economy required of the competitive market for 1,3-propanediol is in the direction of building a single biocatalyst capable of utilizing the lower cost feedstock D-glucose. Even though D-glucose as a feedstock is cheaper than glycerol, the demanding stoichiometric requirements of D-glucose represent a drawback as it requires co-reactants for redox balance than the

use of glycerol. However, the combination of lower cost and yield favour D-glucose as a feedstock over glycerol for the production of 1,3-Propanediol. *K. pneumoniae* (Menzel and Zeng, 1997), *Citrobacter freundii* (Boenigk et al., 1993) and *Clostridium butyricum* (Biebl, 1991) have been used in the biotechnological production of 1,3-propanediol from glycerol; *K. pneumoniae* or *C. butyricum* yielded 56 g/l of 1,3-propanediol from glycerol (Biebl et al., 1992) in batch or fed-batch fermentations. (Chotani et al., 2000) constructed a strain of *E. coli* expressing yeast genes for production of glycerol from glucose and also expressing genes from *K. pneumoniae* for the production of 1,3-Propanediol from glycerol; this strain produced 1,3-Propanediol at levels equal or higher than any of the glycerol to 1,3-propanediol natural organisms.

1.4.1.6 Sweeteners

Sugar alcohols are commonly referred to sweeteners; xylitol, mannitol, and sorbitol are examples reviewed here. The sugar's carbonyl group in sugar alcohols is reduced to the corresponding primary or secondary hydroxyl group and hence they are classed as polyols (Akinterinwa et al., 2008).

Sugar alcohols are widely used in the manufacture of pharmaceuticals, personal care products and animal nutrition (Silveira and Jonas, 2002), and in addition they serve as intermediates in the chemical synthesis of various other products. Metabolic engineering strategies now exist for improving microbial production of sugar alcohols such as xylitol, mannitol, and sorbitol. Annual production of xylitol is between 20,000 and 40,000 ton per year (Granstrom et al., 2007), with a world market estimated to be $340 million at price of $4-5 per kilogram (Kadam et al., 2008) and as such, the microbial production and metabolic engineering of xylitol has recently attracted the most attention of all other sugar alcohols. Xylitol is produced as an intermediate during metabolism of D-xylose in yeasts, the natural producers of xylitol. A study reported *Candida tropicali* capable of producing 12 g/l of xylitol from xylose with glucose as cosubstrate (Kwon et al., 2006), and also xylitol was produced by *Pichia stipitis* mutants, with distruptions in xylitol dehydrogenase (XDH) or D-xylulokinase

(Jin et al., 2005).

Mannitol has a variety of clinical applications and it is also used as a sweetener. The industrial production of mannitol involves a variety of organisms such as bacteria, plants and *Candida magnoliae* (Lee et al., 2003), and currently *lactic acid bacteria* (LAB), *E. coli*, *Bacillus megaterium*, *S. cerevisiae* and *C. glutamicum* are being developed as efficient biocatalysts for production of mannitol (Akinterinwa et al., 2008). Wisselink *et al.* (2004) were able to increase the production of mannitol in *L. lactis*. The study reported the development of an effective mannitol producing *L. lactis* as a result of relieving the bottleneck dephosphorylation of mannitol-phosphate and addition of foreign gene encoding a dephosphorylase-activity in *L. lactis* based on predictions from mathematical modelling of *L. lactis* glycolysis (Wisselink et al., 2004).

1.4.1.7 Nutriceuticals

Nutriceuticals refer to a wide variety of foods or food components, believed to have medical or health benefit (Pszczola, 1992). The products have a range of beneficial actions, from supply of essential mineral or vitamins to protection against diseases.

Lactococcus lactis is a lactic acid producing bacterium employed in the dairy industry for the production of fermented milk products. Hence it a perfect target for the production of nutritional products (Pool et al., 2006) and has since been demonstrated to be the ideal cell factory for the production of nutriceuticals.

Food products such as the dairy products (cheese, butter, butter milk and yoghurt), fermented meat, plants and fruits (such as sausages, silage, sauerkraut, olives and grapes) (Caplice and Fitzgerald, 1999) result from the bacterial acidification, leading to longer shelf-lives (Ross et al., 2002). Lactic acid bacteria (LAB) are almost always the bacteria involved in the fermentation of these products, and the acidifying bacteria (starter culture) may also contribute to flavour, the texture and the nutritional value of the fermented product (Hugenholtz, 2008). Successful metabolic engineering

include the production of the butter aroma compound, diactyl, by redirecting the normal primary pathway in L. lactis towards the production of alpha-acetolactate (precursor of diacetyl), from the usual route starting from lactose or glucose to lactic acid (Hugenholtz, 2008).

Vitamin C (or L-ascorbic acid) is produced naturally by most eukaryotic organisms. It is a water-soluble powerful antioxidant and acts as a scavenger of reactive oxygen species (ROS). There is a wide range of applications for vitamin C, ranging from food, animal feed and beverages to pharmaceutical and cosmetic uses. The world's yearly production of vitamin C was estimated to be between 60 000 to 70 000 metric tons (Chotani et al., 2000), and about 50% of vitamin C synthesis is based on the chemical synthesis. Most of the existing biotechnological approaches involve 2-keto-L-gulonic acid (KLG) as the intermediate (Saito et al., 1997) which is convertible to ascorbic acid by the conventional chemical process (2-KLG, (Chotani et al., 2000). Yeasts lack the natural ability to synthesise L-ascorbic acid, but accumulate low levels of erythro-ascorbic acid (Huh et al., 1998). The first report on biosynthesis of vitamin C was based on a yeast engineered with both plant and animal, capable of covertion D-glucose to L-ascorbic acid (Branduardi et al., 2007), which was further improved by engineering with only the complete plant pathway (Fossati et al., 2010).

Table 1.1: Primary metabolites produced by biotechnological methods

Metabolite	**Organism**	**Carbon Source**	**Strategy**	**Yield**	**Reference**
L-tryptophan	*C. glutamicum*	Glucose	Gene expression and increased availability of PEP and E4P	50g/L	(Katsumata and Ikeda, 1993; Ikeda *et al*, 1994)
L-phenylalanine	*E. coli*	Glucose	Multiple gene manipulations	50 g/L	Backman *et al*, 1990
L-valine	*E. coli*	Glucose	Transcriptome analysis, Genome-scale *in silico* modelling, Multiple gene manipulations	0.378 g/g of glucose	Park *et al*, 2007
Threonine	*E. coli*	Glucose	transcriptome profiling, Genome-scale *in silico* modelling, Multiple gene manipulations	82.4 g/L	Lee *et al*, 2007
Malate	*E. coli*	Glucose	Metabolic flux analysis (genome-scale metabolic model of *E. coli*), Introduction of heterologous reactions, Multiple gene manipulations	9.25 g/L	Moon *et al*, 2008
Ethanol	*S. cerevisiae*	L-arabinose	Introduction of heterologous metabolic pathways, and Evolutionary engineering	0.43 g/g of L-arabinose	Wisselink *et al*, 2007
Isopropanonol	*E. coli*	Glucose	Introduction of heterologous metabolic pathways	4.9 g/L	Hanai *et al*, 2007
Glycerol	*S. cerevisiae*	Glucose	Multiple gene manipulations	0.469 g/g of glucose	Cordier *et al*, 2007
Sialic acid	*C. jejuni*	Lactose	Introduction of heterologous metabolic pathways	25 g/L	Fierfort and Samain, 2008
L-lactic	*S. cerevisiae*	-	Introduction of heterologous reactions	122 g/L	Ishida *et al*,2009
Xylitol	*Candida tropicali*	xylose (glucose: cosubstrate)	Multiple gene manipulations	12 g/L/h	Kwon *et al*, 2006
Succinic acid	*E. coli*	Glucose	Genome-scale *in silico* modelling, comparative genome analysis combined with in silico metabolic analysis	More than 7-fold yield increase	Lee *et al*, 2005
L-lactic acid	*E. coli*	Glucose	Genome-scale *in silico* modelling, Adaptive evolution, Multiple gene manipulations	0.87 to 1.75 g/L	Fong *et al*, 2005

Table 1.2: Secondary metabolites produced by biotechnological methods

Metabolite	**Organism**	**Carbon Source**	**Strategy**	**Yield**	**Reference**
flavonones	*E. coli*	phenylpropanoic acid	Introduction of heterologous metabolic pathway	700 mg/L	Leonard *et al* (2008)
lycopene	*E. coli*	Glucose	systematic (model-based) and combinatorial (transposon-based) methods to identify gene knockout targets, Genome-scale *in silico* modelling, Multiple gene manipulations	18 mg/g DCW	Alper *et al*, 2005a; Alper *et al*, 2005b
Beta-carotene	*E. coli*	Glycerol	Multiple gene manipulations	390 mg/L and 240 mg/L of b-carotene in 50 L and 300 L fermenter respectively	Kim *et al*, 2006
polhydroxyalkanoates (PHA) and novel PHAs	*E. coli*	-	Metabolic flux analysis (MFA), Bioinformatics and Proteome analysis,Multiple gene manipulations	Improved yield	Park and Lee, 2005
Vanillin	*E. coli*	ferulic acid	Introduction of heterologous reactions	16.1 g/L	Barghini *et al*, 2007
Ectoine	*E. coli*	Glucose	Introduction of heterologous reactions	4.6 mg/L of ectoine Rajan *et al*, 2008	
Penicillin	*Penicillium chrysogenum*	-	-	70 g/L	Olano2008
carotenoids	*E. coli*	Glucose	Elementary flux mode analysis,Multiple gene manipulations	Increased yield	Urean2009

Chapter 2

Identification and selection of biotechnological products and host strains

2.1 Introduction

An extensive review of the literature covering the production of commercially valuable products and materials, especially those produced through biotechnological, was carried out at the beginning of this PhD research. This effort was required to understand relevant information for choosing interesting biotechnological products from the large variety of commodity products and materials. While many value products and materials are already being produced biotechnologically, most of them are still based on petroleum resources. The literature also revealed the dynamics in the paradigm shift from the use of random mutagenesis and rational metabolic engineering to the most recent approach that combines rational metabolic engineering with systems biology and Synthetic Biology.

The literature review in section 1.4 (Production of value chemicals and materials) of this thesis contains all products reviewed. Products of interest were chosen based on

commercial need of the product, its market value, current scientific knowledge about the potential biocatalyst host and possibilities for achieving the goals of the research in an academic environment and short time scale for research work. Bioethanol, fumaric acid, lysine, glutamate and trehalose were considered the most interesting products.

2.2 Specific products of interest

2.2.1 Ethanol

Bioethanol is an interesting product based on the following reasons:

1. There is world-wide demand for a replacement for fossil-fuel due to high cost of oil and supply instability, and also because of increasing environmental concerns.

2. Metabolic engineering, as a result of inputs from advances from genomics and technology, has emerged as the leading tool for deriving renewable energy, especially from lignocellulosic fractions of plant biomass.

3. Construction of novel pathways in production strains based on lignoceric feedstocks offer possibilities for *in silico* modelling approach to enhance ethanol production in yeast and other microbial strains.

2.2.2 Trehalose

Trehalose is an interesting product based on the following reasons:

a. Trehalose is an expensive carbohydrate, costing about £1000 per kg.

b. *S. cerevisiae* has a pathway to synthesize trehalose quite efficiently.

c. There has been little work done in on *in silico* modelling to enhance production of trehalose in yeast.

2.2.3 Amino acids

Amino acids have extensive industrial applications based on their nutritional value, taste, physiological activities and chemical characteristics (Kramer, 2005). Further applications include use as animal feed supplements, for cosmetics, in the pharmaceutical industry, and as building blocks for chemical synthesis. About 40% of total amino acids produced are used for food, majorly L-glutamate, L-aspartate, L-phenylalanine, glycine and L-cysteine. The industrial production of amino acids has been mostly carried out by fermentation through the use of bacteria. However, the yeast *S. cerevisiae* is generally regarded as a safe microorganism, judging from its widespread use in the food industry, and this coupled with its high nutritional value in terms of protein and vitamin, explains its increasing use as a supplement and as flavour enhancer (Farfan and Calderon, 2000).

Only a few cases and with limited success stories of overproduction of L-tryptophan (Prasad et al., 1987) involved *Hansenula* or *Candida* yeasts. Studies on gene arrangement and regulation of *Trp* biosynthesis revealed particular features in *S. cerevisiae*, which are largely different from those of *E. coli*. *TRP* genes in *E. coli* are completely clustered in an operon (Yanofsky, 1981), whereas in *S. cerevisiae* they are scattered over the chromosomes. Hence, yeast and bacteria have different gene regulation in the biosynthesis of L-tryptophan. Thus yeast, and in particular *S. cerevisiae*, otherwise well characterized in fermentation processes, has hardly been used for L-Tryptophan production (Prasad et al., 1987). However, *S. cerevisiae* strains overproducing threonine have HOM3 allele encoding a feedback-resistant *aspartate kinase*.

The commercial production of amino acids in yeast are important due to the following reasons:

1. Amino acids are among the major products in biotechnology in both volume and

value, and the global market is growing. Amino acids have extensive industrial applications and also as animal feed supplement, for cosmetics and building blocks of chemical synthesis.

2. Wild-type yeast strains are poor in essential amino acids such as lysine, threonine and methionine. Overproducing these amino acids in yeast strains might be of great interest to the food industry (Farfan and Calderon, 1999) because of the GRAS (generally regarded as safe) status of yeast. Moreover, phenylalanine, glutamate, methionine and lysine are commercially important.

3. There is potential for multiple product opportunities (amino acids) in yeast from one pathway, providing means to reduced technical and commercial development costs.

4. Opportunities exist for enhancing production of various amino acids in yeast through computational methods, especially pathway analysis using elementary flux modes (EFMs), flux balance analysis (FBA) and other systems biology tools.

2.2.4 Fumaric acid

The U.S Department of Energy listed fumaric acid as one of the top 12 most interesting chemical building blocks derivable from biomass (Werpy and Petersen, 2004). The other specific reasons why fumaric acid is an interesting product based as follows:

1. Fumaric acid is very valuable in the food industry.

2. Genetic manipulation has hardly ever been explored for fumaric acid production (Roa Engel *et al*, 2008).

3. Opportunities exist for enhancing production of fumaric acid in yeast through computational methods, especially pathway analysis using elementary flux modes (EFMs), flux balance analysis (FBA) and other systems biology tools.

2.2.5 Xylitol

Xylitol is an interesting product because:

1. Xylitol is expensive and has a large market.

2. Opportunities exist for overproducing xylitol in yeast through in silico methods and other systems biology tools.

2.3 Choice of a production microbial host

Much research work have been carried out in both academic laboratories and the industry on bacteria and yeast as hosts for biocatalysts. The development of these whole-cell biocatalysts has been successful because they combine the advantages of unicellular organisms in terms of rapid growth and ease of genetic manipulation. Yeast emerged as the production host of choice because of the advantages it possesses over the other microbial hosts, especially for safety reasons in the production of products, such as amino acids, for human consumption. Importantly, budding yeast was already used in my host laboratory, the Manchester Centre for Integrative Systems Biology, where I could benefit from the know-how and facilities. Overall *S. cerevisiae* had the most benefits from this study.

Chapter 3

Materials and Methods for computational modelling

3.1 Introduction

3.2 Metabolic pathway analysis

3.2.1 Identification of pathway reactions and modelling strategies

The metabolic network used for most of this study was constructed from the literature (Dobson et al., 2010; Cakir et al., 2004) and from pathway databases (SGD and KEGG). It includes the reactions of the central metabolism of *S. cerevisiae*, consisting of 51 reactions of glycolysis, the pentose phosphate pathway, the TCA cycle, the glyoxylate shunt and oxidative phosphorylation; 85 reactions of the biosynthetic pathways of 17 amino acids and biomass; and 27 transport reactions. The number of reactions and metabolites included in the metabolic network are 163 and 172 respectively (see Appendix A).

The biosynthetic pathways of threonine, glycine and serine have been reported to have

less than the direct flux value of 2 (relative to a glucose uptake of 100 arbitrary units) based on ^{13}C flux measurements of the respiro-fermentative metabolism of *S. cerevisiae* in batch cultures (Gombert et al., 2001) and have therefore been excluded from this network. In any reaction where isoenzymes are involved, only one reaction is represented, while metabolites and cofactors (NAD, NADH, NADP, and NADPH) were compartmentalised. A simplified biomass reaction considering 5 central metabolic intermediates (2-oxoglutarate, phosphoenolpyruvate, 3-phosphoglycerate, pyruvate and oxaloacetate) is included in the reaction network to account for flux directed to biomass production. Redox requirements for the synthesis of biological precursor molecules and the energy requirements for polymerizing the monomers into macromolecules have not been considered. However, an energy requirement for maintenance is accounted for in the model.

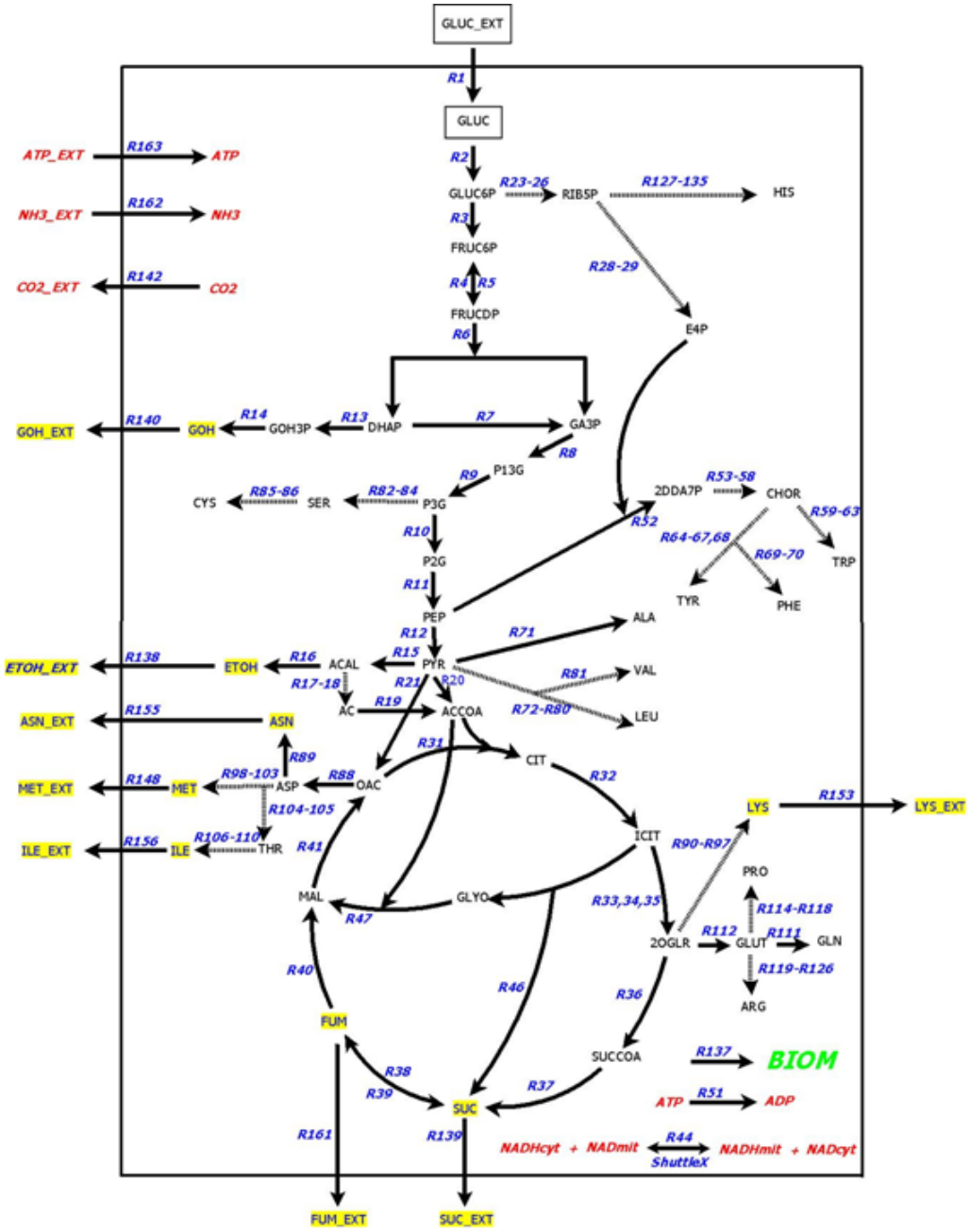

Figure 3.1: **Metabolic network of yeast**. Metabolic network for yeast grown on glucose depicting the reactions of central metabolism, the biosynthetic pathways of 17 amino acids, and reactions of biomass formation and energy interconversion. Input and output metabolites defined as "external" are appended with EXT. Solid arrows indicate steps of single reactions, while dashed arrows indicate steps of several reactions. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A). Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A). aas = amino acids (explained in Appendix A)

Table 3.1: **Stoichiometric models for EFM analysis**. A summary of stoichiometric models used for EFM analysis in terms of reactions, fixed external metabolites and EFM results

	Model simulation		
Stoichiometric Model	**No of reactions**	**External metabolite**	**Application**
M1	163	GLUC, ETOH, SUC, ACETAL, AC, GOH, CO2 and 5 aas	Statistical analysis only
M2	163	GLUC, ETOH, SUC, ACETAL, AC, GOH, CO2 and 2 aas	Statistical analysis only
M3	163	GLUC, ETOH, SUC, ACETAL, AC, GOH, CO2 and 3 aas	Statistical analysis only
M4	163	GLUC, ETOH, SUC, ACETAL, AC, GOH, CO2 and 17 aas	Statistical analysis only
SM1	163	GLUC, TREH, ETOH, GOH, SUC, FUM, BIOM, CO2, NH3 and 12 aas	Statistical analysis only
SM2	163	GLUC, TREH , ETOH, GOH, SUC, FUM, BIOM, CO2, NH3, ATP and 4 aas	Statictical analysis and Metabolic engineering for lysine production
S1	163	GLUC, ETOH, GOH, SUC, FUM, BIOM, CO2, NH3, ATP and 10 aas	Statistical analysis
S2	163	GLUC, ETOH, GOH, SUC, FUM, BIOM, CO2, NH3, ATP and 17 aas	Statistical analysis and Metabolic engineering for ethanol and glutamate production
S3	163	GLUC, ETOH, GOH, SUC, FUM, BIOM, CO2, NH3, ATP and 14 aas	Statistical analysis only
Teusink	17	GLUC, ETOH, SUC, ACETAL, AC, GOH, CO2, TREH	Statistical analysis and Metabolic engineering for trehalose production
Cakir_TREH	69	GLUC, TREH,GLYC, ETOH, GOH, SUC, BIOM, CO2, ATP, FUM	Statistical analysis and Metabolic engineering for ethanol and fumaric acid production

A number of different stoichiometric models (M1, M2, M3, M4, SM1, SM2, S1, S2 and S3) were built for this study, based on the reaction network model (Figure 3.1). The

stoichiometric models are different from each other depending on the number of amino acids and other metabolites fixed (see Table 6.1) as "external metabolites". The fixed "external metabolites" are defined as (a) Inputs - Glucose, ammonia and ATP; or (b) Outputs: carbon sinks (ethanol, glycerol, succinate, acetaldehyde, fumaric acid, and carbon dioxide), biomass, and amino acids of interest. Glucose is the carbon source and ammonia was included in the inputs as a nitrogen source in the minimal medium. Other stoichiometric models used in the study are the reactions of the glycolysis model (Teusink et al., 2000) and the Cakir reaction network (Cakir et al., 2004) which was modified by adding trehalose pathway reactions (Cakir_Treh model). The stoichiometric models were used for statistical analysis, metabolic engineering or both.

3.2.2 EFM modelling and classification of EFMs

COPASI 4.5 (Hoops et al., 2006) was used to compute EFMs for *S.cerevisiae* grown on glucose, using the reactions of the stoichiometric models as input files. The COPASI output files contain the original EFM results (EFM_0) from the stoichiometric models and are referred to as EFM datasets M1, M2, S1, S2, S3, Teusink and Cakir_Treh.

3.2.2.1 Hierarchical and *k*-means clustering

Using the R statistical package (`http://www.r-project.org/`), k-means and hierarchical clustering analyses were performed on EFM data matrix based on with EFMs on the columns and reactions on the rows. The steps involved in this process are outlined in Figure 3.2. In this methodology, Mclust model clustering was used to determine the value of k (number of clusters) which was then used for either k-means or hierarchical clustering analysis of the EFM data. Mclust (a contributed package in R stastistical package) is a model-based approach which apply maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters. The value of k suggested by the analysis of within-group sum of squares is

used for comparison with the mclust predicted values.

k-means clustering has a disadvantage in that a prior knowledge of the appropriate number of clusters is required for the best cluster solution, and a further disadvantage is that since k-means start by randomly allocating observations into clusters, different runs may yield slightly different cluster solutions. Hence, k-means clustering of EFM data were based on the value of k predicted by mclust models (available in R statistical package) and within-group sum of squares. In addition, the following procedures were carried out to validate all clustering analysis results:

(a) Visual inspection:

Clusters generated by k-means clustering were visually inspected for similarity of EFM members in each partition. That is, the group members are checked for any meaningful patterns that represent the distinguishing characteristics in each group. Validity of a cluster solution is indicated when clusters show sensible and expected results.

(b) Reclustering:

By reclustering the data with the same value of k, it is possible to identify a consistent k-means solution. Hence. the k-means clustering was repeated 10 times at a specific value of k predicted by mclust model clustering method and analysis of within-group sum of squares and the most consistent set of clusters were chosen.

(c) Biological relevance:

Biological relevance of the EFMs partitioned into each clusters were verified by looking for EFMs with similar biochemical routes from substrate to a specific external metabolite partitioned into similar clusters. Clusters were compared with each other to find out where modes are characterised by biochemical routes leading to different external metabolites.

R statistical package was also used to investigate the best combination of metrics and distance methods for hierarchical clustering of elementary flux mode data. Using

the hierarchical clustering of the Teusink data matrix (Table 6.1), the metrics tested are Spearman, Pearson's correlation, and Euclidean and distance methods tested are Ward, Single, Complete and Average. As a result of these prior investigations,further hierarchical clustering analyses on the data matrices from models M1, M2, M3 and M4 were carried out on "Euclidean" metrics and "Complete linkage" method. In addition, for the hierarchical clustering of EFM data of these 4 models, dendrograms were "cut" to the value of k predicted by mclust results.

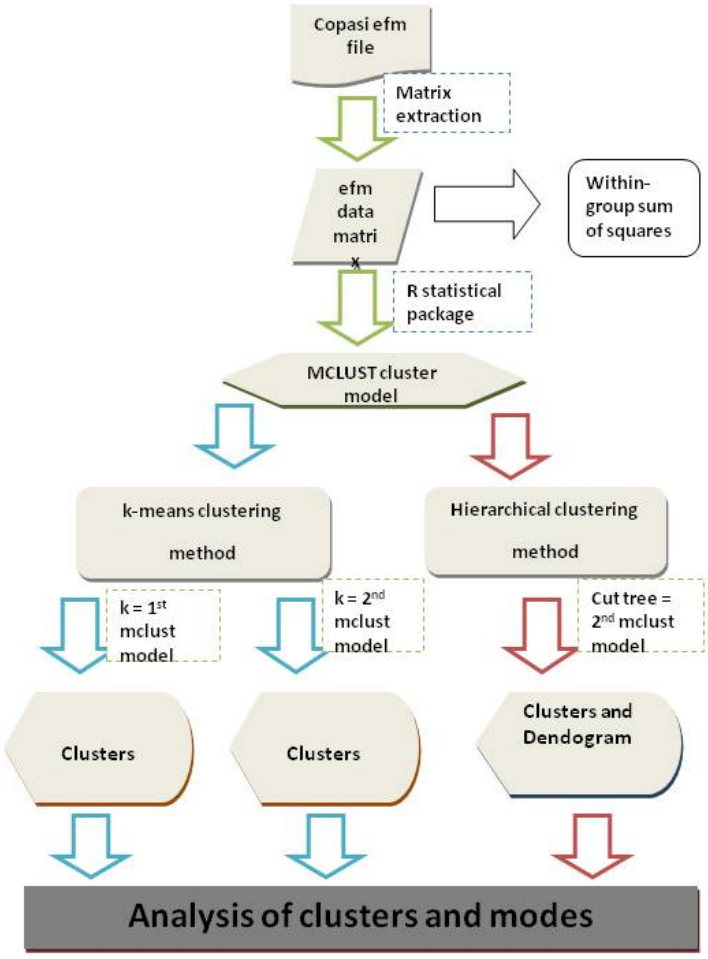

Figure 3.2: **Kmeans and hierarchical clustering of EFM**. Figure depicts a schematic outline of steps involved in the clustering analysis of the EFM data matrix. Mclust model clustering is used to determine the value of k (number of clusters) which was then used for either k-means or hierarchical clustering. The value of k suggested by the analysis of within-group sum of squares is used for comparison with the mclust predicted values.

3.2.2.2 EFM complexity reduction and PAMK clustering

A new methodology (Figure 3.9) was developed to address the shortcomings of the previous approach (see section 3.2.2.1 and Figure 3.2). As the number of EFMs were quite large and difficult to interpret for biotechnological purposes, reduction of complexity in the EFM data was therefore a very important issue to be tackled in this project. This issue was not addressed in the first clustering approach (section 3.2.2.1). Hence, as a focal point in this study, computational extraction and clustering methods were combined to help reduce the complexity in EFM data. EFM subsets obtained were expected to reveal characteristic patterns that may help in quickly locating the most useful modes and reactions (enzymes) for biotechnological purposes in terms of routes from the substrate (feedstock) to products of interest.

Figure 3.3 depicts the computational processing part of the new methodology, that is how the computational data extraction steps (using Java programs) steps interfaced with the clustering analysis step. In addition, it shows that the integration of computational analyses of EFM clusters allows for the determination of mean molar yields of target metabolite in each cluster, and for comparisons of clustered EFMs in terms of metabolites.

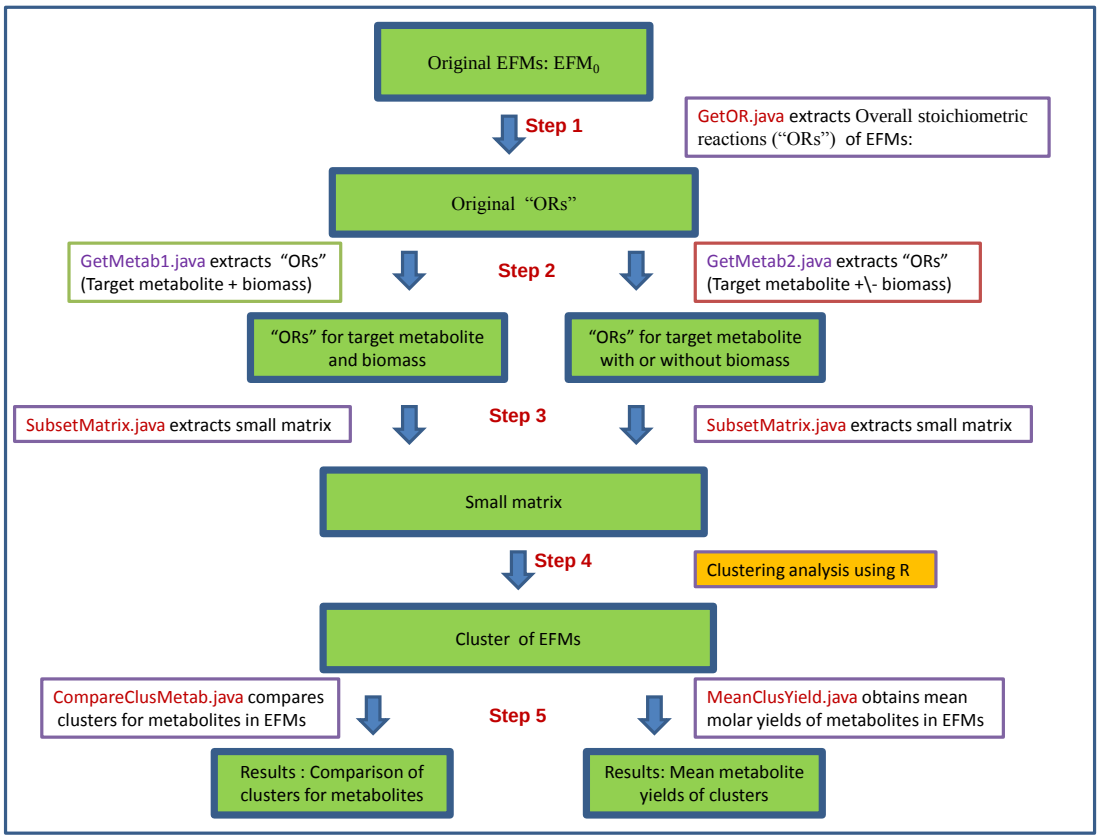

Figure 3.3: **Computational processing of EFM data.** A schematic representation of the entire pipeline for computational processing of EFM data showing the how the computational data extraction (using Java programs highlighted in red and purple) steps are interfaced with clustering analysis. In a systematic and stepwise manner, original EFMs are parsed for overall stoichiometric reaction or "ORs", from which can be extracted two types of subsets of "ORs", one type "ORs" are from EFMs containing target metabolite and biomass and the other type of "ORs" are from EFMs containing target metabolite with or without biomass. Small matrix is then obtained from either type of small subsets of "ORs" for clustering using R statistical package. Clusters of EFMs are then analysed computationally for the mean molar yields of target metabolite in each cluster, and for comparisons of clustered EFMs in terms of metabolites

I wrote 5 Java programs (GetOR.java, GetMetab1.java, GetMetab2.java, CompareClusMetab.java and MeanClusYield.java, all in Appendix B) for the implementation of the computational part of the methodology.

In the first step of the computational extraction methodology, a Java program (GetOR.java, Appendix B) was used to convert the EFMs in the original COPASI EFM output (EFM_0) into their stoichiometrically balanced overall equations ("ORs") and the associated reactions for each EFM. The "ORs" obtained at this stage are the original "ORs". The output file, Figure 6.1 (as an example), shows both the "ORs" and the

associated reactions for each EFM. The algorithm involved in the "ORs" extraction by GetOR.java is explained with the pseudocode in Figure 3.4.

```
Step1:
Programs looks in COPASI EFM output file (EFM dataset)
Step2:
for each EFM, program {
  finds and removes the internal metabolites
  stores stoichiometrically balanced external metabolite equations.
  stores reaction names (including coefficients)
  }
end
Output file of original ''ORs'' saved.
```

Figure 3.4: Pseudocode for GetOR.java

The next step involves the extraction of 2 different subsets of "ORs" from the original "ORs", depending on the output choices. The original "ORs" output file from first step was used as the input file for GetMetab1.java and GetMetab2.java programs (Appendix B); GetMetab1.java extracts only "ORs" corresponding to EFMs starting from external glucose and produces the target metabolite and biomass, while GetMetab2.java extracts "ORs" that are from EFMs starting from external glucose and produces target metabolite, either with or without biomass (that is these EFMs may or may not produce biomass). Figure 3.5 is a pseudocode explaining the how the two programs carry out the extraction of the subsets of "ORs".

```
Step1:
Programs looks in original ''ORs'' output file (from original EFM dataset)
Step2:
foreach line (''OR'') in the input file {
    Get metabolites on the left and right side of equations
    if search terms are found in the left and right
    then
        save current mode to file
    }
endforeach
Output file of ''ORs'' subset saved.
```

Figure 3.5: Pseudocode for GetMetab1.java and GetMetab2.java

The "ORs" subset output file from GetMetab1.java and GetMetab2.java Java programs was used as input for SubsetMatrix.java to extract the coefficients of the reactions to form a data matrix of rows of EFMs and columns of reactions according to the algorithm in Figure 3.6.

```
Step1:
Programs looks in the ''overall reaction'' output file
Step2:
program stores all reactions and assigns each reaction to a column
for each row of EFM {
stores the coefficient values of the reactions to a matching position in the column
Where a particular reaction does not appear in a mode, a value of ''0'' is recorded
end}
Step 2 is repeated for all reactions in every EFM to form a matrix
Output file of small matrix saved.
```

Figure 3.6: Pseudocode for SubsetMatrix.java

Hence, the element in ith row and jth column of the matrix is the coefficient of reaction j for all rows of EFM. Program writes out a matrix data for all EFMs into an output file used later in clustering analysis (see section 3.2.2.1).

The output matrix was used for clustering analysis in R statistical package. After clustering analysis, further computational processing was carried out using CompareClusMetab.java and MeanClusYield.java programs. CompareClusMetab.java was used to compare the clusters for metabolites in the EFMs. Figure 3.7 is a pseudocode explaining the algorithm of CompareClusMetab.java program.

```
Step 1:
Program accepts EFM file and PAMK clustering output file as inputs
Step 2:
foreach <cluster -> mode> in \$clusterModes {
   \$currentClusterAllMetabs = get all metabs from all modes from this cluster
   get common and not common metabs among the cluster modes
   if \$metabolite appears in all modes from current cluster
      then
         add \$metabolite to \$currentClusterCommonMetabs
      else
      add \$metabolite to \$currentClusterOtherMetabs
   endif
write \$currentClusterCommonMetabs to output file
write \$currentClusterOtherMetabs to output file
      }
endforeach
Results output file saved.
```

Figure 3.7: Pseudocode for CompareClusMetab.java

MeanClusYield.java program was used to compare clusters for their mean cluster yields of metabolites. Molar yield is the ratio of the coefficient of a metabolite to the coefficient of glucose substrate. The pseudocode in Figure 3.8 explains the algorithm of MeanClusYield.java program.

```
Step 1:
Program reads cluster file (for extracting the cluster modes)
Step 2:
Program reads ''ORs'' file
Search for external glucose (GLUC\_ext) and other metabolites
 for each metabolite
    for each cluster
       search it's modes in the calculated ModeYield tree
          calculate sum of yields
          calculate and output mean
 end
 Results output file saved
```

Figure 3.8: Pseudocode for MeanClusYield.java

Figure 3.9 presents graphically the computational and clustering methodology employed for the reduction and classification of the EFM data, which enabled *in silico* gene deletion phenotype analysis.

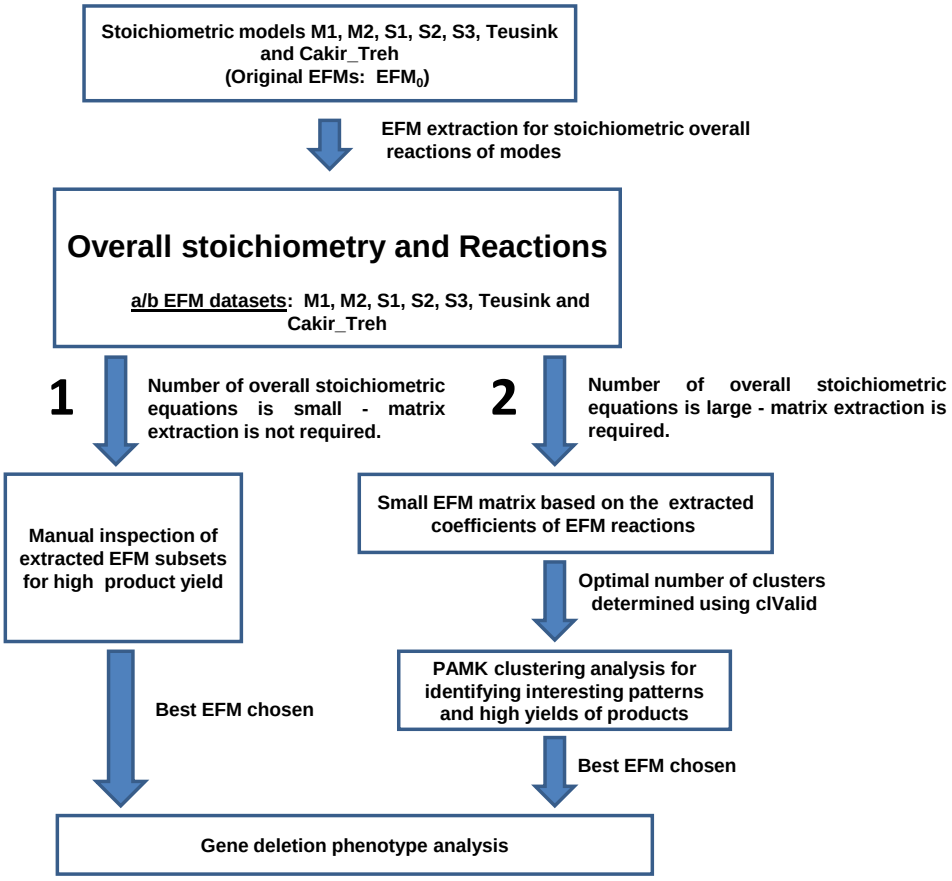

Figure 3.9: **An improved EFM clustering methodology using PAMK.** Schematic overview of steps involved in the computational extraction and clustering analysis of EFM data, leading to gene deletion phenotype analysis. "Route 1" is suitable for small EFM data and it consists of only computational extraction of the overall stoichiometry of EFMs, while "Route 2" is applicable to large EFM data and it combines the computational extraction of overall stoichiometry of EFMs with clustering analysis of EFMs. EFM datasets S1a and S2a contain the overall stoichiometry of EFMs (and EFM reactions) starting from glucose and producing a metabolite of interest and also biomass. EFM dataset S1b contains the overall stoichiometry of EFMs (and EFM reactions) starting from glucose and producing a metabolite of interest, whether or not they also produce biomass.

In the first step of this methodology, COPASI was used to compute EFMs for stoichiometric models SM1, SM2, SM3, S1, S2, Teusink and Cakir_treh models. The output files obtained from the computation (EFM No 6 example in the top part of Figure 6.1), containing the original EFMs EFM_0, was used as the input of a Java script for the extraction of "ORs" of all EFMs (EFM No 6 example in the bottom part of Figure 6.1). Two types of "ORs" files obtained are:

(1) SM1, SM2, SM3, S1a, S2, Teusink and Cakir_treh

(2) S1b

The "ORs" files for SM1, SM2, SM3, S1a, S2, Teusink and Cakir_treh consisted of all overall stoichiometric equations of EFMs starting from glucose and finishing with production of a metabolite of choice (and biomass) while EFM data set S1b comprised of overall stoichiometric equations of EFMs starting from glucose and producing a metabolite of interest (the EFMs may or may not produce biomass). These two different datasets were created for analysis showing the different effects of reducing the EFM data by only one variable (metabolite of interest) and by two variables (metabolite of choice and biomass).

The next step is split into two alternative routes: if the number of "ORs" (corresponding to number of EFMs) is small, manual inspection of extracted EFMs is performed for the identification of the best ones (Route 1, Figure 2), otherwise the "ORs" file was processed by the matrix extraction step (Route 2, Figure 6.1) involving the extraction of coefficients of all reactions into file forming an EFM matrix. The next step along "Route 2" (Figure 2) is clustering analysis for the identification of the best EFM for in silico gene deletion phenotype analysis as described in section 3.2.4.1.

3.2.2.3 Cluster sizes and validation of results

For this study, seven different clustering algorithms (UPGMA or agglomerative hierarchical, kmeans, PAM, Diana, Clara, Fanny and Model based clustering) implemented in the R statistical package, clValid (Team, 2009) were tested. A matrix of EFMs (rows) and reactions (column) was derived from "ORs" ("Route 2", Figure 3.9) and used for clustering analysis. Using clValid, clustering analyses were performed on the EFM data matrix. In order to find out the best clustering algorithm and validating metrics for EFM data, matrices from "ORs" of S1a and S1b were used for the purpose of comparisons between the seven clustering methods (UPGMA or agglomerative hierarchical, kmeans, PAM, Diana, Clara, Fanny and Model based clustering).

The comparison study was based on the optimal number of clusters returned by the internal validating measures (Connectivity, Dunn's index and Silhouette width) available in clValid. The optimal number of clusters for the EFM datasets S1a and S1b were determined by evaluating 2 to 10 clusters using different clustering methods.

The PAM clustering analysis using Dunn's index as internal validation measure, proved to be the best clustering method for obtaining subsets of EFMs yielding biologically meaningful information. All subsequent clustering analyses involving EFM datasets S1, S2, S3 and S4, were carried out using PAM clustering method, Dunn's index and the optimal number of clusters were determined by evaluating 2 to 10 clusters.

3.2.3 Pattern analysis using regular expression

Inspection of the overall stoichiometry of EFMs ("ORs") show that the "ORs" contain patterns which may be used to reduce the dimensionality of EFMs and also help to classify EFMs into classes that permit their use for *in silico* gene deletion phenotype analysis. Pattern analysis based on regular expression was implemented in Matlab. A regular expression (regex or regexp) provides a flexible means of matching text strings to patterns of characters. I wrote a matlab program, ProcessEFM.m (Appendix B), to carry out the pattern analysis of "ORs". The implementation of this pattern matching algorithm using ProcessEFM.m involves the user to specify the substrate name and the name of the product of choice. In this way, the program provides a flexible EFM data classification for different substrate and products of interest. The output files are of two types: the first one contains the different classes of EFMs matching the class patterns and the molar yields for each substrate of choice, and the second file contains the redundant EFMs not matching any of the class patterns.

Figure 3.10 shows the steps for the implementation of pattern analysis of EFMs. The EFM overall stoichiometric ("ORs" file) is used as an input file for the Matlab program. The Matlab program then matches the EFM class pattern supplied by the

user to the "ORs" in the input file. The output file from the program contain the matching classes of EFMs.

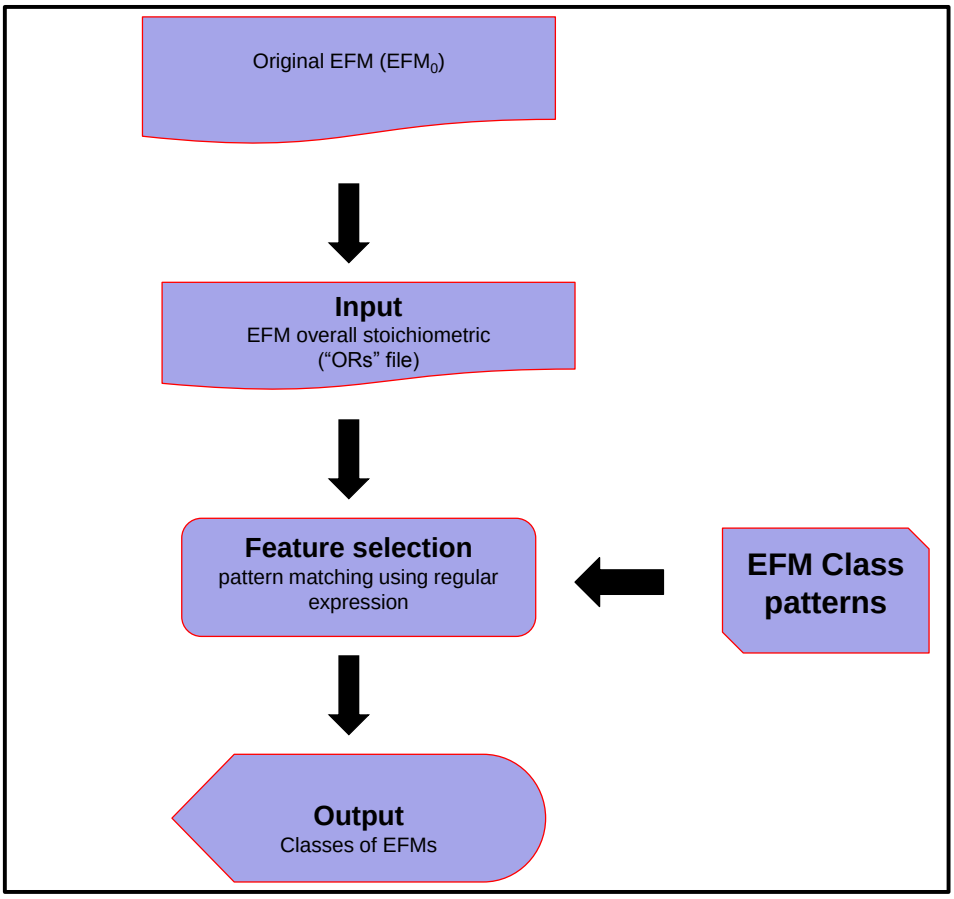

Figure 3.10: **Pattern analysis of EFMs.** The Figure depicts the steps involved in the implementation of pattern analysis of EFMs. The EFM overall stoichiometric ("ORs" file) is used as an input file for the matlab program. The Matlab program then matches the EFM class pattern supplied by the user to the "ORs" in the input file. The output file from the program contain the matching classes of EFMs.

In this study, EFM analysis has been used to simulate the growth of yeast in minimal synthetic medium with glucose as the substrate, and NH_3 is included in the reactions to simulate the presence of ammonia in the medium. In order to obtain useful classes of EFMs from metabolic engineering purposes, classes of patterns have been designed based on different scenarios characterising the different EFMs as follows:

1. Pattern No1: Substrate = BIOM_EXT + Product

This is a class of EFMs requiring only the main substrate and producing biomass

and metabolite of interest.

2. Pattern No2: Substrate + NH_3_ext = BIOM_EXT + Product

This is a class of EFMs requiring main substrate and externally supplied ammonia and producing biomass and metabolite of interest.

3. Pattern No3: Substrate + NH_3_ext = BIOM_EXT + Product + anything else

This is a class of EFMs requiring main substrate and externally supplied ammonia and producing biomass, metabolite of interest and any other metabolite as a side product.

4. Pattern No4: Substrate + NH_3_ext + anything else = BIOM_EXT + Product + anything else

This is a class of EFMs requiring any other metabolite in addition to the main substrate and externally supplied ammonia and producing biomass, metabolite of interest and any other metabolite.

5. Pattern No5: Substrate + NH_3_ext + anything else = BIOM_EXT + Product

This is a class of EFMs requiring any other metabolite in addition to the main substrate and externally supplied ammonia and producing biomass and metabolite of interest.

The output classes (1 - 5) indicate ranking of classes of EFMs to help make important decisions for metabolic engineering purposes. Class 1 EFMs are likely to give molar yields nearest to the theoretical yields of products, followed by EFMs in class 2, and so on, until class 5.

3.2.4 In silico gene deletion phenotype analysis

3.2.4.1 The general concepts

EFMs can be considered as a minimal set of enzymes necessary for the production of specific metabolites that operate at steady state, and represent all the capabilities of a metabolic network, that is, all the phenotypes that can be expressed in the organism. Hence, EFM analysis permits the design of *in silico* phenotype gene deletion studies.

In order to facilitate *in silico* gene deletion phenotype analysis, EFM clusters obtained in section 3.2.2.2 were analysed for cluster mean yields for target metabolites and the EFMs in the different clusters were compared based on their metabolite members. The two analyses of cluster solutions for biological interpretation were based on two stages as follows:

1. Computation of cluster mean molar yields:

   The "cluster mean molar yields" of metabolite of interest were determined for each cluster, in order to find out the cluster in which the EFMs with highest molar yields of the target metabolite are located.

2. Comparisons of different EFMs cluster solutions:

   Clusters were then compared with each other on the basis of different types of EFM members in each cluster. The inspection of the stoichiometric equations of the EFMs provide useful information such as utilisation of additional substrate(s) in addition to the main substrate and the by-products associated with EFMs.

3. Further analysis to choose the best EFM for biotechnological purposes

   In this step, the best EFMs for biotechnological purposes were chosen by applying a number of criteria as discussed below.

Using the results from the analysis of cluster solutions above, potential gene target knockouts were identified in four stages as follows:

1. The best EFM was chosen based on the following criteria:

   (a) EFM with the highest yield of product is selected if the conditions in (b) and (c) are met.

   (b) The EFM must start from the substrate and also produce the product of choice *and* biomass

   (c) A good candidate EFM must require either only glucose or the addition of inorganic compounds such as ammonium sulphate

   This process allowed the removal of cycles, incomplete EFMs (any EFM not of a full length: glucose to target product), EFMs requiring the addition of organic substrate (such as amino acids) as a second substrate and non-biomass producing EFMs.

2. All reactions (genes) not found in the EFM chosen in stage 1 were considered as potential target gene knockouts and hence compiled for *in silico* deletion analysis.

3. The compiled reactions were then carried through in silico gene knockout simulations based on the iterative steps as follows:

   (a) Deletion of a single reaction from the reaction network reaction network

   (b) EFM analysis in COPASI performed and the number of remaining EFMs recorded.

   (c) Further deletion (step a) and EFM analysis (step b) are repeated on the reduced reaction network until the number of recorded EFMs does not change.

   For multiple deletions (double, triple, etc), steps (a), (b) and (c) were repeated iteratively in COPASI for the required combination set of reactions.

4. Ranking of the reactions (genes) used for knockout simulations according to their effectiveness in lowering the number of EFMs lead to identification of

the best target gene knockouts for improving the yield of product of choice. For multiple knockouts, a small number of multiple deletions were chosen to minimise labour and time costs for their construction in the laboratory.

3.2.4.2 Gene Deletion Phenotype Analysis for lysine

EFM set SM2 was carried forward to the *in silico* gene deletion phenotype analysis stage without clustering since this data subset contained only 2 modes and hence not suitable for clustering (Route 1, Figure 3.9). The best mode with the highest yield for lysine and requiring only ammonia in addition to glucose was considered for *in silic*o gene deletion phenotype analysis involving the deletion of reactions (enzymes) which are not found in the mode of choice but are present in all the other modes contained in the original EFM set (EFM_0) of model SM2.

The set of reactions marked for *in silico* deletion was simulated for effects of gene deletion using COPASI EFM analysis by iterative sequential deletion of reactions followed by the determination of the number of EFMs left intact (as described in section 3.2.4.1). The reactions with the largest appreciable EFM reductions were kept as the target 'single genes knockout' for lysine production in yeast. The reactions of the target 'single genes knockout' were then combined with each other to derive 'double-mutant' reaction sets. The 'double-mutant' reaction (and genes) sets that are of known *in vivo* lethality were removed. Each of the remaining 'double-mutant' reaction sets were then simulated for lysine production in the same way described for the single reactions. Similarly, the reactions of the target 'double knockout genes' were then combined with the reactions of the target 'single knockout genes' and each triple set of reactions were simulated for improved yield of lysine in yeast (as carried out for single and double knockouts).

3.3 Modelling using FBA

3.3.1 Formulating and solving an FBA problem

FBA involves the optimisation of a linear objective function with linear equality and/or inequality. An objective function is a function that one desires to be maximised or minimised. Constraints were used to define the limits of allowable values in the solution space. In this study, FBA problem was formulated as follows:

Maximise z = $\mathbf{c}.\mathbf{v}$

Such that $\mathbf{S}.\mathbf{v} = 0$

$LB \leq \mathrm{v} \leq UB$

Given an FBA model, linear programming algorithm gives a set of fluxes that maximises $\mathbf{cv}$ while satisfying the bounds (LB = lower bound, UB = upper bound) and stoichiometric constraints ($\mathbf{Sv}$). Thermodynamic reversibility and mechanistic constraints are enforced by the bounds. The FBA problem is solved as a unique maximum flux through the objective function. In this work, the objective function was either biomass yield or/and yield of any other metabolite.

3.3.2 Steps in FBA

This section describes the FBA methods for strain development towards overproduction of various metabolites in yeast using the yeast genome scale reconstructions. I developed a four-step methodology for developing *in silico* production strains (Figure 3.11).

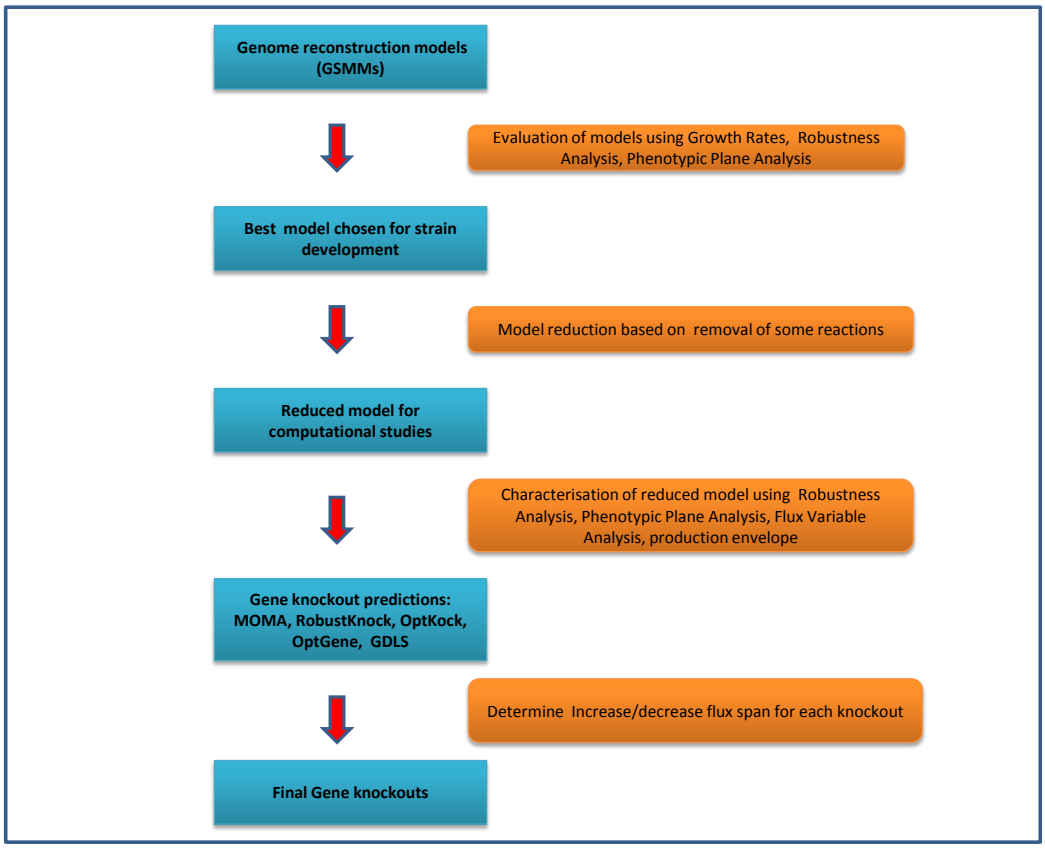

Figure 3.11: **Schematic overview of steps FBA steps for strain development.** First, Genome scale models were evaluated and the best was chosen for further studies. Next, the model chosen was reduced and characterised for production of target products. Reduced model was used for knockout predictions, and the knockout sets were further evaluated to obtain a final set of knockouts.

3.3.2.1 Models and FBA optimisation

Several yeast genome scale reconstructions are currently available, and it was decided to evaluate yeast genome scale reconstructions iND750 (Duarte et al., 2004), iMM904 (Mo et al., 2009) and Yeast 4.38 (Dobson et al., 2010) in order to determine the most suitable model for the purpose of optimal strain development. The models, iND750 and iMM904, have already been functionally validated under different substrate and genetic environments to be predictive of growth rates, metabolite excretion and gene essentiality.

3.3.2.2 Evaluation of models

The best model was chosen based on the characterisation of the networks of the genome scale models by carrying out "growth rate and product secretion optimisation", "robustness analysis", "Phenotypic plane analysis" and "flux variability analysis'. The 3 models were tested under 2 different substrates (glucose and xylose) and 2 environmental conditions (aerobic and anaerobic). COBRA Toolbox 2 software (Schellenberger et al., 2011) in Matlab environment was used for carrying out the "growth rate and product secretion optimisation", "robustness analysis", "Phenotypic plane analysis" and "flux variability analysis". The import and export of a metabolite was simulated as a negative and positive flux respectively. The unit of uptake rate, mmol gD$W^{-1}h^{-1}$, was used for all FBA simulations in this study.

In the growth rate and product secretion analysis, COBRA Toolbox was used to simulate the microaerobic fermentation, oxidative-fermentative growth, and anaerobic growth using iND750 and iMM904 models for assessing yeast growth rates and secretory capabilities for ethanol, fumaric acid, trehalose, lysine and glutamate under specific substrate and environmental conditions. Table 3.2 summarises the environmental conditions simulated for *S. cerevisiae* growth. Simulations involved the addition of ergosterol and zymosterol in order to investigate the metabolic capabilities of *S. cerevisiae* in anaerobic fermentative condition because *S. cerevisiae* in anaerobic conditions require ergosterol and zymosterol for maintenance of cell growth.

Table 3.2: **Environmental conditions simulated for yeast *S. cerevisiae*.** Table summaries the environmental conditions simulated for *S. cerevisiae* growth. LB = lower bound for simulating the uptake of a nutrient from growth medium.

Condition	**Media**	**Bounds for nutrient intake (mmol gDW^{-1} hr^{-1})**
Aerobic	Glucose and oxygen	Glucose: LB = -18.5 Oxygen: LB = -18.5
Microaerobic fermentation	Glucose and oxygen	Glucose: LB = -18.5 Oxygen: LB = -1.5
Oxidative fermentation	Glucose and oxygen	Glucose: LB = -14 Oxygen: LB = -10
Anaerobic	Glucose, oxygen, ammonia, ergosterol, zymosterol, ATPM	Glucose: LB = -14, Oxygen: LB = -0.01, Ammonia: LB = -5, Ergosterol: LB = -10, Zymosterol: LB = -10, ATPM: LB = -20

Robustness analysis was performed by using FBA to analyse the network properties of iMM904 model. The sensitivity of biomass production (objective) to glucose uptake rate and oxygen uptake rate was investigated. In order to determine the effect of varying glucose uptake rate on growth, the values of upper and lower bounds of glucose exchange reaction was set to between 0 and 20 mmol gD$W^{-1}h^{-1}$ while setting a fixed oxygen uptake rate of 20 mmol gD$W^{-1}h^{-1}$. The effect of varying oxygen on growth was performed by setting the values of upper and lower bounds of oxygen exchange reaction to between 0 and 20 mmol gD$W^{-1}h^{-1}$ while glucose uptake rate was fixed at 20 mmol gD$W^{-1}h^{-1}$.

Further analysis of the network properties of iMM904 involved the phenotypic phase plane analysis. In contrast to robustness analysis, which involves the calculation of the network state based on varying one parameter, the results of varying two parameters simultaneously can be plotted as a phenotypic phase plane. In this study, the *S. cerevisiae* phenotypic phase plane was formulated to display the maximum growth rates for different combinations of glucose and oxygen uptake rates. COBRA

Toolbox was used to perform FBA for maximisation of growth for each combination of glucose and oxygen uptake rates, and for plotting two-dimensional and three-dimensional phenotypic phase planes of the results. The glucose uptake rate was varied between 0 and 20 mmol gD$W^{-1}h^{-1}$ and oxygen uptake rate was varied from 0 to 20 mmol gD$W^{-1}h^{-1}$. In order to determine the sensitivity of the FBA solutions, COBRA Toolbox was also used to calculate shadow prices (Edwards et al., 2002) and reduced costs (Varma et al., 1993; Ramakrishna et al., 2001). Shadow prices are the derivative of the objective function with respect to the exchange flux of a metabolite; that is the sensitivity of the objective function to the addition of each metabolite.

Maximise z = $\mathbf{c}.\mathbf{v}$

Such that $\mathbf{S}.\mathbf{v}$ = b

$LB \leq \mathbf{v} \leq UB$

Shadow price (i) = $\frac{dz}{db_i}$ where b_i represents ith metabolite

On the other hand, reduced costs are the derivatives of the objective function with respect to an internal reaction with 0 flux; indicates the degree of change in objective value if a particular reaction with zero flux is given a non-zero value.

Maximise z = $\mathbf{c}.\mathbf{v}$

Such that $\mathbf{S}.\mathbf{v}$ = b

$LB \leq \mathbf{v} \geq UB$

Reduced cost (j) = $\frac{dz}{dv_j}$ where v_j represents jth reaction

Flux Variability Analysis:

The flux distribution calculated by FBA is generally not unique. In many cases it is possible for the biological system to achieve the same objective by using the alternative pathways, meaning that phenotypically alternate optimal solutions are possible. Flux variability analysis (FVA) (Mahadevan and Schilling, 2003) uses FBA to identify alternate optimal solutions. FVA calculates the minimum and maximum

allowable fluxes through a reaction using linear programming, with the objective function flux constrained close to or equal to optimal solution.

Maximize/minimise $\mathbf{v}_i$

Such that $\mathbf{S}.\mathbf{v} = 0$

$LB \leq \mathbf{v} \leq UB$

vobj=$\mathbf{v}$*obj

where $\mathbf{v}$*obj is the optimal value of the objective flux.

Using a COBRA Toolbox function, FVA for the *S. cerevisiae* model under aerobic growth conditions was performed at a minimum growth of 90% of optimal growth. FVA enabled the calculation of minimum and maximum fluxes in the *S. cerevisiae* model.

3.3.2.3 Pre-processing of models

Genome scale reconstructions of yeast iND750, iMM904 and Yeast 4.38 were pre-processed so as to obtain reduced models. The goal for the pre-processing of genome scale reconstructions by the removal of some reactions was to increase the chance for selection of valid gene deletion targets for optimised strains. All reactions with 'zero' minimum and maximum flux values and the dead end reactions were removed from the network. The other reactions removed include spontaneous, diffusion and those not associated with any gene. Only one reaction was allowed for every occurrence of isoenzymes.

3.3.2.4 FBA and strain design

Two COBRA Toolbox functions, OptKnock (Burgard et al., 2003) and GDLS (Lun et al., 2009) were used for the design of production strains based on the reduced iMM904 model (see section 3.3.2.3). In principle, these algorithms use FBA to maximize the flux through each of the exchange reactions in the reduced iMM904 model

for the targeted products. The OptKnock algorithm calculates the knockout reactions sets for maximising the production of a specific product by solving a biLevel MILP (mixed integer linear programming) problem using a Gurobi MILP solver (Gurobi optimisation, Houston, Texas, USA). Similarly, in the GDLS algorithm, a bilevel MILP (mixed integer linear programming) problem was solved using Gurobi solver for the identification of knockout reaction sets for the overproduction of target products. An example of settings for both OptKnock and GDLS (implemented in COBRA Toolbox 2.0) is shown in Figure 3.12.

```
model = readCbModel('iMM904_flux.xml');
model = changeRxnBounds(model, {'EX_o2(e)', 'EX_glc(e)'}, [-18.5, -10], 'l');
selectedRxns = {model.rxns{[490, 446, 695, 691, 693, 379, 380, 848, 415, 696, 698,...
265, 768, 771, 115, 407,697, 465, 789, 839, 828, 840, 200, 55, 516, 518, 517, 90,...
822, 391, 394, 584, 588, 519, 584, 225, 241, 803, 802, 748, 184, 124, 123, 738, 524,...
857, 183, 734, 733, 862, 51, 556, 232, 554, 550, 662, 694, 753, 754, 209, 207, 683,...
143, 134, 484, 483, 492, 36, 34, 791, 792, 140, 131, 503, 504, 87, 505, 835, 834, 47,...
556, 232, 525, 438, 449, 442, 404, 405, 672, 46, 84, 56, 666, 656, 129, 128, 747, 151,...
741, 739, 522, 506, 494, 493 ]}};
options.targetRxn = 'EX_lys(e)';
options.vMax = 1000;
options.numDel = 5;
options.numDelSense = 'L';
constrOpt.rxnList = {'biomass_SC5_notrace', 'ATPM'};
constrOpt.values = [0.05, 8.39];
constrOpt.sense = 'GE';
optKnockSol = OptKnock(model, selectedRxns, options, constrOpt);
[gdlsSolution, bilevelMILPproblem, gdlsSolutionStructs] = GDLS(model, 'EX_lys(e)',...
'minGrowth', 0.05,'selectedRxns', selectedRxns, 'maxKO', 5, 'nbhdsz', 3);
gdlsSolution.KOs
```

Figure 3.12: **Settings for Opknock and GDLS.** Figure shows an example of Opknock and GDLS settings (in Cobra Toolbox 2.0) used for the design of lysine producing *in silico* yeast strains.

Chapter 4

Materials and Methods for laboratory experiments

4.1 Chemicals and reagents

4.1.1 Yeast strains, media and growth conditions

4.1.1.1 *S. cerevisiae* mutant strains for lysine production

Single, double and triple mutants of *S. cerevisiae* for increased production of lysine were either sourced or constructed. Seven single mutant strains of *S. cerevisiae* for increased production of lysine(Δ*alt2* (ALT2 mutant), Δ *kgd1* (KGD1 mutant), Δ*kgd2* (KGD2 mutant), Δ*lsc1* (LSC1 mutant), Δ*lsc2* (LSC2 mutant) and Δ*glt1* (GLT1 mutant) and a control strain (CS, metabolism unrelated HO gene knockout strain) used in this study were obtained from EUROSCARF (Institute of Molecular Biosciences, Wolfgang Goethe-University Frankfurt, Frankfurt, Germany) and are listed in Table 4.1. Lysine accumulating control yeast mutant strains, *lys80* (12T7c Δ*lys80*) and 02940c (lys20fbr and lys21fbr yeast mutants) were kindly provided by (Evelyne Dubois, Institut de Recherches du CERIA, Belgium: (Feller et al., 1999)). Table 4.2 lists the seven S. cerevisiae double mutants, (Δ*alt2*Δ*kgd2*, Δ*alt2*Δ*kgd1*,

Δ*alt2*Δ*lsc2*, Δ*alt2*Δ*lsc1*, Δ*alt2*Δ*glt1*, Δ*kgd1*Δ*kgd2*, Δ*lsc1*Δ*lsc2*) and one control, YLR123CC (a random KAN and NAT double mutant), obtained from the Boone Lab for this study. The five *S. cerevisiae* double mutants (Δ*kgd1*Δ*alt1*, Δ*kgd2*Δ*alt1*, Δ*lsc1*Δ*alt1*, Δ*lsc2*Δ*alt1* and Δ*alt1*Δ*glt1*) constructed by me are listed in Table 4.3.

Table 4.1: **EUROSCARF single mutant and control strains.** *S. cerevisiae* single mutant and control strains obtained from EUROSCARF.

Single Mutant Standard Name	Systematic Name	Background	Genotype
Δalt1	YLR089C	BY4741	Mata; his3Δ1;leu2Δ0;met15Δ0;ura3Δ0;YLR089c::kanMX4
Δalt2	YDR111C	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YDR111c::kanMX4
Δkgd1	YIL125W	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YIL125w::kanMX4
Δkgd2	YDR148C	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YDR148c::kanMX4
Δlsc1	YOR142W	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YOR142w::kanMX4
Δlsc2	YGR244C	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YGR244c::kanMX4
Δglt	YDL171C	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YDL171c::kanMX4
CS	YDL227C	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0;YDL227c::kanMX4

Table 4.2: **Toronto single mutant and control strains.** *S. cerevisiae* double mutant and control strains obtained from the Boone Laboratory (Toronto).

Strain	Genotype
YLR123C	-
Δalt2Δkgd2	YDR111CˆKanR YDR148CˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆLYS2+
Δalt2Δkgd1	YDR111CˆKanR YIL125WˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆSTE3pr-LEU2 LYS2+
Δalt2Δlsc2	YDR111CˆKanR YGR244CˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆLYS2+
Δalt2Δlsc1	YDR111CˆKanR YOR142WˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆLYS2+
Δalt2Δglt1	YDR111CˆKanR YDL171CˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆSTE3pr-LEU2 LYS2+
Δkgd1Δkgd2	YIL125WˆKanR YDR148CˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆLYS2+
Δlsc1Δlsc2	YOR142WˆKanR YGR244CˆNatR can1ˆ::STE2pr-Sp_his5lyp1ˆLYS2+

Table 4.3: **In-house *S. cerevisiae* double mutants.** *S. cerevisiae* double mutants constructed in-house. Double deletion strains have BY4741 background.

Double Mutant	Mutant Name	Background	Genotype
Δkgd1Δalt1	DMOO1	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0;ura3Δ0;YIL125w::kanMX4;YLR089c::KlURA3
Δkgd2Δalt1	DMOO2	BY4741	Mata; his3Δ1;leu2Δ0;met15Δ0;ura3Δ0;YLR089c::kanMX4;YDR148c::KlURA3
Δlsc1Δalt1	DMOO3	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0;ura3Δ0;YOR142w::kanMX4;YLR089c::KlURA3
Δlsc2Δalt1	DMOO4	BY4741	Mata; his3Δ1;leu2Δ0;met15Δ0;ura3Δ0;YLR089c::kanMX4;YGR244c::KlURA3
Δalt1Δglt	DMOO5	BY4741	Mata; his3Δ1; leu2Δ0; met15Δ0;ura3Δ0;YDL171c::kanMX4;YLR089c::KlURA3

4.1.1.2 Synthetic minimal (SD) medium and selective media plates

Reagents were autoclaved for 20 minutes at 121 °C and 15 psi or filtered through a 0.22 µm filter disc (Anachem) for sterilisation.

S. cerevisiae strains were routinely grown in synthetic minimal medium (SD) supplemented with various additions of uracil and amino acids (Table 4.4). The SD medium composed of 0.17% w/v Yeast Nitrogen Base (YNB), 2% w/v glucose and 0.5% w/v ammonium sulphate. Minimal medium plates for uracil (URA) or leucine (LEU) selection were prepared by adding 2% agar and supplements (excluding either uracil or leucine respectively) to SD medium. The complete medium, YPD medium, used in this study composed of 1% w/v yeast extract, 2% w/v bacto-peptone, 2% w/v dextrose, and YPD plates were prepared by solidifying YPD medium with 2% w/v agar.

Table 4.4: **Supplements for SD medium.** Uracil and amino acid supplements for SD medium. URA = uracil, LEU = L-Leucine, HIS = L-Histidine, MET = L-Methionine, ALL = URA + HIS + MET + LEU.

Components of uracil and amino acids				
Supplement	Constituent	Stock concentration (g/100 ml)	volume of stock for 1 Litre of medium (ml)	Final concentration (mg/l)
NC	URA	0.2	10	20
	LEU	1	3	30
	HIS	1	2	20
	MET	1	2	20
2 X LEU	Two times the concentration of LEU and the same concentrations of URA, MET and HIS in NC above			
3 X LEU	Three times the concentration of LEU and the same concentrations of URA, MET and HIS in NC above			
2 X ALL	Two times the concentration of URA, MET, HIS and LEU in NC above			
3 X ALL	Three times the concentration of URA, MET, HIS and LEU in NC above			
NC_PH6	Same concentrations of URA, MET, HIS and LEU in NC above. SD medium also adjusted to PH6.0.			

4.1.2 Oligonucleotide primers

Gene disruption oligonucleotide primers (see Table 4.5) were either sourced or designed based on a PCR gene deletion strategy of (Baudin et al., 1993) to generate

disruption cassette consisting of URA3 selectable marker gene amplified from plasmid pBS1539_kl that was introduced into yeast cells. Confirmation primers (sourced and designed) for the verification of successful gene disruption by URA3 and LEU2 markers are also listed in Table 4.7. Oligonucleotide primers were designed using "Primo Optimum 3.6 Optimal Gene Synthesis And Expression" software in the package BioToolKit 320 (`http://www.changbioscience.com/download/biotoolkit.html`). Confirmation of successful gene disruption by URA3 and LEU2 selectable markers were carried out according to Tables 4.8 and 4.9 respectively.

Table 4.5: **Primers for URA3 and LEU2 cassettes.** Primers for generating URA3 and LEU2 selectable disruption cassettes (5' - 3'). Primers used in generating disruption cassettes for deleting genes for the construction of *S. cerevisiae* mutant strains for production of lysine.

Primer Name	**Primer Sequence**
ALT1_Disrupt_F	GTTTCTGCTTCTCAATTGAACGCATATAAATATATTTCCCCAGTCTTTATTTTGCTCTCTCCACGA TGTCGGTCTGCATTGGATGGTGG
ALT1_Disrupt_R	GATCACATTATTATAATAAACTAGCTATTTAAATGTTTATTGAAGACTGTTCTGCCCCCTTTTATT CAGTTGCACCGTGCCAATGCAG
KGD2_Disrupt_F	ATAAACTTCACTACCACATTTGTTACAACCAAAGACACAACTTCAGATAATTATTTAAACAATGTC GGTCTGCATTGGATGGTGG
KGD2_Disrupt_R	CACAGTAATAGCGGACAAGAATAATCATGAAATCAGATTGGTATGGGCTGCAAATTTCAAATCAGT TGCACCGTGCCAATGCAG
LSC2_Disrupt_F	GATTAATAAGGATTGAGTCAATACAATCGAAAAAAATACTGAAGCATTGCAACTGAACAAAATGTC GGTCTGCATTGGATGGTGG
LSC2_Disrupt_R	CTTTATTAATAGTAAAAAAGCATATATACTTTATTATTAACTCTTTTGTTTTTCTCGAGAAGCTTA GTTGCACCGTGCCAATGCAG
ZWF1_disrupt_F	CCCCTCCTTCTCCCCCTTCCCCCTCTCCAATTGGCTGTATAGACAGAAAGAGTAAATCCAATAGAA TAGAAAACCACATAAGGCAAGATGACTTCTAGTATATCTACATACC
ZWF1_disrupt_R	AACAAAGAGAGTGAGCTTGCAAGATAAAATCACTCGAAAAAAAAAATTTCAGTGACTTAGCCGATA AATGAATGTGCTTGCATTTTTCTATCGACTACGTCGTTAAGGC

Table 4.7: **Confirmation primers for URA3 and LEU2 markers.** A list of primers used for Confirmation of successful gene disruption by URA3 and LEU2 selectable markers. Primers were used for 5 PCR confirmation experiments: 3 positive and 2 negative.

Primer Name	Primer Sequence (5' - 3')	Source
ALT1_A	TGAGACAACTTCACGTACTCTTCTG	YKOs collection
ALT1_D	TAGGTGCCATTGGTAAGAAGTAAAG	YKOs collection
KGD2_A	ACATTAGCACCATTCTACTACAGGG	YKOs collection
KGD2_D	ACATGTTAGGTCAATGGAAAGTCAT	YKOs collection
LSC2_A	CACCAAAGCCAGGTAGATACTAAAA	YKOs collection
LSC2_D	AAACGATAATATGTTCCTGAACTCG	YKOs collection
URA3_F_Cassette	TCATGCAAGTCCGGTTGCATCG	In-house generated
URA3_R_Cassette	CTCTTCCTCCCATATCGTTCTG	In-house generated
ALT1_B	CTGCTGTTCGTTTTGGTTTAATAGT	YKOs collection
ALT1_C	GAACATCCAGGTAAATTCGATAATG	YKOs collection
KGD2_B	TGACCTCAATATCAATTTTATCGGT	YKOs collection
KGD2_C	AATTAACCCTAGAAGATATGACGGG	YKOs collection
LSC2_B	TATATACAGCAGATACTGGCTTCCC	YKOs collection
LSC2_C	TGAAGGGTAACATTGGATGTTTAGT	YKOs collection
ZWF1_A	ATTATTAATGTGGGATTTTTGGCTC	YKOs collection
ZWF1_D	TCAATGATAAGTACAAGTCCAATCG	YKOs collection
ZWF1_B	CTTGAAGAACTGTTCGACCTTAGAG	YKOs collection
ZWF1_C	CGCTGTGTACCTAAAGTTTAATGCT	YKOs collection
ZWF1_LEU2_Fcassette	ACTTCTAGTATATCTACATACC	In-house generated
ZWF1_LEU2_Rcassette	TCGACTACGTCGTTAAGGC	In-house generated

Table 4.8: **Confirmation primer pairs for alt1, kgd2 and lsc2.** An example of different combinations of primer pairs for confirmation of gene disruption - alt1, kgd2 and lsc2 gene deletion. Primer set numbers 1 - 3 are for positive controls and primer set numbers 4 - 5 are for negative controls.

Combination of primer sets for deletion of alt1, kgd2 and lsc2 genes					
Primer set	KGD1xALT1	**KGD2XALT1**	**LSC1XALT1**	**LSC2XALT1**	**ALT1XGLT**
1	ALT1_A R_Cassette	KGD2_A R_Cassette	ALT1_A R_Cassette	LSC2_A R_Cassette	ALT1_A R_Cassette
2	ALT1_D F_Cassette	KGD2_D F_Cassette	ALT1_D F_Cassette	LSC2_D F_Cassette	ALT1_D F_Cassette
3	ALT1_A ALT1_D	KGD2_A KGD2_D	ALT1_A ALT1_D	LSC2_A LSC2_D	ALT1_A ALT1_D
4	ALT1_A ALT1_B	KGD2_A KGD2_B	ALT1_A ALT1_B	LSC2_A LSC2_B	ALT1_A ALT1_B
5	ALT1_C ALT1_D	KGD2_C KGD2_D	ALT1_C ALT1_D	LSC2_C LSC2_D	ALT1_C ALT1_D

Table 4.9: **Confirmation primer pairs for zwf1.** An example of different combinations of primer pairs for confirmation of gene disruption - zwf1 gene deletion. Primer set numbers 1 - 3 are for positive controls and primer set numbers 4 - 5 are for negative controls.

Primer set No	Δkgd2Δalt1Δzwf1	Δlsc2Δalt1Δzwf1	Δalt1ΔgltΔzwf1
1	ZWF1_A	ZWF1_A	ZWF1_A
	R_Cassette	R_Cassette	R_Cassette
2	ZWF1_D	ZWF1_D	ZWF1_D
	F_Cassette	F_Cassette	F_Cassette
3	ZWF1_A	ZWF1_A	ZWF1_A
	ZWF1_D	ZWF1_D	ZWF1_D
4	ZWF1_A	ZWF1_A	ZWF1_A
	ZWF1_B	ZWF1_B	ZWF1_B
5	ZWF1_C	ZWF1_C	ZWF1_C
	ZWF1_D	ZWF1_D	ZWF1_D

4.1.3 Plasmids pBS1539 and pREP41

URA3 and LEU2 selectable markers used as gene disruption cassettes in this study were amplified from Plasmids pBS1539 (Puig et al., 2001) and pPREP41 (Basi et al., 1993) respectively. plasmid pBS1539 (see Figure 4.1)harbours a URA3 marker from *Kluyveromyces lactis* and does not replicate in yeast. The plasmid pPREP41 is from S. Pombe and contains the *S. Cerevisiae* LEU2 gene for auxotrophic selection of transformants on media lacking leucine.

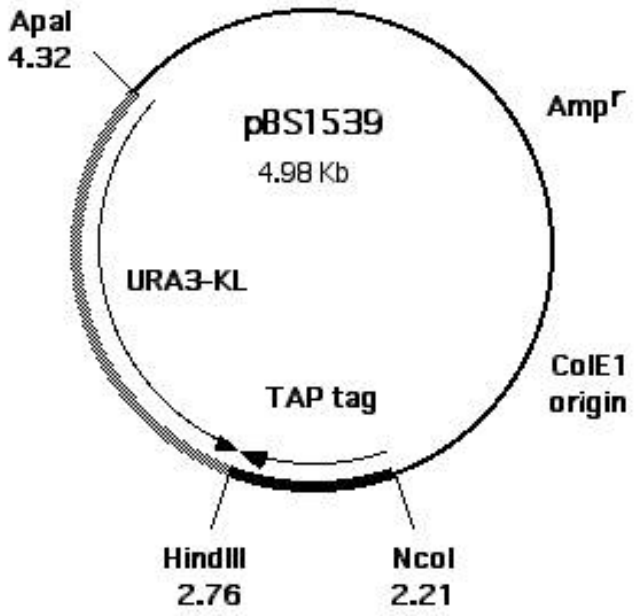

Figure 4.1: **Diagram of pBS1539.** Diagram of pBS1539 from Seraphim Lab (http://www.embl.de/ExternalInfo/seraphin/pBS1539.html)

4.1.4 Reagents for yeast competent cells and transformation

4.1.4.1 50% (w/v) polyethylene glycol (PEG)

100 ml of 50% (w/v) PEG was prepared by dissolving 50 g of PEG (MW 3350, Sigma) in 35 ml of Milli-pore water in a 150 ml glass beaker by magnetic stirring for about 30 minutes. The volume of the PEG solution was then adjusted to 100 ml in a 100 ml graduated cylinder, and then mixed by inversion. PEG solution was autoclaved in a securely capped bottle.

4.1.4.2 Salmon sperm DNA

2 mg/ml salmon sperm DNA was was prepared as follows: 200 mg of high molecular weight DNA (Deoxyribonucleic acid Sodium Salt Type III from Salmon Testes, Sigma D1626) were weighed into 100 ml of TE buffer (10 mM Tris-HCL pH 8.0, 1.0mM EDTA) in a 200ml glass beaker. DNA was dispersed into solution by drawing it up

and down repeatedly in a 10 ml pipette. DNA solution was then mixed vigorously on a magnetic stirrer for 3 hours. DNA solution was then aliquoted into 50 ml falcon and 1.5 ml Eppendorf tubes and stored in a $-20\,^{\circ}$C freezer.

4.1.4.3 10x Tris EDTA (TE) buffer PH 7.7

100 ml solution of 10X TE (PH 7.7) was prepared by mixing 10 ml of 1M Tris.Cl PH 8.0 [100mM], 2 ml of 0.5 M EDTA PH 8.0 [10mM] and 78 ml of milli-pore water in a 250 ml glass beaker.The PH of the solution was then adjusted to 7.7 using a PH meter (Sartorious PB-11).

4.1.4.4 Tris EDTA lithium acetate (TELiAc) PH 7.5

10 ml of 1 M lithium acetate (LiAc, PH 7.5, 100 mM) was prepared by mixing 1 ml of 10x TE pH 7.7 [10mM Tris, 1mM EDTA], 1 ml of 1 M LiAc pH 7.5 [100 mM] and 8 ml sterile milli-pore water in a 50 ml falcon tube.

4.1.4.5 PEG/TELiAc pH 7.5

5 ml solution of (PEG/TELiAc pH 7.5) was prepared mixing 0.5 ml of 10x TE pH 7.7 [10mM Tris, 1 mM EDTA], 0.5 ml of 1 M LiAc pH 7.5 [100 mM] and 4 ml of 50% PEG4000 [40%] in a 50 ml falcon tube.

4.2 Growth characterisation

20 ml SD medium (with NC supplement) (see Table 4.4) in 125 ml flask was inoculated freshly from a plate and incubated overnight at 30 °C with shaking at 200 rpm. The overnight culture was added to 60 ml fresh SD medium in 200 ml flask to give $OD_{600} = 0.2$ and the diluted culture was incubated at 30 °C with agitation at 200 rpm. For each *S. cerevisiae* mutant strain, culture was set up in duplicate, and OD measurements

carried out at different time points on JASCO V-630 spectrophotometer until 30 hrs of growth. Samples for biomass measurements were also taken at mid-log phase and early-stationary phase. Specific growth rate and doubling time for each mutant strain were also determined.

Calculations of growth rate were carried out at the steepest part of *S. cerevisiae* growth curves using the exponential function in Microsoft Excel software. The exponential curve gives the following function:

Number of cells at a specific time = a x $e^{(\mu t_0)}$

a = number of cells at time = 0 (t_0)

b = specific growth rate, μper hour

Doubling time, t_d is the amount of time it takes for a culture to double the number of cells. The calculation of doubling time was carried out by deriving the relationship between μ and t_d as follows:

$\frac{dx}{dt} = \mu x$ (where μ is the growth rate and x is the product)

$\frac{dx}{x} = \mu dt$

Integrate for x_t and x_0 (x_0 is the product at 0 time and x_t is product at a specific time after t(0)):

$[\ln x]_{x_0}^{x_t} = \mu t$

$\frac{x_t}{x_0} = e^{\mu t}$

$x_t = x_0 e^{\mu t}$

i.e., $OD_t = OD_0 e^{\mu t}$

For doubling time:

$$\ln OD_t = \mu t + \ln OD_0$$

$$\mu t = \frac{\ln OD_t}{\ln OD_0}$$

$$\mu t_d = \ln(2X\frac{\ln OD_t}{\ln OD_0}) = \ln(2)$$

$$t_d = \frac{\ln(2)}{\mu}$$

$$= \frac{0.6932}{\mu}$$

The best concentrations of amino acid supplements in SC minimal medium for optimal growth of single mutant strains were determined by calculating the doubling times and growth rates of the strains. This investigation was carried out by growing strains, CS and Δ*glt1*, for 30 h in 20 ml of SC medium under 6 different conditions based on different concentrations of supplemented amino acids (leucine, histidine and methionine) and uracil in the SD medium supplemented with either 2 X LEU, 3 X LEU, 2 X ALL, 3 X ALL or NC_PH6 (see Table 4.4)

4.3 Cultivation of *S. cerevisiae* for GC/MS

4.3.1 Culture medium and batch cultures

All *S. cerevisiae* mutants were grown on minimal medium (SD)supplemented with 3 X ALL (see section 4.1.1.2). Mutant strains were grown in triplicates and each replicate culture was assigned a number. Cultures were then randomised into different batches to reduce biological and analytical variations. Batches of cultures were started with 30 minute-interval between each batch, and the end points of batch cultures or transfer to the next stage of experiment were adjusted accordingly.

4.3.2 Steps for growing mutant strains

125 ml flasks containing 30 ml of culture medium were inoculated freshly from YPD cultural plates of *S. cerevisiae* mutant yeast strains and allowed to grow overnight at 30 °C with shaking at 200 rpm. Batch cultures of mutant strains were prepared by inoculating pre-warmed (30 °C) 60 ml of culture medium in 250 ml flasks to initial concentration of OD = 0.2 with washed inoculum from 30 ml starter culture and cultures allowed to grow until the exponential phase was reached. Next, the exponential phase cultures were were used to inoculate fresh batches of pre-warmed (30 °C) 60 ml culture medium in 250 ml flasks to initial concentration of OD = 0.2 and exometabolome and endometabolome samples were from the growing cultures at the start of log phase (7h of growth) and mid-log phase exometabolome (9h). All samples were stored in (−80 °C) freezer until analysed.

4.3.3 Determination of biomass

Eppendorfs containing pellets were weighed and then pellets were allowed to dry by leaving Eppendorfs open for 24 hours. Eppendorfs were weighed again each subsequent 24 hours until the weights of Eppendorfs remained unchanged from previous measurements. Then dry biomass was removed from each Eppendorf tube and the Eppendorf tubes were weighed in order to determine the weights of empty Eppendorf tubes. Dry weight of biomass for each culture was determined by subtracting the weight of each empty Eppendorf tube from the last and unchanged weight of the corresponding Eppendorf tube.

4.4 GC/MS quantitative metabolome measurements

This section describes the methods used for preparation of the samples for quantitative exometabolome (footprinting) and endometabolome analysis. Quantitative measurements were carried out on exometabolome and endometabolome samples (mid

log-phase) of *S. cerevisiae* mutant and control strain using GC/MS for lysine and other metabolites (glycerol, fumaric acid, glutamate, alpha-ketoglutarate and phenylalanine).

Rick Dunn (Manchester Centre for Integrative systems biology) provided methods for GC/MS sample preparation and carried out the GC/MS analysis.

4.4.1 Footprinting

For exometabolome analysis (footprinting), 1.0 ml of culture sample was syringe-filtered (0.22 μm) into a clean Eppendorf tube and the filtered sample was then stored at −80 °C.

4.4.2 Quenching of intra-cellular metabolism

Preparation of quenching solution: 2.5 l of quenching solution (methanol and water 60:40 ratio) was prepared by mixing 1.5 l of HPLC methanol with 1.0 l of HPLC grade water, and 40 ml aliquots of the quenching solution in 50 ml falcon tubes were stored in the −80 °C fridge.

For each mutant culture, triplicate samples were collected and quenched, each was prepared by adding 14 ml of culture sample into the centre of 40ml of quenching solution placed in dry-ice. Collection of triplicate samples ensured that the required minimum of 20 mg dry weight biomass for each mutant sample for GC/MS analysis. The quenched samples were then centrifuged at −9 °C to pellet biomass which was then kept at −80 °C, and 5 ml of supernatant were kept to check for metabolite leakage.

4.4.3 Extraction of intra-cellular metabolites

For each sample, 500 µl of HPLC grade methanol:water (80:20, −48 °C in dry ice) were added to cell pellets from the quenching step to vortex mix and solubilise the cells into the extraction solution. The mixture was then transferred into a 2 ml Eppendorf tube on dry ice and the tube was then placed in liquid nitrogen for 60 seconds. Next, the tube was allowed to thaw on dry ice. At this stage, the process of freezing in the liquid nitrogen followed by thawing on dry ice was repeated two more time before the tube was then centrifuged at 32,368 g for 10 minutes and supernatant collected into a clean Eppendorf. The entire process up to this stage was then repeated by washing the remaining of pellet from falcon tubes with fresh 500 µl of HPLC-grade methanol:water which was then transferred into the same Eppendorf tube used in the first process.

4.4.4 Quantitative GC/MS analysis

With the exception of pyridine (extra dry), methoxylamine hydrochloride and N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) which were obtained from Acros Organics (Loughborough, UK), all other materials were purchased from Sigma-Aldrich (Gillingham, UK) unless otherwise stated.

4.4.4.1 Derivatisation of metabolites for GC-ToF-MS analysis

The dried extracts were redissolved in 50 µl of 20 mg.ml^{-1} O-methoxylamine hydrochloride in pyridine, vortexed, and incubated at 60 °C for 30 minutes in a block heater. 50l of N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) was then added. and incubated at 60 °C for 30 minutes. On completion, 20 µl of retention index marker solution was added (0.2 mg.ml^{-1} docosane, nonadecane, decane, dodecane and pentadecane in pyridine) prior to centrifugation at 13,363 g for 15 minutes. The resulting supernatant (90 °C) was transferred to GC-MS vials for analysis.

4.4.4.2 GC-ToF-MS Analysis

Analyses were performed with a Gerstel MPS-2 autosampler (Gerstel, Baltimore, USA) coupled to an Agilent 6890N Gas Chromatograph with a split/splitless injector and Agilent LPD split-mode inlet liner (Agilent Technologies, Stockport, UK) which was interfaced to a LECO Pegasus III (4D) GCxGC-MS operated in GC-MS mode (Leco Corp., St. Joseph, MO). A 30m x 0.25 mm x 0.25 um VF17-MS bonded phase capillary column (Varian, Oxford, UK) was used at a constant helium carrier gas flow of 1 ml.ml^{-1}. A temperature programme was performed to provide elution of metabolites of differing volatility and polarity. The oven temperature was held at 70 °C for 4 minute followed by a temperature ramp of 20 °C /min to a temperature of 300 °C and finally the temperature was held for 4 min. 1 µl sample injections were performed. The injector was operated at 280 °C with a 4:1 split ratio, and a 25 ml.ml^{-1} gas saver flow switched on after 15 seconds. The transfer line was maintained at 240 °C. The mass spectrometer was operated at 70eV ionisation energy with a source temperature of 220 °C, acquiring m/z 45-600 at 20Hz.

4.4.4.3 Raw data processing

Raw data processing was performed using LECO's ChromaTOF v3.25 software (Leco Corp., St. Joseph, MO). Deconvolution was performed for each sample applying the following parameters; peak width of 1.8 s, S:N ratio of 10:1 and baseline of 1.0. A reference database was constructed containing relevant information for each metabolite of interest and included retention time, electron impact mass spectrum and single quantification ion. Chromatographic peaks in the reference database were searched for in each sample and if detected within specified ranges (retention time difference less than 10 seconds and mass spectral match score greater than 70%) the chromatographic peak area was calculated applying the quantification ion to define the peak limits. These data were exported for further data analysis.

4.5 Metabolite profiling

4.5.1 Sample preparation for GC/MS analysis

Six randomised replicates samples of *S. cerevisiae* double mutants and control strain were carried through metabolic profiling analysis using GC/MS. Cultivation of *S. cerevisiae* strains and sample preparation for GC/MS were carried out as described in section 4.3. However, only the endometabolome samples were collected at log phase of growth for metabolic profiling of *S. cerevisiae* strains.

The sample OD was employed to normalise for differences in the number of cells extracted so that the sample volume lyophilised (and therefore the sample solution analysed) was equivalent to the same number of cells for all samples. The lowest OD was 1.6 and the volume lyophilised for this OD was 800 μL. For example, for two samples with an OD of 1.6 and 2.0 the extraction solution volumes lyophilised were 800 and 640 μL, respectively. An internal standard (0.13 mg/ml; 100 μL) was mixed with each sample followed by sample lyophilisation for 18 hours.

4.5.1.1 Biological experiment

Metabolic profiling experiment of six randomised replicate samples of each of the five *S. cerevisiae* double mutants Δkgd1Δalt1 (M1), Δkgd2Δalt1 (M2), Δlsc1Δalt1 (M3), Δlsc2Δalt1 (M4) and Δalt1Δglt1 (M5) together with the control strain (CS) sample was carried out. Cultivations of *S. cerevisiae* strains and collection of log-phase endometabolome samples for GC/MS were carried out as described in section 4.3. However, only four mutant samples (M2, M3, M4 and M5) and CS were available for GC-MS analysis because mutant M1 showed very poor growth and so was eliminated from further investigations.

4.5.1.2 GC/MS analysis of samples

Metabolic profiling of intracellular extracts was performed with the analysis of all samples in a random order and with intermittent QC samples to allow appropriate quality assurance processes to be performed. All samples were chemically derivatised applying a two-stage process of heating the sample with a 20mg/ml O-methoxylamine in pyridine solution at 40 °C for 90 minutes followed by heating the resultant solution with MSTFA at 40 °C for 90 minutes. The samples were analysed using an Agilent 6890 GC and 7673 autosampler coupled to a LECO Pegasus III ToF mass spectrometer using the optimal settings previously determined for *S. cerevisiae* (O'Hagan et al., 2005).

The raw data were processed using LECO ChromaTof V2.12 and its associated chromatographic deconvolution algorithm with the following settings, baseline 1.0, data point averaging of 3 and average peak width of 1.8s. Data were exported to a single excel worksheet to construct a data matrix of chromatographic peak (with associated metabolite identification, retention time and quantification mass) *versus* sample and with the matrix infilled with chromatographic peak areas where a metabolite peak was detected.

4.6 Polymerase chain reaction (PCR) procedures

In this study, routine PCR procedure (below) was used for amplification of disruption cassettes from plasmids and PCR based gene deletion, while direct colony PCR was used for the verification of gene deletion using direct PCR from whole yeast cells.

4.6.1 Routine PCR procedure

PCR reactions were carried out in a total volume of 50 µl. 5 µl of 10X Taq buffer,1 µl of 10mM dNTPs, 1 µl of forward primer, 1 µl of reverse primer and 0.5 µl of template

DNA were pipetted into a 0.2 µl PCR tube. Sterile water was used to make the volume to 49.75 µl. The content of the tube was then thoroughly mixed by vortexing. 0.25 µl of Taq polymerase was added into the mixture and the tube was gently mixed by tapping. Master mixes of reagents were carried out routinely to accommodate for multiple-sample analyses.

Amplification of DNA samples were performed using a thermal cycler machine (Eppendorf MasterCycler Personal). The cycling profile of the PCR programme consisted of initial denaturation at 95 °C for 30 seconds (s), followed by 30 cycles of denaturing at 95 °C for 30 s, followed by annealing at 54 °C for 1 minute (min), and extending at 68 °C for 2 min, and then finally, completion of unfinished extension at 68 °C for 10 min.

Subsequent modifications of this protocol involved the use of 2 µl of yeast spheroblast instead of 0.5 µl of template DNA, and the volumes of the other reagents were readjusted accordingly.

4.6.2 Direct PCR from whole yeast cells

The direct PCR from whole yeast cells using zymolyase (lyticase) method (Ling et al., 1995) modified by Namjin Chung (`http://www.duke.edu/web/ceramide/protocols/0003.html\`) was used in this study.

An average-size yeast colony on a plate was touched with a sterile pipette tip. The pipette tip was rinsed with 10 µl Zymolyase solution by pipetting up and down three times. The cell suspension was then incubated at 37 °C for 5 minutes. Next, 2 µl of spheroplasted yeast cells were used instead of 1 µl of template DNA in a PCR reaction as in the "Routine PCR" protocol above, and the volumes of the other reagents were readjusted accordingly.

4.6.3 Agarose gel electrophoresis

Electrophoresis using a 1% agarose (Genetic analysis grade, Fisher Bioreagents, Fisher Scientific) gel in 300 mL of 1X TAE buffer (containing 40 mM Tris acetate PH 7.7 and 1 mM EDTA) and 10 µl of a 10mg/ml ethidium bromide (Sigma) was used to detect the DNA products generated by PCR amplification. A comb was placed at one end of molten agarose gel to create wells into which was loaded samples and gel was allowed to cast as a thin, rectangular slab. For an electrophoretic run, the gel was submerged in 1X TAE buffer contained in the Bio-RAD wide mini SUBTM cell tank. 10 µl of PCR products buffered in 1 µl of loading buffer was loaded into each gel well. Sizing of DNA bands was achieved by running a DNA marker along with the PCR products. 1 µl of GeneRulerTM 1 kb DNA ladder (Thermo Scientific Fermentas) buffered in 1 µl of 6X loading buffer (30% glycerol, 50mM EDTA, 0.25% bromophenol blue), and the buffered marker was mixed with 8 µl of sterile water and then the mixture was loaded into one of the wells. Electrophoresis was carried out at 100 volts for 45 minutes. Gels were then viewed on transilluminator and photographed.

4.7 PCR based gene deletion of *S. cerevisiae*

Primers were designed for PCR-generated disruption cassette consisting of selectable marker gene (URA3 or LEU2) amplified from from plasmid pBS1539_kl based on the PCR gene deletion based on the method of (Baudin et al., 1993). *S. cerevisiae* double mutants were constructed by disrupting the appropriate genes in yeast single mutants using a PCR-generated disruption cassette consisting of URA3 selectable marker gene amplified from from plasmid pBS1539_kl.

4.7.1 Primer design for *S. cerevisiae* mutants

This section describes the primer design strategies for constructing 5 *S. cerevisiae* double mutants based on EUROSCARF *S. cerevisiae* single mutants.

Table 4.10 lists the single mutants and the corresponding deletions required to construct *S. cerevisiae* double mutant strains.

Table 4.10: **Parents and single mutants for double mutants.** A list of single mutants and their corresponding deletions required for constructing*S. cerevisiae* double mutants for lysine production.

Construction of double mutants for lysine			
No	Double mutant required	Parent Single Mutant	Single deletion required
M1	Δkgd1Δalt1	Δkgd1	Δalt1
M2	Δkgd2Δalt1	Δalt1	Δkgd2
M3	Δlsc1Δalt1	Δlsc1	Δalt1
M4	Δlsc2Δalt1	Δalt1	Δlsc2
M5	Δalt1Δglt	Δglt	Δalt1

Table 4.11: **Parents and double mutants for double mutants.** A list of double mutants and their corresponding deletions required for constructing *S. cerevisiae* triple mutants for lysine production.

Construction of triple mutants for lysine			
No	Triple mutant required	Parent Double Mutant	Single deletion required
LT2	Δkgd2Δalt1Δzwf1	Δkgd2Δalt1	Δzwf1
LT4	Δlsc2Δalt1Δzwf1	Δlsc2Δalt1	Δzwf1
LT5	Δalt1ΔgltΔzwf1	Δalt1Δglt	Δzwf1

4.7.1.1 Description of primers and PCR strategy

The primer design strategy for PCR based deletion using a selectable marker(URA3 selectable marker as example) for the construction of *S. cerevisiae* double mutants is

depicted in Figure 4.2.

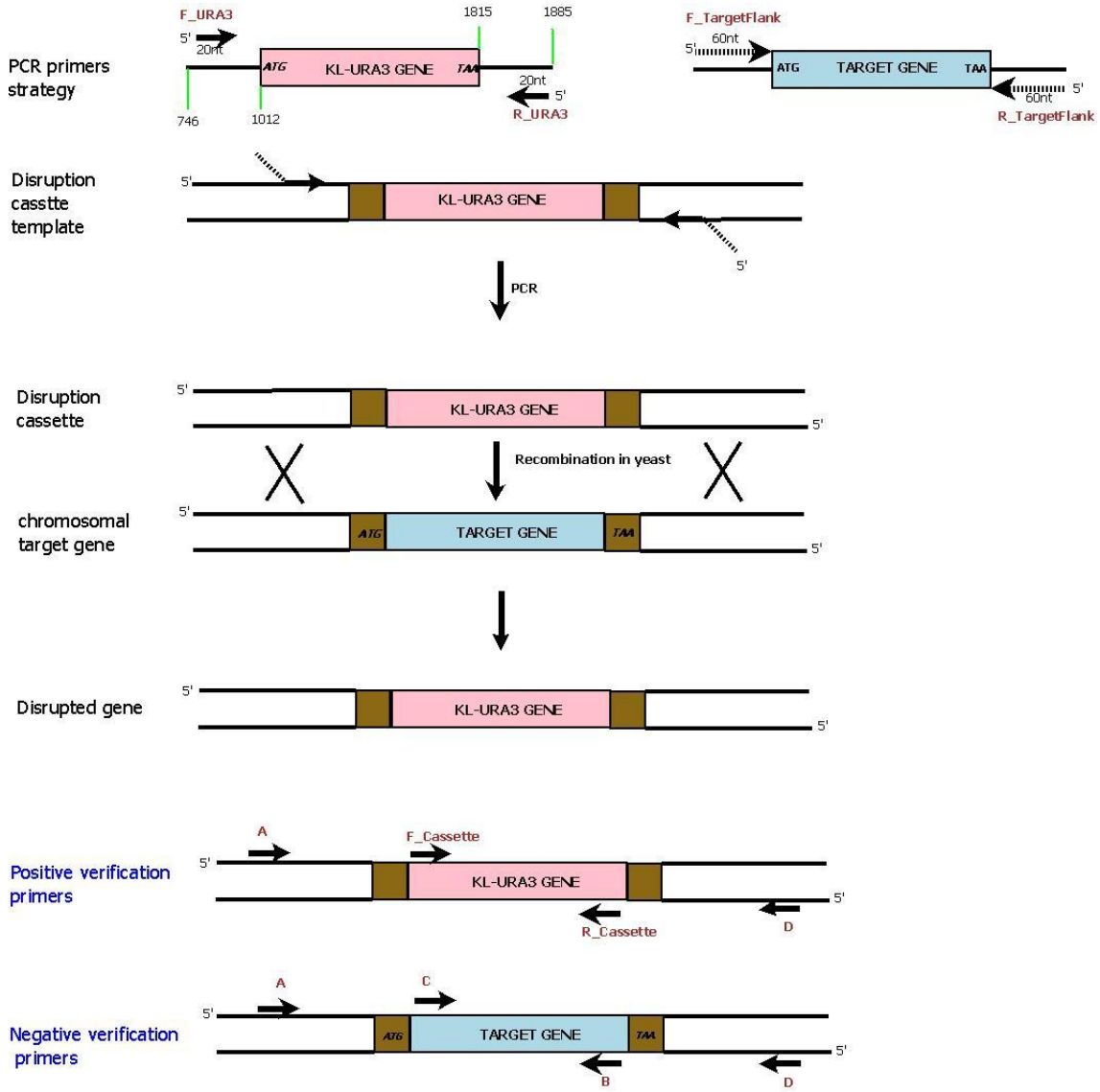

Figure 4.2: **Primer design strategies for PCR based gene deletion.** Primer design strategies for PCR based gene deletion of *S. cerevisiae*.

The generalised strategy for primer design for gene deletion and confirmation of results in all gene deletion experiments are described below:

1. YKO primers

   UP_45 and DOWN_45 PRIMERS are ORF deletion primer sequences in the Yeast Knock-Outs (YKO's) from *Saccharomyces* Genome Deletion Project (`http://www-sequence.stanford.edu/group/yeast\_deletion\_project/deletions3.`

`html`) consisting of primers used to make the MATa mating type strains (BY4741). UP_45 PRIMERS are 45 bases directly upstream of each yeast open reading frame, including the ATG while DOWN_45 are the 45 bases directly downstream of each yeast open reading frame, including the stop codon. All primers were analysed for GC content, melting temperature, hairpin loops and dimers; Artemis viewer was used to visualise the sequences for manual checks and NCBI Blast searches were also carried out on the primers to ensure that primers bind to the appropriate regions of the target gene in *S. cerevisiae*.

2. Extended YKO primers: Extended YKO primers were made by extending each of the UP_45 and DOWN_45 PRIMERS at the 5' end with the addition of between 15 -18 nucleotides, with final lengths ranging from 61 to 68. All primers were analysed for GC content, melting temperature, hairpin loops and dimers; Artemis viewer was used to visualise the sequences for manual checks and NCBI Blast searches were also carried out on the primers to ensure that primers bind to the appropriate regions of the target gene in *S. cerevisiae*.

3. Kl-URA3 primers: F_URA3 and R_URA3 primers were designed by finding 20 bases upstream and 21 bases downstream respectively of the ORF of URA3 gene. A single set of F_URA3 and R_URA3 primers was valid for gene disruption in all yeast strains. All primers were analysed for GC content, melting temperature, hairpin loops and dimers; Artemis viewer was used to visualise the sequences for manual checks and NCBI Blast searches were also carried out on the primers to ensure that primers bind to the appropriate regions of the kl-URA3 gene in pBS1539.

4. URA3_disruption cassette primers: ALT1_Disrupt, KGD2_Disrupt,LSC2_Disrupt, TDH1_Disrupt,TDH2_Disrupt and TDH3_Disrupt are gene disruption cassette primers consisting of F_URA3 and R_URA3 primers added to UP_45 and DOWN_45 PRIMERS at the 3' end respectively. The lengths of primers range from 84 to 112 nucleotides.

5. SC-LEU2 primers F_LEU2 and R_LEU2 primers were designed by finding 22 bases upstream and 19 bases downstream respectively of the ORF of LEU2 gene. A single set of F_LEU2 and R_LEU2 primers was valid for gene disruption in all yeast strains. All primers were analysed for GC content, melting temperature, hairpin loops and dimers; Artemis viewer was used to visualise the sequences for manual checks and NCBI Blast searches were also carried out on the primers to ensure that primers bind to the appropriate regions of the SC-LEU2 gene in pREP41.

6. LEU2_disruption cassette primers: ZWF1 disrupt are gene disruption cassette primers consisting of F_LEU2 and R_LEU2 primers added to UP_45 and DOWN_45 PRIMERS at the 3' end respectively. The lengths of primers range from 109 to 112 nucleotides.

7. Verification primers: YKO collection primers A and D were obtained from regions 200 to 400 bases upstream or downstream, respectively, of the open reading frame. Primers B and C were chosen from regions within the open reading frame and were designed to give PCR products with sizes of 300-1000 bases when used with A or D, respectively

   (a) Positive - Confirmation set 1: A primer: A_confirmation_primer_sequence R_Cassette primer

   (b) Positive - Confirmation set 2: D primer: D_confirmation_primer_sequence F_Casette

   (c) Positive - Confirmation set 3 A primer: A_confirmation_primer_sequence D primer: D_confirmation_primer_sequence

   (d) Negative - Confirmation set 4: A primer: A_confirmation_primer_sequence B primer: B_confirmation_primer_sequence

   (e) Negative - Confirmation set 5 D primer: D_confirmation_primer_sequence C primer: C_confirmation_primer_sequence

4.7.2 pBS1535 and pREP41 disruption cassettes

Gene disruption cassettes were amplified from plasmids pBS1539_kl and pREP41 EGFPC using disruption primer sets (Table 4.5) and Routine PCR (see section "Routine PCR").

4.7.3 Yeast transformation with disruption Casssettes

Transformation of yeast was carried out according to modified method of Gietz and Woods (Gietz and Woods, 2006).

4.7.3.1 Incubation

20 ml YEPD in 125 ml flask was inoculated freshly from a plate and incubated overnight at 30 °C in a shaker incubator (NewBrunswick Scientific I26). The overnight culture was added to 20 ml fresh YEPD in 125 ml flask to give $OD_{600} = 0.25$ and the diluted culture was incubated with agitation until cells have doubled to $OD_{600} = 0.8$ - 1.0.

4.7.3.2 Collection of cells

Cells were harvested in sterile 50ml Falcon tubes by centrifugation (Fischer Scientific accuSpin Micro17R) at 1,008 g for 5 min at room temperature. Next, cells were washed twice with 20 ml sterile millipore water, centrifuging as before. The cells were then resuspend in 1 ml sterile millipore water and then transferred to 1.5 ml eppendorf tube.

4.7.3.3 Preparation of competent cells

Cells were washed twice with 1 ml fresh TELiAc and collected by centrifuging (Fischer Scientific accuSpin Micro17R) at 1,008 g for 5 minutes. Next, the cells were

resuspended into a total volume of 1 ml TELiAc to give $2X10^9$ cells / ml.

4.7.3.4 Transformation

Carrier DNA (sheared salmon sperm) was heated in the incubator at 95 °C for 10 mins and then snap-cooled in ice. 8 µg of plasmid DNA (disruption cassette) and 10 µl of carrier DNA were pipetted into a 1.5 mL eppendorf tube. Next, 50 µl of yeast cells ($\sim 10^8$cells) were added into the mixture, followed by addition of 300 µl PEG/LiAc and the tube was vortex-mixed briefly. The transformation mixture was incubated at 30 °C for 30 min, followed by heat-shock of the cells at 42 °C for either 2 minutes, 15 minutes or 20 minutes.

4.7.3.5 Plating

The transformation mixture was centrifuged for 5 min at 1,008 *g* and the supernate discarded. A second centrifugation was carried out and the remaining PEG/TELiAc was removed by pipetting. Cell pellet was resuspended in 100 µl of sterile water and then 90 µl of transformed yeast cells were spread on URA- plates (selective media plates lacking URA) for 7 days of incubation at 30 °C.

Chapter 5

Constraint-based metabolic engineering

In this chapter, the results of constraint based analysis of iMM904 for the design of optimised *S. cerevisiae* strains are presented.

5.1 Results of Phenotypic flux plane analysis

Figure 5.1 shows the iMM904 three-dimensional phenotypic phase plane results for FBA maximisation of growth for each combination of glucose and oxygen uptake rates. The surface of the figure show 4 distinct regions of flat planes, indicative of phenotypically distinct metabolic states. In addition, the surface of the iMM904 three-dimensional phenotypic phase plane indicate the maximum growth rates possible at various combinations glucose and oxygen uptake rates.

Phase 1 is indicative of no growth as there is insufficient glucose in the system for ATP maintenance reaction. In phase 2, growth is hampered by the amount of oxygen, until the point between phase 2 and 3 is reached where maximum growth occurs. The remaining 2 phases show fermentative characteristics as growth becomes slower due to insufficient oxygen for glucose oxidation. The production of four of the target

products (lysine, fumaric acid, glutamate and trehalose) in *S. cerevisiae* is known to be favourable under aerobic cultivation of the organism, and production of ethanol in the same organism is more favourable under anaerobic growth.

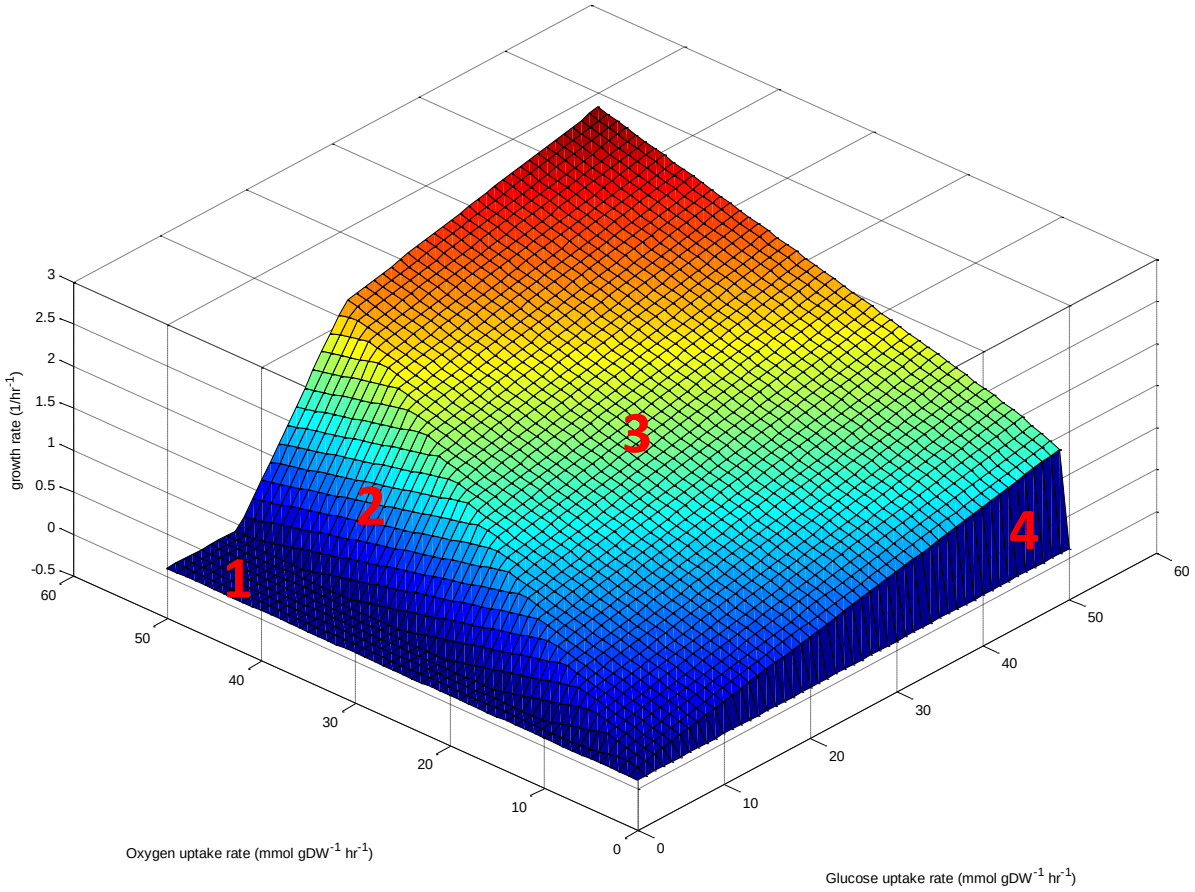

Figure 5.1: **3-D Phenotypic phase plane for iMM904**. Figure depicts a three-dimensional phenotypic phase plane for iMM904 model based on three dimensions of glucose uptake rate, oxygen uptake rate and growth rate.

In further characterising the phenotypic space of iMM904 so as to be as close as possible to experimental conditions, Table 5.1 shows the fluxes to biomass and the secretory profiles of the target products under four different growth conditions (aerobic, microaerobic, oxidative fermentation and anaerobic). This exploratory study indicated that highest production of biomass occurred in aerobic condition as expected, and the highest ethanol secretion was under microaerobic condition. There

are also appreciable amounts of fluxes to lysine in both aerobic and anaerobic conditions. Glutamate did not show any flux value above 0, and fluxes to trehalose decreased from aerobic to anaerobic conditions. Fluxes to fumaric acid did not follow any particular pattern.

Table 5.1: **Growth rates and product secretion.** Growth rates and product secretion for ethanol, lysine, fumaric acid, glutamate and trehalose.

	Fluxes					
Condition	Biomass	Ethanol	lysine	Fumaric acid	Glutamate	Trehalose
Aerobic	1.00268	18.58010	0.28697	5.43600	0	0.02300
Microaerobic	0.46309	30.84090	0.13254	0.17400	0	0.01100
Oxidative fermentation	0.63156	16.94870	0.18075	2.76300	0	0.01500
Anaerobic	0.17940	25.78130	0.05135	0	0	0.00210

5.2 Knockout predictions from OptKnock and GDLS

The results of OptKnock and GDLS strain design for production of ethanol are presented in Table 5.2. Using a full model of yeast iMM904 model and under microaerobic growth condition, GDLS identified 5 knockouts (glucokinase, hexokinase, homoserine O-trans-acetylase, phosphoserine transaminase and ribulose 5-phosphate 3-epimerase) predicted to have growth rate of 0.24 and a synthetic flux of 36.6 for ethanol. In comparison, OptKnock found only one knockout, citrate synthase, predicted to grow at the rate of 39.3. A knockout list for ethanol production under anaerobic condition (simulated ammonia,ergosterol and zymosterol in medium) by GDLS predicted lesser growth rate of 0.140 and ethanol excretion rate of 26.2 than under microaerobic growth condition. In this case, OptKnock did not converge to any solution, and hence the knockout list was empty. A knockout list based on a reduced iMM904 model, including O-acetylhomoserine (thiol)-lyase, Glucokinase, hexokinase, pyruvate carboxylase and phosphoglycerate dehydrogenase, was generated by GDLS; knockout set predicted a growth rate of 0.160 and excretory rate of 17.3 for ethanol under oxidative fermentation.

Table 5.2: **FBA strains for ethanol.** Table shows the *in silico* designed strains for ethanol production using OptKnock and GDLS. OUR = oxygen uptake rate (mmol gDW^{-1} hr^{-1}) and GUR = glucose uptake rate (mmol gDW^{-1} hr^{-1}). GLU5K = glutamate 5-kinase, GLUK = Glucokinase, HEX1 = hexokinase (D-glucose:ATP), ME2m = malic enzyme (NADP) mitochondrial and PSERT = phosphoserine transaminase, RPE = ribulose 5-phosphate 3-epimerase, HEX1 = hexokinase (D-glucose:ATP), ME2m = malic enzyme (NADP), mitochondrial, AHSERL2 = O-acetylhomoserine (thiol)-lyase, PC = pyruvate carboxylase, PGCD = phosphoglycerate dehydrogenase, HSERTA = homoserine O-trans-acetylase and CSp = citrate synthase.

	OptKnock	GDLS
Full model, GUR = -18.5, OUR = -1.5:		
Biomass flux	-	0.2437
Synthetic flux	39.3872	36.6885
Knockout cost	5	5
Knockouts	CSp	GLUK, HEX1, HSERTA, PSERT and RPE
Full model, GUR = -14, OUR = -0.01, ammonia = -5, ergosterol = -10, zymosterol = -10:		
Biomass flux	-	0.1383
Synthetic flux	-	26.3054
Knockout cost	5	5
Knockouts	0	GLU5K, GLUK, HEX1, ME2m and PSERT
Reduced model, GUR = -14, OUR = -1:		
Biomass flux	-	0.1608
Synthetic flux	-	17.3687
Knockout cost	-	5
Knockouts	-	AHSERL2, GLUK, HEX1, PC,PGCD

Table 5.3 shows the knockout lists for *in silico* design of lysine producing *S. cerevisiae*. The GDLS list, citrate lyase, pyruvate decarboxylase, oxoglutarate dehydrogenase, isocitrate synthase and malic enzyme (NADP), mitochondrial, predicts the excretion of lysine at a rate of 0.0100 and biomass production rate of 0.722 in aerobic condition. Under the same environmental conditions, the OptKnock solution include 3 knockout targets (citrate synthase, isocitrate lyase and pyruvate decarboxylase) for achieving a synthetic lysine flux of 6.5039, but no solution for biomass was indicated. The lack of biomass flux, and also the presence of large synthetic flux, in the OptKnock solution, may indicate an unreliable strain knockout strategy for developing lysine producing strains.

Table 5.3: **FBA strains for lysine.** Table shows the in silico designed strains for lysine production using OptKnock and GDLS. OUR = oxygen uptake rate (mmol gDW^{-1} hr^{-1}) and GUR = glucose uptake rate (mmol gDW^{-1} hr^{-1}). AKGDam = oxoglutarate dehydrogenase,acetyl-CoA synthetase, ME2m = malic enzyme (NADP), mitochondrial, PYRDC = pyruvate decarboxylase, CSp = citrate synthase, ICL = isocitrate lyase, and PYRDC = pyruvate decarboxylase.

	OptKnock	GDLS
Full model, GUR = -10, OUR = -18.5:		
Biomass flux	-	0.7220
Synthetic flux	6.5039	0.0100
Knockout cost	5	5
Knockouts	CSp, ICL and PYRDC	KGDam, CSp, ME2m and PYRDC

GDLS and OptKnock generated knockout lists for fumaric acid, trehalose and glutamate under various growth condition which predicted growth rates, but zero excretory rates for the target products

5.3 Discussion

In the phenotypic phase plane diagram (Figure 5.1), the yeast model iMM904, demonstrated the metabolic phenotypes useful for simulating *in silic*o design of strains for target products. Although it is a validated model for metabolic engineering purposes, characterisation of the phenotypic space for secretion of target products was important. The investigation to find out the right combinations of carbon source and oxygen led to simulations under 4 different conditions with varying amounts of glucose (carbon source) and oxygen.

Although, the GDLS knockout set, Glucokinase, hexokinase (D-glucose:ATP), malic enzyme (NADP) mitochondrial and phosphoserine transaminase, and ribulose 5-phosphate 3-epimerase, predicted the highest excretion rate of ethanol, the inclusion of both glucokinase and hexokinase just like any other knockout target demands caution. Glucokinase and hexokinase are two of the three enzymes phosphorylating glucose in the first irreversible step of glycolysis, and the deletion of the two might have disastrous consequences.

The GDLS (Isocitrate lyase, pyruvate decarboxylase, oxoglutarate dehydrogenase, isocitrate synthase and malic enzyme (NADP)) and OptKnock (citrate synthase, Isocitrate lyase and pyruvate decarboxylase) proposals for a lysine producing strain had two exact matches (citrate synthase and pyruvate decarboxylase), and most of the other enzymes are from the TCA cycle (oxoglutarate dehydrogenase,acetyl-CoA synthetase, malic enzyme (NADP), mitochondrial and citrate synthase,Isocitrate lyase). Considering the pathway for improving lysine production, succinyl-CoA ligase and oxoglutarate dehydrogenase, are among two of the TCA cycle enzymes identified by elementary flux mode analysis for enhanced production of lysine. Oxoglutarate dehydrogenase catalyses a crucial step in the TCA cycle and predicted by OptKnock, GDLS and EFM analysis as a point of intervention for lysine production in *S. cerevisiae*. This result further testifies to the credibility of the demonstrated increased lysine yield, based on EFM-designed strains, one of which included the deletion of oxoglutarate dehydrogenase. In addition, the close agreements of the three methods lends more credence to the usefulness of application of mathematical modelling to strain development.

5.4 Conclusion

A genome scale model (iMM904) of yeast was characterised for production capacities and in silico phenotype gene deletion analysis for target products. Design of *in silico* strains were achieved for enhanced ethanol and lysine production in *S. cerevisiae*.

Chapter 6

Elementary flux mode analysis

6.1 Introduction

Elementary flux mode analysis was used to study the metabolic network of yeast grown on glucose under aerobic condition. EFMs were computed for stoichiometric models derived from the metabolic network of reactions (see section 3.2). The ultimate purpose of EFM analysis in this project was to use its results for *in silico* phenotype gene deletion analysis, leading to prediction of knockout sets for enhancing the production of target products in *S. cerevisiae*. However, before the EFM data could be applied for this purpose, computational and clustering methodologies (section 3.2.2) were developed to reduce the dimensionality of the EFM data and also to cluster EFMs into groups that facilitate a quick and easy use of the EFM data for metabolic engineering purposes. A method based on regular expression was also developed to classify EFMs based on pre-defined classes. The results of the studies involving the computational processing, classification and clustering of the EFM data, including the *in silico* gene deletion knockout strategy for lysine production in *S. cerevisiae* are presented and discussed in this chapter.

6.2 Making sense of the elementary flux data

6.2.1 EFM analysis results

The results of COPASI computation of EFMs in stoichiometric models used in this study (Table 6.1) revealed the following numbers of detected EFMs: 325 for M1; 98 for M2; 16 for M3; 3090 for M4; 151 for SM1; 28 for SM2; 361 for S1; 1935 for S2; 2497 for S3; 26 for the Teusink and 69 for Cakir_Treh. Most of the EFMs obtained represent biosynthetic pathways starting from glucose and leading to various amino acids and other metabolites including ethanol, glycerol, acetate, succinate, malate, succinate, citrate etc. In addition, the production of these metabolites occurred in a number of different EFMs. Specifically, Table 6.1 shows the number of EFMs leading to particular target products. As examples, the production of glutamate by 199 EFMs, lysine by two EFMs, ethanol by 253 EFMs, fumaric acid by 985 EFMs and trehalose by five EFMs, in M1, SM2, S1, S3 and Teusink models respectively.

Table 6.1: **EFM analysis of stoichiometric models.** A summary of EFM analysis of stoichiometric models in terms of reactions, fixed external metabolites and EFM results. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

	Simulation output	
Stoichiometric Model	**No of EFMs**	**Metabolite (EFMs)**
M1	325	TYR (293), PHE (293), ALA (200), GLUT (199), ASP (63)
M2	98	PHE (83)
M3	16	GLUT (1), ASP (1)
M4	3090	SUC (1448), GOH (2899), ETOH (1777)
SM1	151	ETOH (138)
SM2	28	LYS (2)
S1	361	ETOH (253)
S2	1935	ETOH (1405), GLUT (1101)
S3	2497	FUM (985), LYS (814)
Teusink	26	TREH(5)
Cakir_TREH	69	ETOH (64)

6.2.2 Extraction of overall reaction and data matrix

In the first step of the computational extraction methodology (section 3.2.2.2, Figure 3.3), GetOR.java was used to convert the original EFMs into their stoichiometrically balanced overall equations ("ORs") and the associated component reactions for each EFM. Figure 6.1 is an example from a typical output file from GetOR.java program. EFM number 147 from the stoichiometric model S1 and its corresponding overall stoichiometry is presented as an example in Figure 6.1.

Elementary Flux mode:

Mode 147 Irreversible

27 * R1:HXT1,2,4: GLUC_ext = GLUC
27 * R6:FBA1: FRUCDP = GA3P + DHAP
10 * R7:TPI1: DHAP = GA3P
37 * R8:TDH1,2,3: GA3P + NADcyt = P13G + NADHcyt
37 * R9:PGK1: P13G + ADP = P3G + ATP
35 * R10:GPM1,2,3: P3G = P2G
35 * R11:ENO1,2: P2G = PEP
2 * R32:ACO1: CIT = ICIT
-12 * R40:FUM1: FUM = MAL
-12 * R41:MDH1: MAL + NADmit = OAC + NADHmit
12 * TRA28:FUM: FUM = FUM_EXT
12 * TRA2:ETOH: ETOH = ETOH_EXT
17 * TRA4:GOH: GOH = GOH_EXT
1 * TRA1:BIOMASS: BIOM -> BIOM_EXT
1 * R137:BIOM: 2 * 2OGLR + 2 * PEP + 2 * P3G + 3 * PYR + 2 * OAC -> BIOM
8 * R44:ShuttleX NADHcyt + NADmit -> NADcyt + NADHmit
2 * R33:IDH1,2: ICIT + NADmit -> 2OGLR + NADHmit + CO2
2 * R31:CIT1,3: OAC + ACCOAmit -> CIT
16 * R21:PYC1,2: PYR + ATP + CO2 -> OAC + ADP
2 * R20:PDA1,2,PDB: PYR + NADmit -> ACCOAmit + NADHmit + CO2
12 * R16:ADH1,4G: ACAL + NADHcyt -> ETOH + NADcyt
12 * R15:PDC1,2,5: PYR -> ACAL + CO2
17 * R14:GPP: GOH3P -> GOH
17 * R13:GPD1,2: DHAP + NADHcyt -> GOH3P + NADcyt
33 * R12:PYK1,2G: PEP + ADP -> PYR + ATP
27 * R4:PFK1,2G: FRUC6P + ATP -> FRUCDP + ADP
27 * R3:PGI1: GLUC6P -> FRUC6P
27 * R2:GLK1,HXKG: GLUC + ATP -> GLUC6P + ADP

Overall reaction:

Mode 147: 27 * GLUC_ext -> BIOM_EXT + 12 * ETOH_EXT + 17 * GOH_EXT + 12 * FUM_EXT

Figure 6.1: **An example of EFM and its overall stoichiometry.** Figure shows the reactions (and their coefficients) involved in the EFM, and also the corresponding overall stoichiometry. Sequence numbers of the reactions in the EFM are denoted by "R"s followed by the abbreviated reaction names. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A). Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A).

6.2.3 Clustering analysis of EFM data

6.2.3.1 Methodology based on Mclust and Wss approaches

In order to classify the EFMs into groups using hierarchical and kmeans clustering methods, a methodology was developed (Figure 3.2) which involved Mclust and Wss prediction and validation of optimal cluster number followed by clustering analysis. Mclust (a contributed package in R stastistical package) is a model-based approach which apply maximum likelihood estimation and Bayes criteria to identify the most likely model and number of clusters. Wss is within groups sum of squares. Both Mclust and a plot of the within groups sum of squares by number of clusters were used to find the optimal number of clusters.

Mclust and Wss methods for predicting the optimal number of clusters (k) were tested. Table 6.2 shows the optimal number of clusters predicted by mclust model clustering and within-group sum of squares for the Teusink model and models M1, M2, M3 and M4. For the Teusink model, mclust indicated two prediction models for the number of clusters (Table 6.2) out of which the best was a model with diagonal, equal volume and shape with 8 components, and the second best model contained 3 components. A plot of the within-group sum of squares carried out for the range of 1 to 10 cluster solutions suggested the best optimal number of clusters as either 2, 3, 5, 6, 7 or 8 clusters (Table 6.2) for the Teusink model.

Table 6.2: **Mclust and Wss.** Prediction of number of clusters by two methods - Mclust clustering method and within-group sum of squares on different stoichiometric models with EFMs as observations. Number of clusters suggested by second best model are in parenthesis. In addition, the numbers separated by comma represent the number of EFMs in a cluster group.

	Stoichiometric models and cluster solutions				
	Teusink	**M1**	**M2**	**M3**	**M4**
mclust	8 (3)	9	9	9	8
Wss	2,3,5,6,7,8	2,3,6	2,8	8	4,6

As Mclust model was able to suggest a more specific k value (8) for clustering the Teusink model than the Wss method (Table 6.2), hence the hierarchical clustering of the Teusink model was carried out by "cutting" the dendrograms based on $k = 8$ as predicted by the Mclust model.

6.2.3.2 Metrics and methods for Hierarchical clustering analysis

The best combination of metrics and methods for hierarchical clustering of the EFM data was investigated by carrying out hierarchical clustering analyses involving different metrics (Spearman, Pearson's correlation, and Euclidean) and distance linkage methods (Ward, Single, Complete and Average) on the Teusink data matrix. Table 6.3 indicates that the combination of Euclidean metrics with either "complete" or "single" method gave the same pattern of 8 clusters (9 1 5 5 2 1 2 1) on the Teusink glycolysis model. Figure 6.2 is a dendogram showing the 8-cluster solution using Euclidean metrics and complete method. Hence, further hierarchical clustering analyses on the data matrices from EFM datasets M1 - M4 were based on Euclidean metrics and the Complete method.

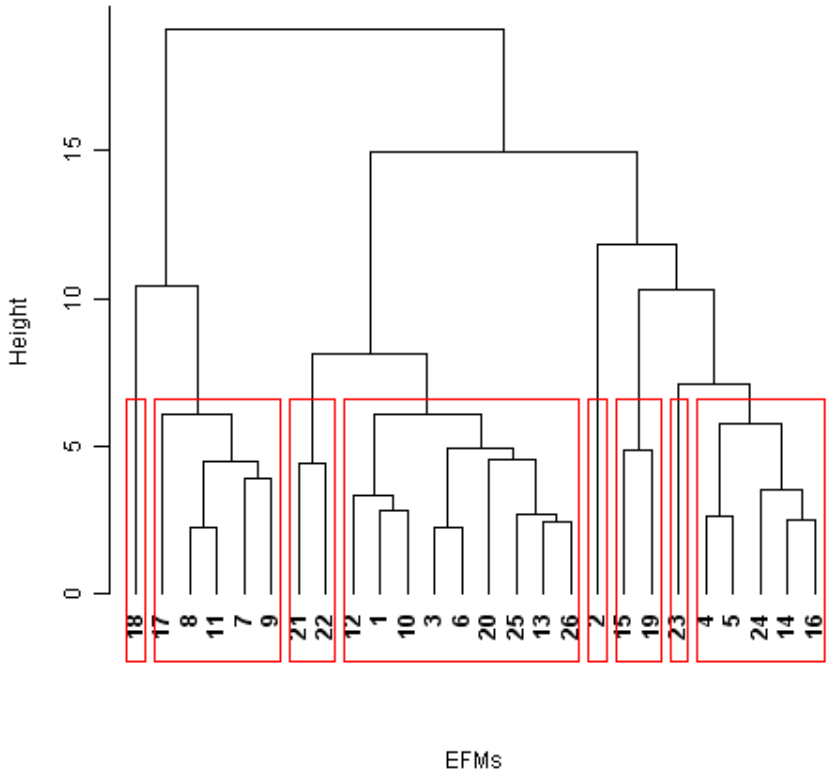

Figure 6.2: **Dendogram of hierarchical clustering of EFMs.** Dendogram of an 8-cluster solution of hierarchical clustering of EFMs in Teusink model.

6.2.3.3 *k*-means against hierarchical clustering of EFM data

Hierarchical clustering (using the Euclidean metric and complete method) and *k*-means clustering were carried out on the EFM data from the Teusink model and the stoichiometric models M1 - M4. Table 6.4 summarises the results of the comparisons of the EFMs in the different cluster groups based on different cluster solutions: Teusink (8 clusters), M1 (9 clusters), M2 (9 clusters), M3 (9 clusters), and M4 (8 clusters). It is apparent from Table 6.4 that the two clustering methods, *k*-means and agglomerative hierarchical clustering, yielded different clustering patterns when data from the four different models were subjected to clustering analyses by these two methods. Partitioning by the hierarchical method for both small and large datasets (M1 - M4) seems to be characterised by lumping of most EFM members into one

Table 6.3: **Cluster solutions for the Teusink model.** Cluster solutions for the Teusink model based on hierarchical clustering, using different metrics and distance methods. The numbers separated by comma represent the number of EFMs in a cluster group.

		Metrics and distance methods					
	Number of clusters	**Spearman with Complete**	**Pearson with Complete**	**Euclidean with Ward**	**Euclidean with Single**	Euclidean with Complete	Euclidean with Average
Teusink glycolysis model	8	8,7,1,3,2,2,2,1	4,8,6,1,2,3,1,1	3,1,5,7,5,1,1,2	1,9,1,1,1,1,1,1	9,1,5,5,2,1,2,1	9,1,5,5,2,1,2,1

group, leaving only a few (one EFM each in 6 cluster of M3)in the remaining clusters. k-means clustering partition large numbers into one group also, especially in M4.

Figure 6.3 shows a clustplot of 8 cluster solutions against first two principal components for the Teusink model based on k-means clustering, showing partitioning of 26 elementary flux modes into 8 cluster groups and cluster numbers appear in either red, blue or purple colour.

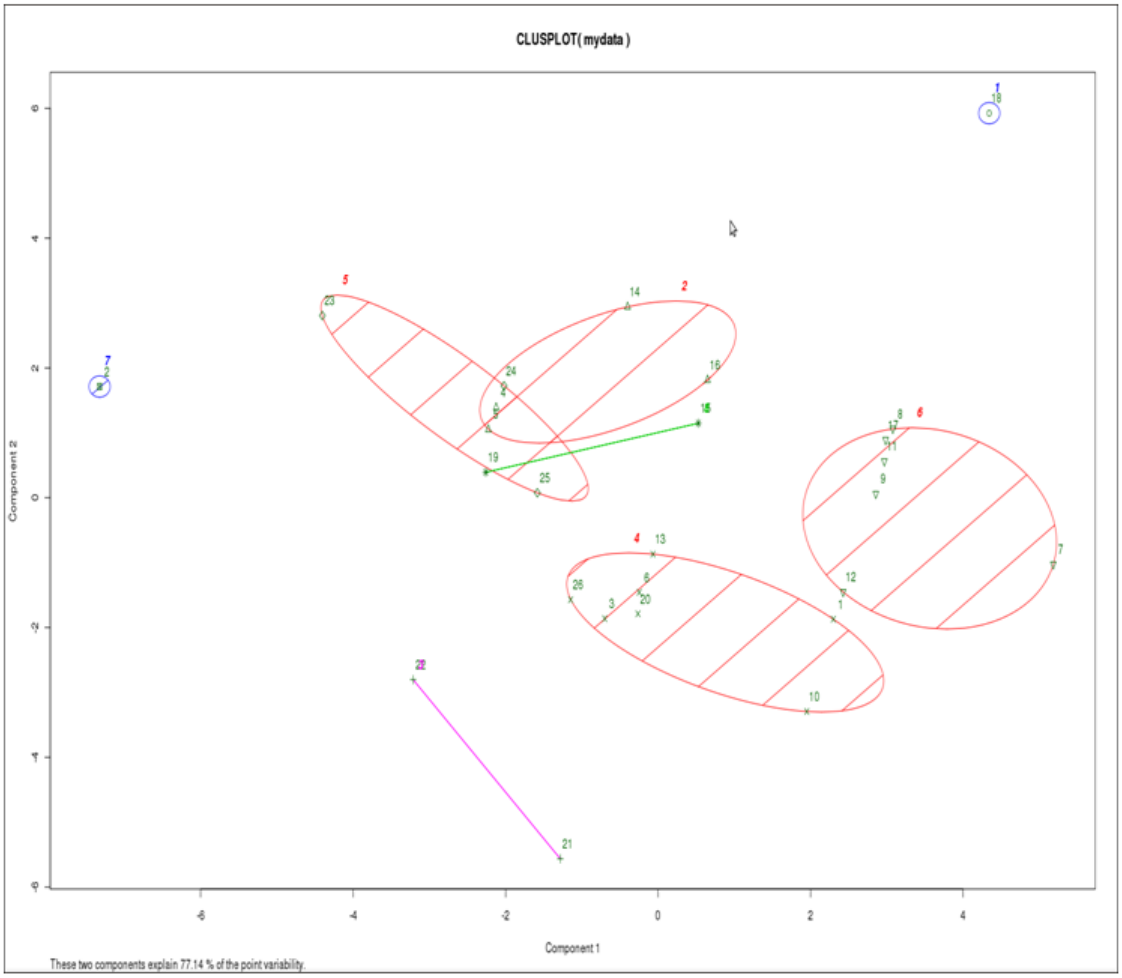

Figure 6.3: **Clustplot of 8 cluster solutions.** Clustplot of 8 cluster solutions obtained by plotting the first two principal components on the cluster solutions for the Teusink model. Cluster numbers appear in blue, red or purple colour.

Table 6.4: **Cluster solutions for 5 stoichiometric models.** Cluster solutions for 5 stoichiometric models based on hierarchical and k-means clustering, showing partitioning of elementary flux modes into different cluster groups. The numbers separated by comma represent the number of EFMs in a cluster group.

	Teusink	**M1**	**M2**	**M3**	**M4**
Agglomerative Hierarchical clustering(Euclidean,Complete)	8 clusters: 9,1,5,5,2,1,2,1	9 clusters: 303,2,6,2,4,2,2,2,2	9 clusters: 74,6,6,2,2,2,2,2,2	9 clusters: 5,1,3,1,1,2,1,1,1	8 clusters: 3061,5,6,4,6,4,2,2
K-means clustering	8 clusters: 1,7,1,2,1,2,5,7	9 clusters: 8,27,26,2,34,16,12,154,46	9 clusters: 4,16,7,9,20,16,12,10,4	9 clusters: 1,1,6,1,2,2,1,1,1	8 clusters: 31,88,407,22,241,24,314,1963

6.2.3.4 Methodology involving clValid validation

This section reports the findings from "clValid" comparisons of 7 different clustering algorithms based on EFM matrix from from stoichiometric model S1 (Table 6.1) corresponding to only EFMs producing both ethanol and biomass (i.e, EFM matrix S1a). As EFM matrix S1a was fairly large, the strategy in "route 2" of the methodology illustrated in Figure 3.9 was applicable. Hence, the results of clustering EFM matrix S1a (contains rows of 64 EFMs and a subset of EFMs with glucose as substrate and also producing ethanol and biomass) using 7 different methods, are presented here. The optimal numbers of clusters were determined by evaluating 2 to 10 clusters, and internal validation of cluster solutions was achieved using Dunn's index and Silhouette. Table 6.5 depicts the optimal numbers of clusters suggested and validated by Dunn's index and Silhouette in clValid package for EFM matrix S1a and also for the Teusink matrix which was added for comparisons.

Table 6.5: **Comparisons of seven clustering methods.** Comparisons of seven different clustering algorithms and the optimal number of clusters suggested by clValid for clustering EFMs in S1 and the Teusink models.

	Agglomerative Hierarchical	Diana	K-Means	PAM	Clara	Fanny	Model (Mclust)
Teusink	10	10	10	10	10	-	8
EFM dataset S1a	4	4	4	4	4	3	5

Tables 6.6 - 6.12 show the cluster solutions as result of Diana, Clara, Model, Kmeans, Fanny, PAM and Agglomerative Hierarchical clustering, respectively, of EFM matrix S1a.

Table 6.6: **4-cluster solution using Diana clustering method.** A summary of a 4-cluster solution from Diana clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/Non-core members) to each cluster, and also the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	57	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.406
2	4	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-ALA_EXT; L-LYS_EXT; SUC_EXT	None	0.258
3	2	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-ALA_EXT; L-LYS_EXT; SUC_EXT	None	0.258
4	1	BIOM_EXT; ETOH_EXT; FUM_EXT; GOH_EXT; L-LYS_EXT	None	0.413

Table 6.7: **4-cluster solution using Clara clustering method.** A summary of a 4-cluster solution from Clara clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/Non-core members) to each cluster, and also the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	31	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.392
2	7	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	FUM_EXT; L-ALA_EXT; L-GLUT_EXT	0.352
3	19	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-LYS_EXT	0.449
4	7	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; SUC_EXT	0.280

Table 6.8: **hierarchical 4 cluster solutions for model M1.** A summary of a 4-cluster solution from model-based (Mclust) clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/Non-core members) to each cluster, and also the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	26	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLUT_EXT	0.360
2	8	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-GLUT_EXT	0.359
3	24	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-LYS_EXT; SUC_EXT	0.472
4	6	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-ALA_EXT; L-LYS_EXT; SUC_EXT	None	0.258

Table 6.9: **4-cluster solution using k-means clustering method.** A summary of a 4-cluster solution from k-means clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/Non-core members) to each cluster, and also the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	7	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; SUC_EXT	0.280
2	7	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	FUM_EXT; L-ALA_EXT; L-GLUT_EXT	0.352
3	19	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-LYS_EXT	0.449
4	31	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.392

Table 6.10: **4-cluster solution using Fanny clustering.** A summary of a 4-cluster solution from Fanny clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/Non-core members) to each cluster, and also the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	25	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.345
2	10	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-LYS_EXT	0.471
3	13	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-GLUT_EXT; SUC_EXT	0.301
4	16	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLUT_EXT; L-LYS_EXT	0.491

Table 6.11: **4-cluster solution using PAM clustering method**A summary of a 4-cluster solution from PAM clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/non-core members) to each cluster, and in terms of the average molar yields of ethanol produced by the EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	50	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.414
2	7	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	FUM_EXT; L-ALA_EXT; L-GLUT_EXT	0.352
3	4	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; SUC_EXT	0.292
4	3	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-ALA_EXT; L-LYS_EXT; SUC_EXT	None	0.264

Table 6.12: **4-cluster solution using agglomerative hierarchical clustering.** A summary of a 4-cluster solution from agglomerative hierarchical clustering of EFM matrix S1a, in terms of external metabolites common (core members) and not common (additional/non-core members) to each cluster, and in terms of the average molar yields of ethanol produced by the EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A)

Cluster	No of EFMs	Common/core metabolites	Non-core members	Ethanol yield
1	50	BIOM_EXT; ETOH_EXT; GOH_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.414
2	7	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	FUM_EXT; L-ALA_EXT; L-GLUT_EXT	0.352
3	6	BIOM_EXT; CO2_EXT; ETOH_EXT; GOH_EXT; L-ALA_EXT; L-LYS_EXT; SUC_EXT	None	0.258
4	1	BIOM_EXT; ETOH_EXT; FUM_EXT; GOH_EXT; L-LYS_EXT	None	0.413

The results of Diana and Hierarchical methods (Tables 6.6 and 6.12) are similar in that most of the EFMs are partitioned into cluster 1 (with 50 or more), and cluster 4 in either case is made up of only one member (EFM number 359) with ethanol yield of 0.413. The "outlier" EFM is interesting since it is a high ethanol producing mode, compared with the all other EFMs; this EFM requires glutamine as a co-substrate and in addition to ethanol, produces biomass and by-products (lysine, fumaric acid, ethanol and glycerol). The reason for the agreements in the Hierarchical and Diana 4-cluster solutions may be because Diana clustering method is an example of divisive hierarchical approach to clustering which is implemented by dividing clusters until each cluster contain a single observation, while agglomerative hierarchical is a similar algorithm but works from the opposite direction. Even though the classification of EFM 359 into one cluster by hierarchical and Diana clustering methods may seem interesting, this may prove to be of limited use since further reduction of the 50 EFMs in cluster 1 may be required to reveal further the intrinsic groupings and characteristics of the EFM data. To achieve this, it is necessary to "cut" the tree further, followed by re-clustering, and even then there is no guarantee how many of the other EFMs would be redistributed to yield more meaningful clusters for biotechnological purposes.

Tables 6.7 and 6.9 show the same 4-cluster solutions (31, 7, 19, 7 and 7, 7, 19, 31 respectively) for the 64 EFMs in EFM S1a by Clara and Kmeans algorithms. The 4-cluster solutions (26, 8, 24, 6) obtained from Model clustering algorithm (Table 6.8) are also similar to those of Clara and Kmeans. The PAM partitioning of the EFM data is similar to that of Hierarchical method in two clusters of 50 and 7 EFM members (clusters 1 and 2 respectively) with ethanol molar yields of 0.414 and 0.352 respectively.

Out of the 7 clustering methods tested, only PAM-based 4-cluster solutions for EFM dataset S1a (Table 6.11) showed cluster separations based on different "core metabolites" as follows: cluster 1 (biomass, ethanol and glycerol), cluster 2 (biomass, carbon dioxide, ethanol, glycerol and lysine), cluster 3 (biomass, ethanol, glycerol and lysine) and cluster 4 (biomass, carbon dioxide, ethanol, glycerol, alanine, lysine and succinate). The cluster solutions for the 6 other clustering methods (Tables 6.6 - 6.10 and 6.12) indicated that more than one cluster shared the same "core metabolites". In addition, the cluster mean yields for ethanol in the in PAM-based cluster solutions were more representative of ethanol yields for each EFM than cluster mean yields from other clustering algorithms. These findings may suggest that cluster separation was better using PAM than with the other 6 clustering algorithms. Clusters were further distinguished from each other based on the types of metabolites constituting the "additional/non-core membership of each cluster, which indicates the by-products associated with the EFMs in each cluster. Furthermore, if ETOH had been chosen as the main product, a core member of every cluster, the information in the "core metabolites" or "additional/non-core members" column would be instructive as to the different types of potential by-products produced by EFMs from different clusters.

It can be concluded that from the analysis of the results of comparisons of 7 clustering methods for the clustering of EFM data that only the PAM clustering algorithm partitions the EFM data into clusters with similar substructures distinguishable according to different "core memberships" (ethanol, biomass and either glycerol or amino acids), while also permitting opportunity to consider the types of metabolites

constituting the "additional/non-core membership" of each cluster. Furthermore, PAM clustering does not partition the EFM data into a cluster with a single member and PAM Cluster mean yields of the target metabolite were more reliable than the cluster mean yields from the other clustering methods. Hence, PAM clustering method allows the interpretation of the EFM classes to find biological meaning applicable to gene deletion phenotype analysis. Henceforth, PAM clustering algorithm was judged to be the most suitable algorithm for clustering the EFM data.

6.2.4 Complexity reduction in EFM data

6.2.4.1 Computational extraction of high-dimensional variables

The effects of dimensionality reduction in EFM dataset S1a was investigated by comparing the results of PAM clustering of EFM dataset S1a reported (Table 6.11) with the results of PAM clustering of EFM dataset S1b 6.13. EFM dataset S1a was computed from model S1, and it comprises of all EFMs leading from glucose to the external metabolite of interest (ethanol) and biomass; Ethanol and biomass are regarded as the first and second biological variables of interest, respectively, for data complexity reduction in this case. EFM dataset S1b, also computed from model S1, comprises of all EFMs leading from glucose to the external metabolite of interest (ETOH), whether or not they also produce biomass; biomass is not regarded as the second biological variable of interest for data complexity reduction in this case. Validation based on Dunn's index suggested a 2-cluster partitioning of EFM dataset S1b using PAM and agglomerative hierarchical clustering.

As shown in Table 6.13, 254 EFMs were partitioned into two clusters of 243 EFMs with ethanol as the criterion for core-membership, and 11 EFMs with ethanol, glycerol and lysine as the criteria for core-membership, with cluster mean yields for ethanol of 0.481 and 0.305 respectively. The usefulness of these results for biotechnological purposes is limited as the EFMs have been partitioned into only two clusters. This ineffective data partitioning is reflective of inadequate data reduction as the structures

of the EFM dataset S1b was not sufficiently exposed for effective partitioning using clustering analysis. The information contained in the additional/core memberships of this 2-cluster solution suggest that further partitioning could have been effected by using another member such as biomass.

Table 6.13: **2-cluster solution using PAM clustering method.** A summary of a 2-cluster solution using PAM clustering for EFM dataset S1b. EFM clustering results show external metabolites that are common (core members) and the external metabolites that are not common (additional/non-core members) to each cluster and the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A)

Clusters	No of EFMs	Core members	Additional/Non-core members	ETOH yield
1	243	ETOH_EXT	BIOM_EXT; CO2_EXT; FUM_EXT; GOH_EXT; L-ALA_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	0.481
2	11	ETOH_EXT; GOH_EXT; L-LYS_EXT	BIOM_EXT; CO2_EXT; FUM_EXT; L-ALA_EXT; L-GLUT_EXT; SUC_EXT	0.305

By way of comparison, the results of the 4-cluster partitioning of EFM dataset S1a using PAM (Table 6.11) is more useful for making biotechnological decisions than the results of the 2-cluster partitioning of EFM dataset S1b using the same clustering method (Table 6.13). Table 6.11 indicates that there was a better separation of clusters in a PAM-based 4-cluster solution for EFM dataset S1a. In addition, all clusters had ETOH and BIOM as part of core members, but were distinguished from each other based on the rest of the metabolites constituting the core membership of each cluster. Clusters were further distinguished from each other based on the cluster ethanol mean yields which are representative of the constituent EFMs.

6.2.4.2 Regular expression method

Figure 6.4 depicts an example of output of ranked EFM "ORs" (reduced to fit page) as a result of pattern analysis on S2 dataset (containing 1935 EFM "ORs") to find the best EFMs for ethanol.

```
Group 1: (No. of EFMs = 0)
No               EFM                                        Product Yield        Biomass Yield
        None

Group 2: (No. of EFMs = 0)
No               EFM                                        Product Yield        Biomass Yield
        None

Group 3: (No. of EFMs = 2)
No               EFM                                        Product Yield        Biomass Yield
53        23 * GLUC_ext -> BIOM_EXT +
        4 * FUM_EXT + 12 * ETOH_EXT +
        17 * GOH_EXT + 8 * CO2_EXT                           0.521739        0.043478
55        27 * GLUC_ext -> BIOM_EXT +
        12 * FUM_EXT + 12 * ETOH_EXT +
        17 * GOH_EXT                                         0.444444        0.037037

Group 4: (No. of EFMs = 0)
No        EFM        Product Yield        Biomass Yield
        None

Group 5: (No. of EFMs = 765)
No               EFM                                        Product Yield        Biomass Yield
307        235 * GLUC_ext + 180 * L-GLN_EXT +
        120 * ACAL_EXT -> 100 * L-ALA_EXT +
        90 * L-LYS_EXT + 5 * BIOM_EXT +
        120 * ETOH_EXT + 205 * GOH_EXT +
        260 * CO2_EXT + 80 * L-GLUT_EXT                      0.510638        0.021277
315        355 * GLUC_ext + 180 * L-GLN_EXT ->
        100 * L-ALA_EXT + 90 * L-LYS_EXT +
        5 * BIOM_EXT + 120 * ETOH_EXT +
        325 * GOH_EXT + 380 * CO2_EXT +
        80 * L-GLUT_EXT                                      0.338028        0.014085
```

Figure 6.4: **An example of output of ranked and classes of EFM "ORs".** Figure depicts an example of output of ranked and classes of EFM "ORs" (reduced) as a result of pattern analysis to find the best EFMs for lysine.

This example was based on finding all EFM "ORs" which produce ethanol according to the class specifications in section 3.2.3. The outputs (Figure 6.4) are arranged according to the user's specifications of the defined classification implemented in the algorithm. Out of 1935 EFM "ORs", 767 were ranked into two groups of 2 and 765 respectively. There were no EFMs "ORs" ranked into the first and second groups. In order for EFM "ORs" to be ranked in groups 1 and 2, it is required that the EFMs consume either only glucose or glucose together with ammonia, respectively, and are both producers of target metabolite, ethanol and biomass, with no by-products. However, two EFM "ORs" (EFM No 53 and EFM No 55) were placed into group 3; they both consume glucose as the main substrate, and produce ethanol, biomass,

fumaric acid and glycerol. The two modes produce the same amount of biomass (molar yield = 0.04), but the molar yield of ethanol (0.52) produced in EFM No 53 was higher than in EFM No 55 (0.04).

In Group 5, the EFMs consume other substrates (fumaric acid, acetaldehyde and amino acids) in addition to glucose as the main substrate, and they produce other products, apart from ethanol. The two EFMs (EFM No 53 and EFM No 55) in group 3 are attractive from the biotechnological point of view. The EFMs in groups 5 are less attractive in that since they consume other metabolites in addition to glucose and also produce more by-products.

6.3 EFM modelling for strain development

6.3.1 Biological interpretation of a medium EFM data

In this section, a medium-sized EFM dataset was used to illustrate how the methodology involving cValid can be used to accomplish the tasks of finding, interpreting and using the patterns revealed by clustering analyses of EFMs for aiding important decisions that can facilitate *in silico* gene deletion phenotype analysis for ethanol production in yeast.

Data extraction for biomass and ethanol yielded 767 out of 1935 EFMs contained in the original EFM data from stoichiometric model S2 (Table 6.1), and these were included in EFM dataset S2a. The extracted EFMs represent 60.4% reduction from the thousands of EFMs in the original EFM data. Validation based on Dunn's index and silhouette suggested a 5-cluster partitioning of EFM dataset S2a using the PAM clustering method. Table 6.14 shows that there were similarities in the core memberships of clusters 1 and 2 (based on biomass and ethanol), clusters 3 and 5 (biomass, ethanol and lysine), and only cluster 1 was partitioned based on biomass, ethanol, glycerol and lysine. However, the 5 EFM clusters indicate different mean ethanol yields. Clusters without glycerol in their core membership (clusters 1, 2, 3

and 5) have cluster mean ethanol yield of greater than 1.0. Although, the non-core members for clusters 1, 2, 3 and 5 indicate that some of the constituent members produce glycerol. EFMs in cluster 4 had the lowest cluster mean ethanol yield of 0.345. The EFMs in all clusters produce biomass, ethanol and also by-products such as glycerol, glutamate, succinate, lysine, aspartate and alanine.

Table 6.14: **PAM clustering for EFM data set S2a.** A summary of a 5-cluster solution using PAM clustering for EFM data set S2a. EFM clustering results show external metabolites that are common (core members) and the external metabolites that are not common (additional/non-core members) to each cluster and the average molar yields of ethanol produced by EFMs in each cluster. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A)

Clusters	No of EFMs	Core members	Additional/Non-core members	ETOH Yield
1	404	BIOM_EXT; ETOH_EXT	CO2_EXT; FUM_EXT; GOH_EXT; L-ALA_EXT; L-ASN_EXT; L-ASP_EXT; L-GLN_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	1.501
2	212	BIOM_EXT; ETOH_EXT	CO2_EXT; FUM_EXT; GOH_EXT; L-ALA_EXT; L-ASN_EXT; L-ASP_EXT; L-GLUT_EXT; L-LYS_EXT; SUC_EXT	1.877
3	110	BIOM_EXT; ETOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; GOH_EXT; L-ALA_EXT; L-ASN_EXT; L-ASP_EXT; L-GLUT_EXT; SUC_EXT	1.918
4	19	BIOM_EXT; ETOH_EXT; GOH_EXT; L-LYS_EXT	CO2_EXT; FUM_EXT; L-ALA_EXT; L-GLUT_EXT; SUC_EXT	0.345
5	22	BIOM_EXT; ETOH_EXT; L-LYS_EXT	FUM_EXT; GOH_EXT; L-ALA_EXT; L-ASN_EXT; L-ASP_EXT; SUC_EXT	2.133

Inspection of the overall stoichiometry of the EFMs partitioned into clusters clusters 1 - 5 revealed that, in addition to glucose as the main substrate, other metabolites were consumed in most of the EFMs in the different clusters. Acetaldehyde was used as the co-substrate by most EFMs in cluster 1, while glutamine was consumed in addtion to glucose in the EFMs parttioned into cluster 4. A combination of co-sustrates, glutamine and acetaldehyde, glutamine, alanine and acetaldehyde, was used in most of the EFMs in clusters 2, 3 and 5, respectively. The overall stoichiometric equations of EFM number 53, EFM number 49 and EFM number 284 are shown in Table 6.15 as examples from clusters 1, cluster 2 and cluster 3, respectively. Overall stoichiometry of the EFMs in cluster 1 reveals that 2 (EFM numbers 53 and 55) out of 404 EFM members of this cluster produce ethanol and glycerol requiring either only glucose

as substrate. EFM 53 and 55 requires only glucose and produces ethanol (0.52 and 0.44 molar yield respectively) and biomass (0.04 and 0.04 molar yield respectively). Although the molar ethanol yields from these two EFMs are much lower than the mean cluster ethanol yield of 1.50.

Table 6.15: **"ORs" for EFMs 53, 49 and 284.** Table depicts the overall stoichiometry and the molar yields of ethanol for EFMs 53, 49 and 284. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A)

EFM Number (and Cluster)	Overall reaction	Molar yield (Ethanol)
53 (Cluster 1)	23 * GLUC_ext → BIOM_EXT + 4 * FUM_EXT + 12 * ETOH_EXT + 17 * GOH_EXT + 8 * CO2_EXT	0.52
49 (Cluster 2)	93 * GLUC_ext + 36 * L-GLN_EXT + 108 * ACAL_EXT → 18 * L-LYS_EXT + 36 * L-ASP_EXT + 9 * BIOM_EXT + 108 * ETOH_EXT + 45 * GOH_EXT	1.16
284 (Cluster 3)	255 * GLUC_ext + 130 * L-GLN_EXT + 10 * FUM_EXT → 80 * L-ALA_EXT + 60 * L-LYS_EXT + 5 * BIOM_EXT + 80 * ETOH_EXT + 235 * GOH_EXT + 260 * CO2_EXT + 60 * L-GLUT_EXT	0.31

The EFM overall stoichiometry permits the ranking of EFMs into groups that facilitate biological interpretation and important biotechnological decisions. It is possible to rank the EFM overall stoichiometry according to either (a) EFMs which require only glucose as substrate for production of ethanol, (b) EFMs which require a second substrate in addition to glucose for production of ethanol, (c) EFMs which produce specific by-product(s) in addition to ethanol. Hence, the ranked EFMs enables the best decision process and prevents choosing members of clusters 1 - 5 simply because these EFMs produce the highest molar ethanol yield. In essence, the phenotypic solution space represented by the EFMs requiring co-substrates is not biologically feasible without the external addition of the co-substrate as indicated in the overall stoichiometry. The biochemical routes requiring only glucose are cheaper options than those with a requirement for organic materials in terms of production cost. The production of ethanol in EFM 53 and 55 appears to be particularly attractive options considering the lower cost of the required substrate, and also they are biologically realisable routes producing biomass and an ethanol. For the above reasons, EFMs 53 and 55 are good candidates for in *silico* phenotype deletion studies.

As stated earlier, EFMs can be considered as a minimal set of enzymes necessary for the production of specic metabolites that operate at steady state, and represent all the capabilities of a metabolic network, that is, all the phenotypes that can be expressed in the organism. Hence, EFM analysis permits the design of *in silico* phenotype gene deletion studies. Using EFM 53 as an example, a list of potential gene knockout targets should include all reactions not in EFM 53, but found in the other EFMs of clusters 1 - 5. These reactions are then carried through in silico gene knockout simulations based on the iterative steps of (1) deletions of single reaction (gene) or multiple reactions (genes) from the reaction network followed by (2) EFM analysis in COPASI, until the EFM results show that only EFM 53 remains as a biochemical route. Ranking the reactions (genes) used for knockout simulations according to their effectiveness in lowering the number of EFMs will lead to identication of the best target gene knockouts for improving the yield of ethanol.

6.3.2 *In silico* gene knockout simulations

6.3.2.1 Single gene deletion for lysine

Table 6.16 shows that only 2 modes were left from 2 rounds of extractions of EFM subsets from the 68 EFMs contained in the original EFM set (E0) of model SM2 (Table 6.1 using "Route 1" of the methodology illustrated in Figure 3.9).

Table 6.16: **EFM subsets from the original EFMs for model SM2.** The results of computational extractions EFM subsets from the original EFMs for model SM2. Table depicts the number of EFMs left after a two-stage computational extraction of EFM subsets from the original EFM dataset

EFM set	Number of Modes
Original set (E0)	28
After First extraction (subset 1)	24
After Second extraction (subset 2)	2

Table 6.17: **"ORs" for EFMs 11, 17 and 28.** Table depicts the overall reactions and the molar yields of lysine for modes 11, 17 and 28. Abbreviations of metabolites are explained in a file (Metabolite abbreviations.xls in Appendix A).

Mode	Overall reaction	Molar yield
17	4 * NH3_ext + 78 * GLUC_ext → 34 * CO2_EXT + 12 * FUM_EXT + 40 * ETOH_EXT + 59 * GOH_EXT + 3 * BIOM_EXT + 2 * L-LYS_EXT	0.03
28	4 * NH3_ext + 95 * GLUC_ext → 46 * FUM_EXT + 40 * ETOH_EXT + 59 * GOH_EXT + 3 * BIOM_EXT + 2 * L-LYS_EXT	0.02
11	8 * NH3_ext + 37 * GLUC_ext → 26 * CO2_EXT + 12 * FUM_EXT + 20 * ETOH_EXT + 28 * GOH_EXT + 4 * L-LYS_EXT	0.1

Further examination of the overall reactions of these two modes (numbers 17 and 28, Table 6.17) show that they are both characterised by glucose as substrates in reactions where two biological objectives of interest, lysine and biomass, are produced, albeit along with by-products (ethanol, fumaric acid and glycerol). Table 6.17 also shows that in the overall reactions of mode number 11 (not part of EFMs from subset 2, but part of subset 1), glucose is the substrate and only one of the biological objectives of interest, lysine, is realised in the biochemical route. Although mode 11 has higher lysine yield (0.1) than modes 17 and 28, the lack of production of biomass indicates no growth and hence the possibility that this mode is not feasible.

Mode 17 was chosen for further *in silico* work towards improving the yield of lysine in yeast since the two biological objectives, lysine and biomass, are realisable in this biochemical route, and also it requires slightly less amount of the main substrate, glucose, than mode 28.

Table 6.18 shows that the best single gene knockouts are R71:ALT2, R36:KGD1,2, R37:LSC1,2 and R113:GLT1 as a result of *in silico* simulation of the effects of genetic modification based on EFM analysis after the deletion of each of the 17 candidate single gene deletants for enhancing production of lysine in *S. cerevisiae*. The points of genetic intervention and the resultant flow of flux are shown in Figure 6.5. The limited in vivo effectiveness of these single deletants (R71:ALT2, R36:KGD1,2, R37:LSC1,2 and R113:GLT1) can be gauged from the numbers of operational EFMs left (32 - 34) after each deletion, far from the expected result of only 1 EFM.

Table 6.18: ***In silico* single gene deletion analysis for lysine.** Results of *in silico* single gene deletion analysis based on model SM2 for improved lysine production. Table depicts the effectiveness of the deletion of a reaction (gene) on the network of reactions as "No of reactions left". The more effective a gene deletion, the less the number of EFMs, which in turn indicates the degree of elimination of the competing pathways. Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A).

Deleted reaction (gene)	No of EFMs left (Original EFM total = 28)
R23:ZWF1	48
R24:SOL1,2,3,4	48
R25:GND1,2	48
R26:RKI1	48
R27:RPE1	48
R28:TKL,TKI	48
R29:TAL1	48
R30:TKI,TKL	48
R33:IDH1,2	48
R36: KGD1,2	34
R37:LSC1,2	34
R67:ARO8a	49
R68:ARO9a	49
R69:ARO8b	49
R70:ARO9b	49
R71:ALT1	32
R113:GLT1	34

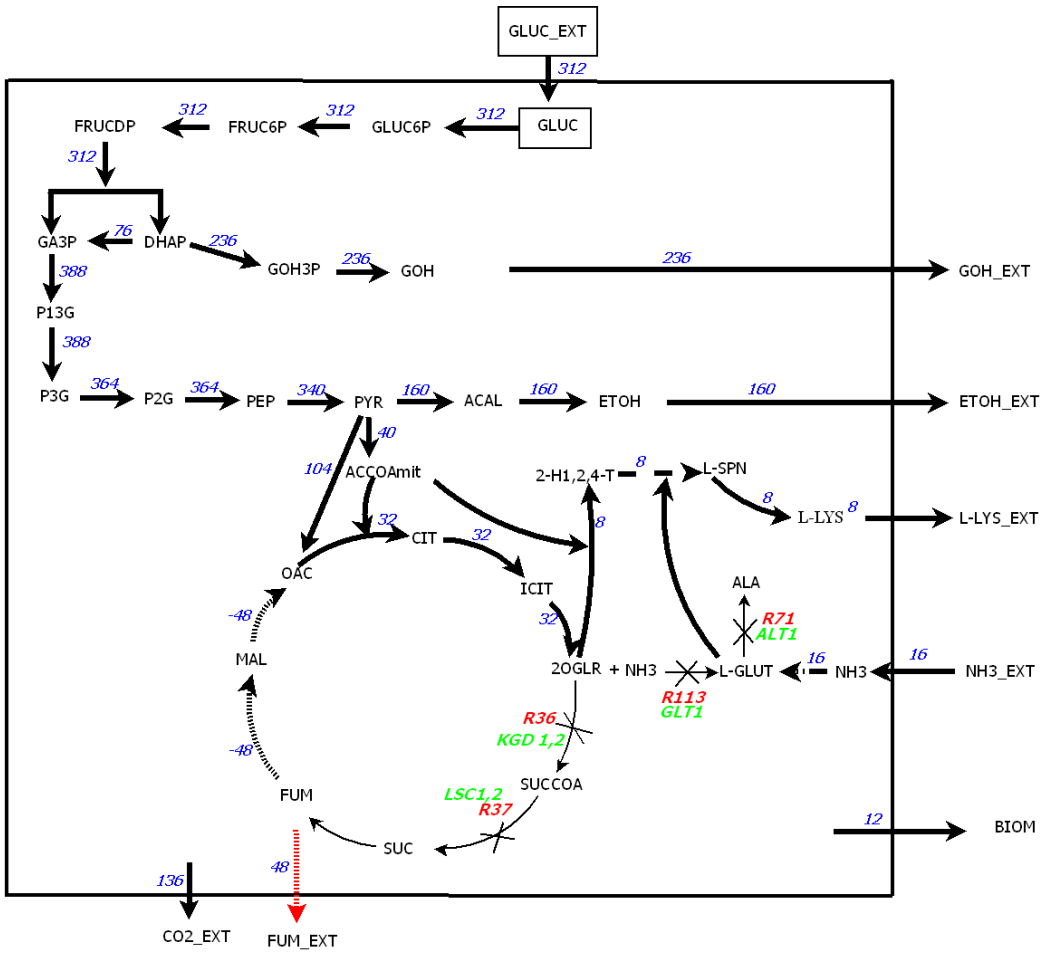

Figure 6.5: **A network of reactions with points of intervention for flux redirection.** A network of reactions showing the points of intervention for redirecting carbon flux towards increasing the yield of lysine based on *in silico* single gene knockouts. R36 is reaction number 36 representing the genes for the enzymes KGD1 and KGD2; R37 is reaction number 37 representing the genes for the enzymes LSC1 and LSC2; R71 is reaction number 71 representing the gene for the enzymes ALT1 and R113 is reaction number 113 representing the gene for the enzyme GLT1. Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A). Each of the numbers in blue represents the flux value of a reaction.

6.3.3 Double gene deletion for lysine

After 136 double-combinations of genes were simulated *in silico* for the effects of double-gene deletion, the combinations of R36 X R71 (KGD1/KGD2 X ALT1), R37 X R71 (LSC1/LSC2 X ALT1) and R71 X R113(ALT1 X GLT1) were the best in terms of the number of operational EFMs left, 18, 18 and 22, respectively. Although it was not possible to carry out *in silico* simulation of the double combinations of isoenzymes with the model, it was judged best to include the following two double

combinations to the set obtained by simulation: KGD1 X KGD2 and LSC1 X LSC2. Hence, in order to account for the isoenzymes of R36 and R37 for the purpose of wet experiment validation of mutant strains, the following 5 double deletants were chosen: KGD1 X ALT2, KGD2 X ALT1, LSC1 X ALT1, LSC2 X ALT1 and ALT1 X GLT1. Lysine yields are expected to be higher in these yeast double mutants than in the yeast single mutants.

6.3.3.1 Triple gene deletion for lysine

Triple knockout strategy involving *in silico* simulations of the effects of the triple deletions based on "crossing" one gene more with each of the best double deletants (section 6.3.3) resulted in the best triple mutants as follows:

1. R71:ALT2 R36:KGD1,2 R23:ZWF1
2. R71:ALT2 R37:LSC1,2 R23:ZWF1
3. R71:ALT2 R113:GLT1 R23:ZWF1

Table 6.19 indicate that the *in silico* simulated genetic effects of these triple deletion are the closest to the expected result as only 2 operational modes are left, and it was anticipated that these outstanding effects will be replicated experimentally.

Table 6.19: ***In silico* triple gene deletion analysis.** Results of *in silico* triple gene deletion analysis based on model SM2 for improved lysine production. Table depicts the effectiveness of the deletion of a reaction (gene) on the network of reactions as "No of reactions left". The more effective a gene deletion, the less the number of EFMs, which in turn indicates the degree of elimination of the competing pathways. Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A).

Deleted triple reactions (genes)	No of EFMs left (Original EFM total = 28)
R71:ALT2 R36:KGD1,2 R23:ZWF1	2
R71:ALT2 R37:LSC1,2 R23:ZWF1	2
R71:ALT2 R113:GLT1 R23:ZWF1	2
R71:ALT2 R36:KGD1,2 R33:IDH1,2	4
R71:ALT2 R37:LSC1,2 R33:IDH1,2	4
R71:ALT2 R113:GLT1 R33:IDH1,2	4

However, due to the consideration for the isoenzymes of KGD and LSC, the final list was expanded as follows:

1. R71:ALT2 R36:KGD1 R23:ZWF1
2. R71:ALT2 R36:KGD2 R23:ZWF1
3. R71:ALT2 R37:LSC1 R23:ZWF1
4. R71:ALT2 R37:LSC2 R23:ZWF1
5. R71:ALT2 R113:GLT1 R23:ZWF1

6.4 Discussion

The capabilities of *S. cerevisiae* to produce the target metabolites (ethanol, lysine, glutamate, trehalose and fumaric acid) was investigated using EFM analysis. The phenotypic solution space for the production of each target product was defined

by fixing the target product as "external" metabolite. The alternative biochemical routes, both biologically and non-biologically feasible, from the substrate to the target product characterising each solution space were represented as EFMs. However, the complex nature of the EFM data, in terms of EFM number and biochemical composition, necessitated the development of methodologies for an easy and quick access to biotechnologically useful information in the data.

A method for classification of EFMs based on motif finding has been reported (Peres et al., 2011); even though this method was able classify EFMs based on motifs of EFM reactions, it does not lend itself to classifying EFMs for metabolic engineering purposes. Hence, the motivation for developing a novel methodology was the need to simplify the use of EFM data for enhancing microbial strain development, especially when large EFM datasets are involved.

The methodology based on Mclust and Wss (Figure 3.2) was fraught with a number of shortcomings, among which was the fact that even though the k value predicted by Mclust was better that that of Wss, the final clustering step using Kmeans method was unstable. Kmeans clustering method provided different cluster solutions of the same data and reclustering was necessary to get fairly reproducible cluster solutions. The partitioning of the EFM data by Hierarchical clustering method was also poor in this case. Hence, a better methodology permitting reliable prediction of k value and more stable cluster solutions was necessary.

A second methodology, involving clValid validation (Figure 3.9), allowed for the EFM data to be decomposed into manageable subsets of EFMs, allowing fast detection of alternative biochemical routes in the metabolic network for the development of improved yeast strains that can produce specific metabolites of interest. In the first step of this methodology, EFM datasets were compiled using the COPASI software. Next, the complexity of these datasets was computationally reduced (section 3.2.2.2, Figure 3.3), whereby only the EFMs producing biomass and a metabolite of interest were retained for further analysis. The effectiveness of this step was demonstrated in section 6.3.1 with 60.4% reduction in the dimensionality of a medium-sized EFM

dataset.

Detecting the true structures in the full feature space of biological data is difficult as a result of impact of noise and the problem may be reduced by applying a drastic reduction of variables with the highest dimension across the datasets (Handl et al., 2005). External metabolites and biomass are variables with high dimension in the EFM dataset, and as such can be useful in reducing the dimensionality of EFM data. Both biomass and the metabolite of interest are variables with high dimension in the EFM dataset. Using these two biological objectives to reduce the EFM dataset helps meaningful partitioning by clustering analysis in the next step. The extraction step was also useful in removing futile cycles and other EFMs that are not complete routes from the "input" external glucose to the target external metabolites.

Out of the 7 clustering methods (UPGMA, K-means, PAM, SOM, FANNY, CLARA and DIANA) tested, PAM was the optimal clustering method for obtaining subsets of EFMs, yielding biological interpretation of EFM data for biotechnological purposes. The results of analysis showed that PAM analysis, using Dunn's index and Silhouette width as internal validation measures, is the best clustering method for obtaining subsets of EFMs yielding biologically useful information for gene deletion phenotype analysis from the EFM data. As shown in Table 6.11 (section 6.2.3.4), clusters are well separated by PAM method judging from the different core-members used for partitioning of EFMs. The mean cluster yields for ethanol in EFM dataset S1a represent a good guide for the ethanol yields of individual EFMs in different groups.

PAM is a more robust version of K-means based on the search for k representative objects or medoids among the observations of the dataset (Kaufman and Rousseeuw, 1990); it minimises a sum of dissimilarities instead of a sum of squared Euclidean distances as is the case with Kmeans. Medoids are objects in the cluster whose average dissimilarity to all the objects in the cluster is minimal. This is perhaps why PAM performed better than the other clustering methods, since the "metabolite" objects found in different EFMs were partitioned based on the minimisation of the dissimilarities between them.

The EFM classification approach based on regular expression is a quick and flexible approach for reducing the dimensionality of EMF data for metabolic engineering purposes as demonstrated with S2 dataset in section 6.2.4.2. This approach is advantageous over clustering analysis in that the most useful classes are exposed quickly, ranked in order of biotechnological usefulness. The location of most economically feasible EFMs (EFM No 53 and EFM No 55) in EFM S2 dataset demonstrated the usefulness of this methodology.

The usefulness of the inspection of the overall stoichiometric reactions for EFMs in aiding the identification of the best EFMs was demonstrated both with a medium-sized and small-sized EFM datasets. In the case of the small-sized dataset, *in silico* phenotype gene deletion analysis was carried out based on the EFM identified through the inspection of the overall stoichiometric reactions, and which eventually led to wet experiment validation of single, double and triple mutants for enhanced lysine production.

6.5 Conclusion

The results of studies presented in this chapter indicate that the general modelling and data reduction approaches contributed in obviating the enormous problems associated with trying to obtain the EFMs from large reaction network models and interpreting the resulting of large number of EFMs. The biological significance of the approaches for quick and efficient deciphering of the EFM data for information useful towards designing a target gene knockout strategy was outlined. It was possible to find biologically and economically feasible EFMs for high yield of products based on overall stoichiometry of EFMs. The approach for classifying EFMs based on pre-defined classes also added an important feature to the metabolic engineering toolbox based of EFM analysis.

Chapter 7

Construction and validation of *S. cerevisiae* mutant strains for lysine production

This chapter covers the experimental work for model validation, which is the "synthesis stage" of the metabolic engineering pipeline (Figure 1.1) employed in this study for lysine production in yeast. The results of the construction of *S. cerevisiae* mutant strains and the experimental validation of the mutants for improved lysine yield are presented and discussed.

7.1 *S. cecevisiae* single mutants and lysine production

Two types of experiments, mutant growth characterisation and metabolite measurements using GC/MS, were carried out for the validation of the single gene knockoutout mutantss (Δalt1, Δalt2, Δkgd1, Δkgd2, Δlsc1, Δlsc2 and Δglt) predicted by *in silico* analysis for increased lysine production. The aim of the mutant growth characterisation experiments was to determine the growth characteristics in 3XALL

medium, namely the specific growth rates, doubling times, the time intervals(in hours) covering the log and stationary phases of growth. The aim of the experiment involving metabolite measurements using GC/MS was to grow the single mutants and collect log and stationary phase sample for growth, exometabolome and endometabolome measurements along time points in 3XALL medium.

7.1.1 Growth curve characteristics for single mutants

The log phase and early stationary phase were found to be from 4 to 10 hours and 10 to 18 hours of growth respectively.

In order to determine the appropriate medium pH and concentrations of amino acid and uracil supplements to be added to SD medium for growing SC and *S. cerevisiae* mutants, an experiment was designed, involving control strain (CS) and Δ*glt1*, to investigate the effects of different concentrations of amino acids at pH6 on growth and doubling times of the *S. cerevisiae* single mutants (Table 7.1).

Table 7.1 indicates that the doubling times of CS and mutant GLT1 in NC medium are 101 minutes and 113 minutes respectively. Furthermore, the smallest doubling time for CS is in NC_PH6 medium (90 minutes) is the same as the reported time for wild type yeast in YPD medium, and the smallest doubling time for mutant GLT1 is in 2XLEU medium (79 minutes). However, the doubling time for mutant GLT1 in NC_PH6 medium is 103 minutes which is an improvement in the doubling time for the same mutant grown in NC medium (113 minutes). 2XLEU medium reduced the doubling time for mutant GLT1 but increased the doubling time for CS slightly by 7 minutes when compared with growth of both strains in NC. Further observation indicates that CS had the same doubling time in both NC and 2XALL media. The closest doubling times for CS and mutant GLT1 occurred in the 3XALL medium, 107 minutes and 110 minutes respectively. More importantly, the growth curves for CS and mutant GLT1 in 3XALL medium showed more reliable growth characteristics than in any other medium, and hence 3XALL medium was chosen as the growth

medium for further growth experiments of yeast single mutants.

Table 7.1: **Doubling times for single strain.** Table shows doubling times for CS and single mutant strains grown in different culture media. Doubling times were calculated from the steepest areas of the growth curves (6 - 10 hours of growth. Abbreviations of enzymes (genes) are explained in a file (Reactions abbreviations.xls in Appendix A).

Strain	Culture medium	Doubling time (minutes)
CS	NC	101
GLT1	NC	113
CS	2XLEU	110
GLT1	2XLEU	79
CS	3XLEU	108
GLT1	3XLEU	126
CS	2XALL	101
GLT1	2XALL	124
CS	3XALL	107
GLT1	3XALL	110
CS	NC_PH6	90
GLT1	NC_PH6	103

7.1.2 GC-MS analysis for lysine production by single mutants

The excretion of 5 metabolites into the culture medium by 7 *S. cerevisiae* mutant and CS strains during the exponential growth phase in 3XALL medium is depicted in Figure 7.1 and Table 7.2. Table 7.2 shows the mean and standard deviation values

for intracellular and extracellular lysine in CS and double mutants. The results of GC-MS quantified metabolites (lysine, 2-oxoglutarate, glutamate, glycerol, fumaric acid and phenylalanine), at exponential growth phase of the *S. cerevisiae* single mutants, indicate that out of the 6 yeast mutant strains, only Δlsc2 and Δglt1 excreted significantly more lysine ($P < 0.05$) than CS into the culture medium. Mutant strains Δlsc2 and Δglt1 excreted 35.4umol/l and 39.8 umol/l of lysine respectively, while CS excreted 5 umol/l of lysine, representing about 5 fold higher excretion of lysine by both Δlsc2 and Δglt1 than that of CS. The amounts of lysine excreted by Δ*lsc1*, Δ*alt2*, Δ*kgd1* and Δ*kgd2 S. cerevisiae* mutant strains were found not to be significantly higher than that of CS ($P > 0.05$). However, Δ*LYS20*Δ*LYS21*, an *S. cerevisiae* mutant strain known for intracellular accumulation of lysine (Feller et al., 1999) and used as a positive control did not show any significantly higher level of excreted lysine than CS. The results also indicate that neither the single mutants yeast strains (Figure 7.1) nor the control mutant strain (Δ*LYS20*Δ*LYS21*) excreted significantly higher amounts of 2-oxoglutarate, glutamate, glycerol and fumaric acid. Furthermore, undetectable levels of lysine, 2-oxoglutarate, glutamate and phenylalanine were observed in blank media samples (not inoculated with any *S. cerevisiae* strain) as expected.

Table 7.2: **Mean and standard deviation values for lysine in CS and single mutant strains.** Mean and standard deviation values for extracellular secretion and intracellular accumulation of lysine by CS and single mutant strains during mid-log phase of growth. Mean values are in micromoles/litre. SD represents standard deviation.

	Extracellular mean (SD)	Intracellular mean (SD)
LSC2	35.4 (8.9)	2.4 (1.4)
GLT1	39.8 (16.5)	4.9 (0.4)
KGD1	11.1 (10)	1.2 (0.7)
LSC1	9.3 (4.4)	1.29 (1.1)
KGD2	14.1 (9.2)	2.9 (2.2)
LYS20/LYS21	10.1 (4.6)	82.0 (25.8)
CS	5 (1.4)	1.1 (0.8)

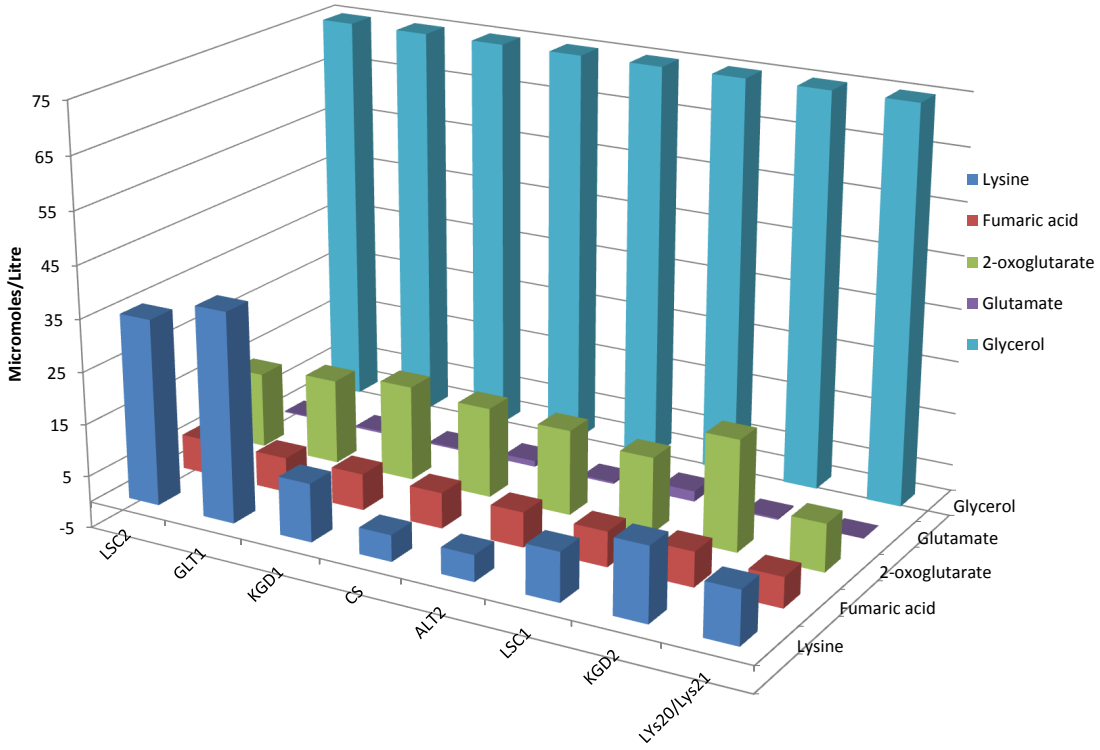

Figure 7.1: **Excretion of lysine by single mutants.** Figure shows excretion of lysine (and 4 other metabolites) by CS and 6 single mutant strains, into the culture medium during mid-log phase of growth.

Figure 7.2 shows the intracellular accumulation of 5 metabolites by 6 *S. cerevisiae* single mutant strains during the exponential growth phase in 3XALL medium. Intracellular accumulation of lysine during the exponential growth phase of the *S. cerevisiae* single mutants $\Delta lsc2$ and $\Delta glt1$, and also $\Delta LYS20\Delta LYS21$, were found to be significantly higher ($P < 0.05$) than in CS. The results of intracellular accumulation of showed similar pattern to the results obtained for the excretion of lysine. Only $\Delta glt1$ accumulated significantly higher amounts of phenylalanine than CS during this growth phase, and also none of the remaining metabolites (2-oxoglutarate, glutamate, glycerol and fumaric acid) showed significant accumulation in any of the single *S. cerevisiae* mutants, including $\Delta LYS20\Delta LYS21$ strain.

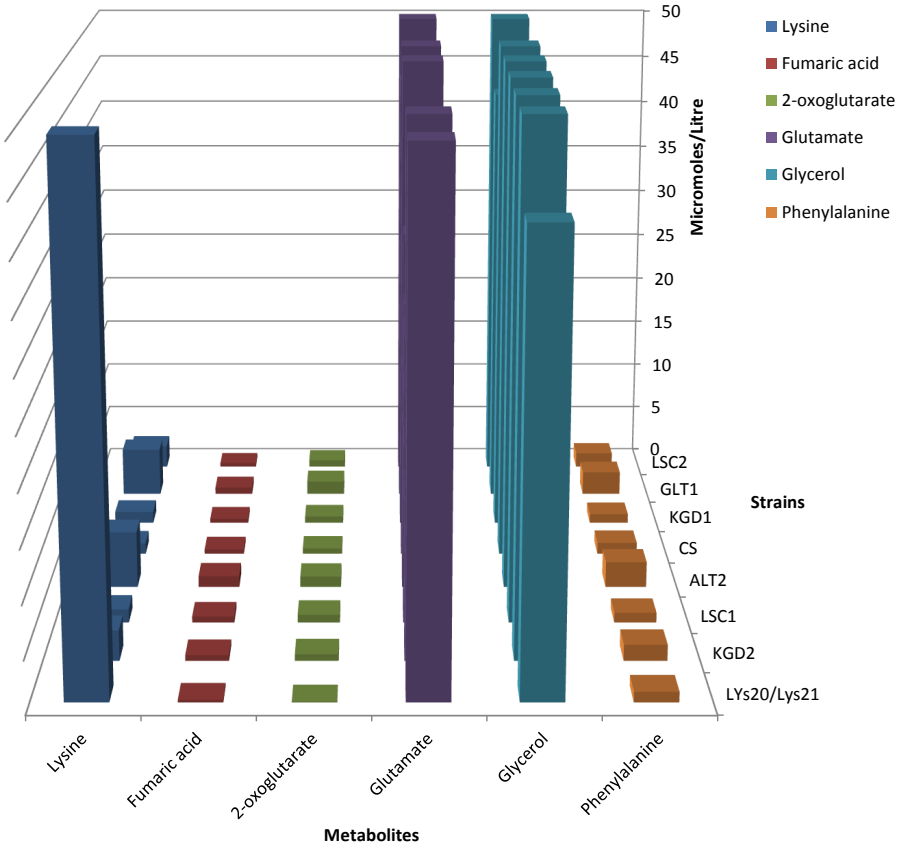

Figure 7.2: **Intracellular accumulation of lysine by single mutants.** Figure show intracellular accumulation of lysine (and 5 other metabolites) by CS and 6 single mutant strains, during mid-log phase of growth.

GC-MS results also reveal that, at early stationary phase of growth, none of the metabolites (lysine, 2-oxoglutarate, glutamate, glycerol, fumaric acid and phenylalanine) was excreted at significantly high levels by any of the single mutants yeast strains $\Delta lsc1$, $\Delta lsc2$, $\Delta glt1$, $\Delta kgd1$, $\Delta kgd2$ and $\Delta alt2$, while only $\Delta LYS20\Delta LYS21$ excreted a higher level of phenylalanine compared with CS.

7.2 Toronto double mutants

In order to reduce the amount of work involved in the experimental validation of *in silico* designed strains, it was decided to obtain *S. cerevisiae* single and double mutants from the Charlie Boone Lab (Toronto, Canada).

7.2.1 GC/MS results for lysine by Toronto double mutant strains

GC-MS analysis of the endometabolome of the Torornto *S. cerevisiae* strains was measured for metabolites, including lysine, and the results are shown is Figure 7.3. These intracellular metabolome results were unexpected and confounding. The most striking features are the lower intracellular yields of 2-oxoglutarate for all double mutants than for YLR123C (a negative control with same genotype as the double mutants but with double mutations not related to amino acid metabolism). The higher intracellular levels of serine and alanine in all mutants (except *kgd1xkgd2*) than in YLR123C are also remarkable. Intracellular concentrations of aspartate, fumaric acid and lysine are also lower for all 7 mutants than YLR123C, while there are no appreciable differences between the intracellular levels of glycerol, phenylalanine, glutamate (except lower values for except *alt2* x *kgd2*) and succinic acid contents (except *alt2* x *kgd2* and *kgd1* x *kgd2*).

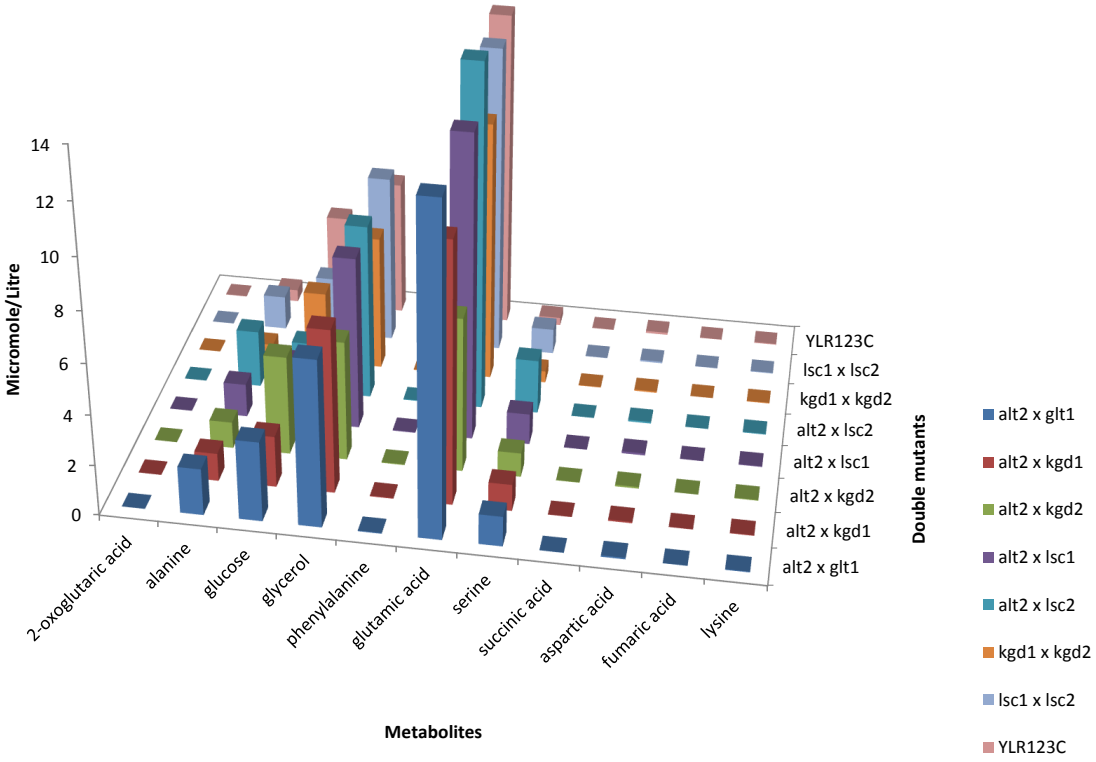

Figure 7.3: **Endometabolome measurements for metabolites in Toronto strains.** Results for endometabolome measurements for lysine and 10 other metabolites, in Toronto *S. cerevisiae* strains during mid-log phase of growth.

7.3 In-house constructed double mutants

Due to the problems encountered with interpretation of the GCMS results of the Toronto *S. cerevisiae* strains (Table 4.2), it was decided to carry out in-house construction of the double mutants($\Delta kgd1\Delta alt1$, $\Delta kgd2\Delta alt1$, $\Delta lsc1\Delta alt1$, $\Delta lsc2\Delta alt1$ and $\Delta alt1\Delta glt1$) shown in Table 4.3. Each of the Euroscarf single mutant strains, served as a recipient single mutant strain on which one deletion was added to produce a double mutant according to Table 4.10. The construction of triple mutants ($\Delta kgd2\Delta alt1\Delta zwf1$, $\Delta lsc2\Delta alt1\Delta zwf1$ and $\Delta alt1\Delta glt\Delta zwf1$) was based on the in-house constructed double mutants ($\Delta kgd2\Delta alt1$, $\Delta lsc2\Delta alt1$ and $\Delta alt1\Delta glt$, which

were used as recipient double mutant strains (Table 4.11).

In the simple and efficient one-step approach for direct gene deletion in *S. cerevisiae* (Baudin et al., 1993) used in this study, the general principle involved transforming yeast cells with DNA fragment consisting of a "gene disruption cassette" that provides a selectable phenotype (URA3 or LEU2 prototrophy), surrounded by 90 - 118 base pairs of sequence flanking the sequence to be deleted. The purified PCR product was used to transform yeast cells, and recombinants that have inherited the disruption cassette were selected. Cells that have correctly integrated the disruption cassette were identified by detecting PCR products generated using primers complimentary to sequence within the cassette and primers flanking the site of integration of the selectable marker. PCR products of expected size were obtained if the disruption cassette was inserted into the genome by homologous recombination.

7.3.1 Construction of *S. cerevisiae* double mutant strains

7.3.1.1 Double gene deletion transformants

Amplification of 3 types of disruption cassettes from plasmid pBS1539_kl using forward and reverse primers pairs (Figure 4.5) for the disruption of target genes (*ALT1*, *KGD2* and *LSC2*)in order to create 5 different *S. cerevisiae* double mutants. Colonies of strains obtained on URA- selection plate were checked for transformants after 7 days of transforming *S. cerevisiae* strains with plasmid pBS1539_kl DNC (disruption cassette). Table 7.3 show the number of colonies obtained for each of the strains heat-shocked for 2 minute and 15 minute during the process of transformation.

Table 7.3: **Double mutant colonies.** Table shows the number of double mutant colonies from two different heat-shocks. 22 colonies in total were picked for further incubation

Double Mutant strain	colonies	No of colonies for further incubation
2 minute heat-shock:		
M1	1	1
M4	1	1
M5	2	2
15 minute heat-shock:		
M1	1 (very big), 60 (small)	3
M2	8 - big	3
M3	11- big	3
M4	3 - big	3
M5	10 (big), 10 (small)	6

18 colonies for all 5 strains (Tables 7.3) were picked and re-plated in selective media plates (URA-) in order to increase the chance of obtaining the correct transformants. After 6 days of incubation of the 18 re-plated isolates, colony PCR was carried out on some of the colonies, and success of gene disruption was verified using 5 sets of verification primers (3 positive and 2 negative). Results indicate that the 5 double yeast mutants have been successfully constructed by the distribution of the appropriate genes in yeast single mutants. Figure 7.4 shows an example of a successful gene deletion for the creation of *S. cerevisiae* double mutants from this study.

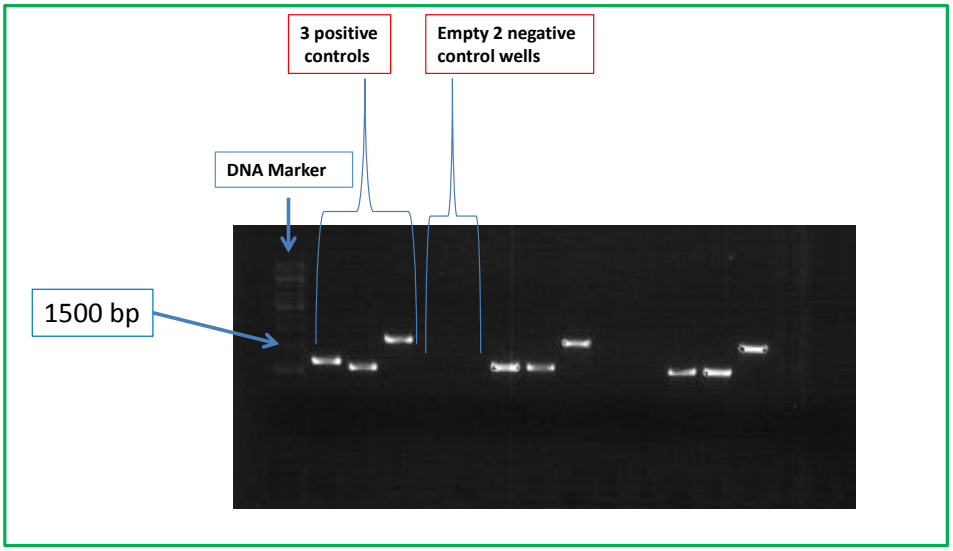

Figure 7.4: **PCR confirmation of gene deletion for double mutants.** Figure shows the result of PCR confirmation of gene deletion for double mutants. Gene deletion was successful as three bands of correct sizes (1300 bp, 1300 bp and 1653 bp) for positive controls and no band for negative controls were obtained.

Out of the 30 transformant colonies (re-plated) screened for all 5 mutant strains (22 for 15 min heat-shock and 8 for 2 min heat-shock), there were successful gene disruptions in the 2 minute heat shock category for strain M4 (2 transformants) and M5 (2 transformants), and also the genes were successfully disrupted for the 15 minute heat shocked strains for M1 (2 transformants), M3 (2 transformants), M4 (4 transformants) and M5 (6 transformants). It was not clear whether double mutant number M2 had been successful since the PCR band for A-Rcasssette was missing; although no bands were obtained in the negative control. However, repeated PCR analysis still did not provide a clear-cut answer as to whether or not gene deletion was successful for strain M2. Due to the low number of transformants for strains M1 and M3, and no transformants for Strain M2 for some isolates, more colonies from the second URA- selective plate were screened. As a result of this further screening, 4 transformants were isolated for strain M1 and 3 transformants each were obtained for strain M2 and M3.

7.3.2 Growth characteristics for double mutant strains

Table 7.4 summarises the doubling times and growth rates of CS and the 5 constructed S. cerivisiae double mutant strains in NC medium.

Table 7.4: **Doubling times and specific growth rates for double mutant strains.** Doubling times and specific growth rates for double mutant strains grown in 3XALL medium. Doubling times and Specific growth rates were calculated from the steepest parts of the growth curves.

Double Mutant	Doubling time (minutes)	Specfic growth rate
ΔKGD1ΔALT1	169	0.0041
ΔKGD2ΔALT1	182	0.0038
ΔLSC1ΔALT1	187	0.0037
ΔLSC2ΔALT1	177	0.0039
ΔALT1ΔGLT1	177	0.0039
CS	161	0.0043

The growth characteristics indicate that the yeast control strain (CS) has a higher growth rate and hence lower doubling time (0.0043 and 161 minutes respectively) than the yeast double mutants. The yeast double mutant, LSC1 X ALT1, showed the lowest growth rate of 0.0037 and the longest doubling time of 187.3 minutes. LSC2 X ALT1 and ALT1 X GLT1 grew at the same rate of 0.0039 and exhibited the same doubling time of 177 minutes.

7.3.3 GC/MS results for lysine by double mutants

Figure 7.5 and Table 7.5 show the GC/MS intracellular measurements for lysine in double mutants. Table 7.5 shows the mean and standard deviation values for intracellular lysine in CS and double mutants.

Table 7.5: **Mean and standard deviation values for extracellular and intracellular lysine in CS and double mutant strains.** Mean and standard deviation values for extracellular secretion and intracellular accumulation of lysine by CS and double mutant strains during mid-log phase of growth. Mean values are in micromoles/litre. SD represents standard deviation.

	Extracellular mean (SD)	Intracellular mean (SD)
KGD1 x ALT1	141.7 (6.2)	129.4 (48.5)
KGD2 X ALT1	169.2 (2.7)	107.7 (45.8)
LSC1 X ALT1	151.6 (15.1)	150.2 (123.4)
LSC2 X ALT1	140.2 (4.4)	226.8 (87.4)
ALT1 X GLT1	158.6 (23.5)	295.7 (147.9)
CS	132.4 (8.9)	46.7 (10.1)

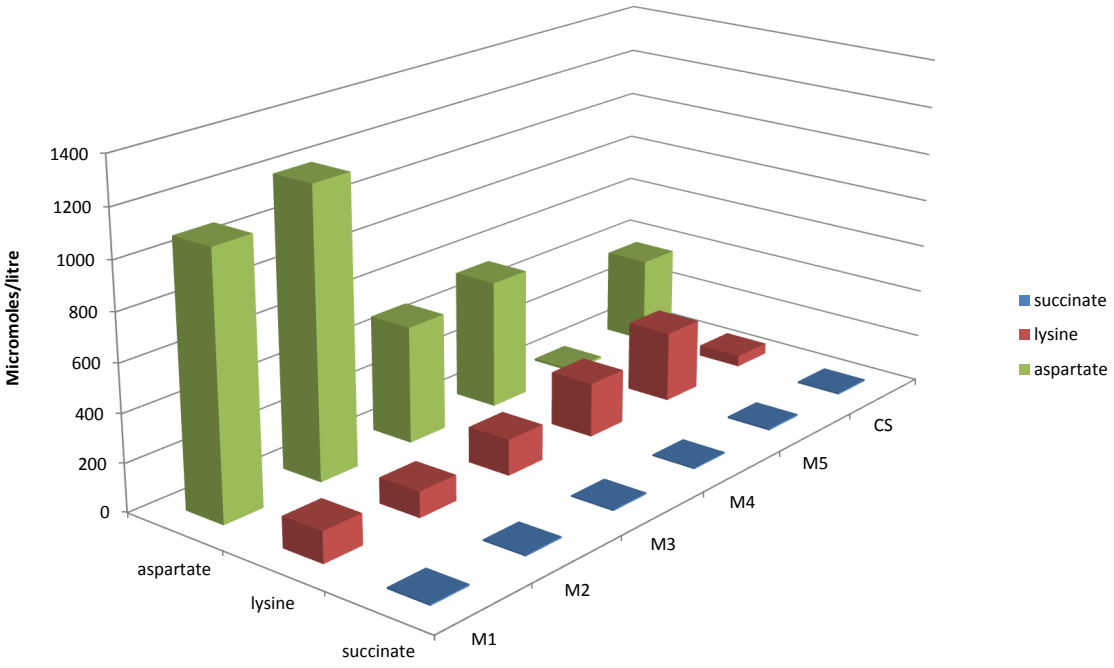

Figure 7.5: **Excretion of lysine by CS and 5 double mutant strains.** Figure shows excretion of lysine (and two other metabolites) by CS and 5 double mutant strains, into the medium during mid-log phase of growth.

The results of endometabolome measurements of the five double mutants showed

that two-fold increase in flux towards lysine production was demonstrated by *S. cerevisiae* double mutant Δ*kgd1*Δ*alt1* (M1), while both *S. cerevisiae* double mutants, Δlsc2Δalt1 (M4) and Δ*alt1*Δ*glt1* (M5) showed about four-fold increase in lysine production more than the control strain. Furthermore, the results showed that Δ*kgd1*Δ*alt1* and Δ*kgd2*Δ*alt1* produced three times more aspartate than CS, while all mutants except Δ*kgd2*Δ*alt1* produced one and a half times more glutamate than CS. Double mutant strains Δ*kgd1*Δ*alt1*, Δ*lsc1*Δ*alt1* and Δ*lsc2*Δ*alt1* also produced more than twice the amount of glycerol than CS, while Δ*alt1*Δ*glt1* produced about one and a half times more glycerol than CS. The concentrations of a number of metabolites (α-ketoglutarate, arginine, alanine, phenylalanine and fumaric acid) were found to be below the limits of detection by GC/MS analyses of the samples for double mutants and CS strains.

7.4 *S. cerevisiae* triple mutants

7.4.1 Construction of *S. cerevisiae* triple mutant

Amplification of the disruption cassettes from plasmid pREP41 EGFPC using disruption primers, ZWF1_disrupt_F and ZWF1_disrupt_R, for the disruption of target genes according to Figure 4.11 (in section 4.7.1) was carried out in order to create 3 different *S. cerevisiae* triple mutants. Figure 7.6 shows that a band of expected size 2391 bp was obtained.

7.4.1.1 Triple gene deletion transformants

Colonies of strains obtained on LEU- selection plate were checked for transformants after 7 days of transforming *S. cerevisiae* strains with plasmid pREP41 EGFPC (disruption cassette). Tables 7.6 show the number of colonies obtained for each of the strains heat-shocked for 2 , 15 and 20 minutes after transformation.

Table 7.6: **Triple mutant colonies.** Number of colonies for triple mutants

Triple mutant Required	Parent double mutant	No of colonies for different heat shock times		
		2 minutes	15 minutes	20 minutes
$\Delta kgd2\Delta alt1\Delta zwf1$	$\Delta kgd2\Delta alt1$	1	8	6
$\Delta lsc2\Delta alt1\Delta zwf1$	$\Delta lsc2\Delta alt1$	3	10	13
$\Delta alt1\Delta glt\Delta zwf1$)	$\Delta alt1\Delta glt$	4	6	16

However,the results of screening for transformants indicate low effieciency of transformation as only one transformant was obtained out of 70 colonies screened with colony PCR. Figure 7.6 shows the only successful gene deletion for the creation of a triple mutant, ΔAlt1Δlsc2Δzwf1, from this study.

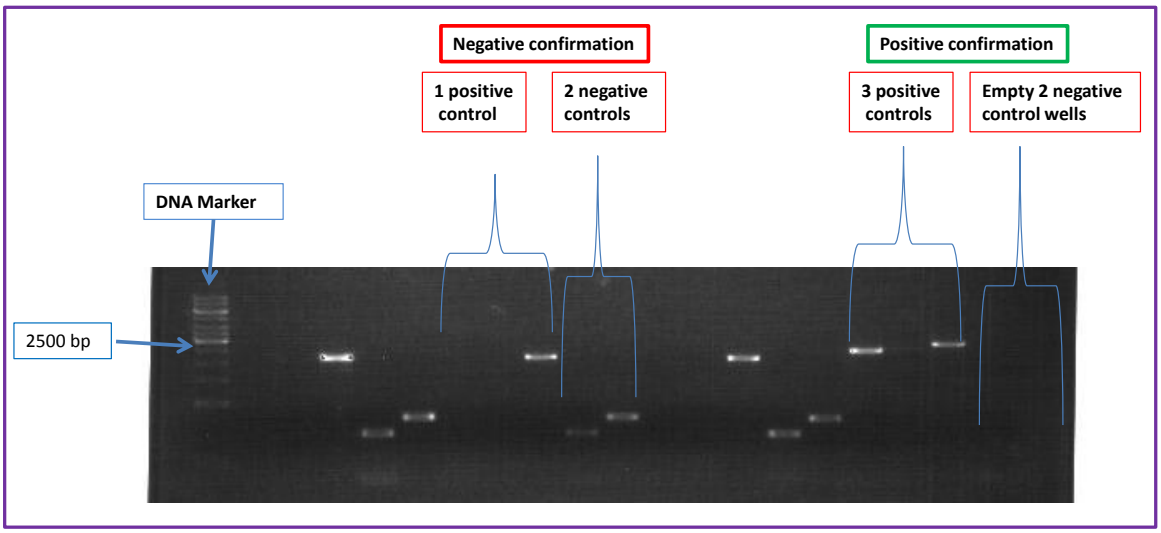

Figure 7.6: **PCR confirmation of gene deletion for triple mutants.** Figure shows the results of a positive (triple mutant $\Delta lsc2\Delta alt1\Delta zwf1$) and negative PCR confirmations of gene deletion for triple mutants. Gene deletion was successful as three bands of correct sizes (2500 bp, 2500 bp and 2700 bp) for positive controls and no band for negative controls were obtained. For the negative confirmation, 2 negative controls are positive and only 1 band of positive control visible.

Unfortunately, the triple mutant, $\Delta lsc2\Delta alt1\Delta zwf1$, showed very poor growth in 3XALL medium and could not be investigated further by GC-MS for metabolome

measurements. In addition, when disruption primers were changed to allow for auxotrophic selection of transformants on media lacking histidine instead of leucine, no triple mutant tranformant was isolated.

7.5 Discussion

In the genetic engineering experiments carried out for the studies reported here, disruption cassettes were generated by PCR-amplification of selectable markers. Plasmids that are non-replicating in yeast were used as template in order to prevent the possibility for the plasmid PCR template for the generation of disruption cassette being able to replicate autonomously in the transformed yeast cells, leading to a high number of transformants inheriting the plasmid instead of the disruption cassette. Long flanking sequences (greater than 60 base pairs) were used to increase the frequency of homologous recombination and also to enhance the efficiency of gene disruption. Rigorous PCR verification of a successful PCR-based gene deletion was carried out with three positive and 2 negative controls.

Single mutants *LSC2* and *GLT1* accumulated and excreted several folds higher amounts of lysine than CS. This finding was supported by the fact that the blank media, also analysed by GC-MS, contained an undetectable level of lysine. In addition, most of the metabolites measured did not show any concomitant increase in appreciable amounts, apart from single mutant *GLT1* which accumulated higher amounts of phenylalanine than CS. The fact that *LYS20/21* accumulated lysine intracellularly was also replicated in this experiment; it was also noted that *LYS20/LYS21* excreted a higher amount of phenylalanine compared with CS.

Importantly, *LYS20/21* does not excrete more lysine than the control. All of the increase in concentration of that metabolite is intracellular. The results of the metabolome measurement of the single mutants probably indicated a modest general effects of genetic perturbation, due to single deletions, in the metabolic pathways of *S. cereviasiae*. Although the gene deletions successfully redirected flux appreciably

in two out of the six single mutants, the modest effect on the general metabolism may point to system robsustness or gene dispensability. Possibly, the effects of the disrupted gene functions in each of the isoenzymes of LSC and KGD might have been buffered by the other isoenzyme still functioning, or by alternative pathways.

The interpretation of the results of the Toronto double mutant strain was rather difficult. Obviously, the effects of the multiple genetic perturbations were more pronounced than the for the single gene deletions. Conspicuously, the intracellular concentrations of 2-oxoglutarate, aspartate, fumaric acid and lysine were low, and the intracellular concentrations of serine and alanine were high, for all mutants. It appears logical that a decreased level of 2-oxoglutarate will be accompanied by low lysine yield (as 2-oxoglutarate is the first metabolite in the lysine biosynthetic pathway). Increased flux towards the production of alanine in all mutants was baffling in that if it is assumed that the deletion of ALT2 gene led to increased ALT1, thereby pushing up intracellular alanine concentrations in five of the six double mutants with increased alanine concentration, the same assumption does not hold for the increased alanine concentration in mutant $\Delta lsc1\Delta lsc2$. Hence, there are number of pertinent questions as follows:

1. Were the *in silico* double knockout predictions for lysine increase in *S. cerevisiae* wrong?

2. Did lysine feedback inhibition of LYS20/21 gene (the first enzyme in the lysine biosynthetic pathway) by increased amount of lysine in the medium occur at any stage?

3. Was the genotype background of the Toronto or incomplete LYS2 gene restoration the reason for these results?

4. Were the results simply due to robustness of *S. cerevisiae* metabolism in response to genetic perturbation?

5. Other reasons?

The successful in-house construction of five double mutants for lysine production helped to answer one or two of the above questions. It is possible for multiple-perturbations to provide a significantly richer and more biologically plausible functional annotation of the genes comprising the metabolic network of the yeast (Deutscher et al., 2006). The results from the five *S. cerevisiae* double mutants generated in-house demonstrated 2.2 fold, 3.8-fold and 4-fold increased production of lysine by $\Delta kgd2\Delta alt1$, $\Delta lsc2\Delta alt1$ and $\Delta alt1\Delta glt$ mutants respectively. These results supported the findings from studies involving single mutant strains for lysine production. Apart from the demonstrated flux redirection to lysine pathway in these three mutants strains, there is also evidence for the effects of multiple-perturbations. Increased intracellular concentrations of glutamate in four out of the five mutant strains may be explained by the deletion of *ALT1* results in accumulation of glutamate since glutamate can no longer be converted into alanine. This is reflected in four of the in-house created double mutants. In these four cases, the flux redirection might have favoured increased reaction between oxaloacetate and glutamate, leading to increased levels of both aspartate and 2-oxoglutarate. The increased level of aspartate from this reaction might explain the finding that $\Delta kgd1\Delta alt1$ and $\Delta kgd2\Delta alt1$ produced three times more aspartate than CS. 2-oxoglutarate is the first metabolite in the lysine biosynthetic pathway and hence increased glutamate is beneficial to increased flux to lysine biosynthetic pathway (Figure 8.15). On the other hand, all Toronto strains produced more alanine than CS, which is the opposite of what was expected. It is unclear whether or not the genotype background or the incomplete LYS2 gene restoration of the Toronto strains were the reasons for the unexplained increase in alanine. Hence, it is pertinent to conclude here that the available results do not provide sufficient evidence, either directly or indirectly, to enable the inference of a logical conclusion for increased alanine in all the Toronto yeast double mutants.

Since flux redirection might have resulted in accumulation of oxaloacetate and some of which might have been converted back to pyruvate and consequently increased level of pyruvate might have inhibited the conversion of glyceraldyde-3-phosphate

to 3-phosphoglycerate; increased glyceraldyde-3-phosphate level may then favour its conversion into dihydroxyacetone phosphate by TPI and eventually to glycerol. LSC2 is a beta subunit of *succinyl-CoA ligase*, which is a mitochondrial enzyme of the TCA cycle that catalyses the nucleotide-dependent conversion of succinyl-CoA to succinate. The deletion of LSC2 succeeded in disrupting the TCA cycle and hence redirected flux in the direction of lysine pathway via 2-oxoglutarate. ALT1 is involved in alanine biosynthesis. Disruption of *ALT1* gene in addition to *LSC2* gene ensured the availability of the much needed glutamate which could have been used instead for alanine production.

Vacuolar transport of lysine in yeast has "in" direction and hence the assumption that yeast accumulates lysine intracellularly may have been supported by the results that most of the double mutant strains did not excrete significant amounts of lysine into the culture medium.

The only successful triple mutant in this study failed to grow in the minimal medium used for the other strains. The construction of triple mutants from the previously generated double mutants required the deletion of ZWF1, which encodes the enzyme Glucose-6-phosphate dehydrogenase (G6PD), catalysing the first step of the pentose phosphate pathway. The deletion of ZWF1 would have ensured the redirection of flux form the pentose phosphate pathway to glycolysis and to the remaining part of TCA cycle uninterupted by the other two deletions. However, the failure of the only triple mutant generated to grow appreciably in minimal medium may point to synthetic lethality of ZWF1 on the other genes or the known toxic effects of lysine accumulation and the intermediate metabolites (such as α-aminoadipate semialdehyde in the lysine biosynthetic pathway.

7.6 Conclusion

The results from the studies reported in this chapter have demonstrated the usefulness of the application of modelling in realising the desired effects in the metabolic

engineering of *S. cerevisiae* for production of lysine. Enhanced production of lysine in two out of six *S. cerevisiae* single mutants and three out of five *S. cerevisiae* double mutants was successfully validated by experiments. However, metabolic regulation imposed by the feedback inhibition of lysine and systems-wide metabolic interactions are among the bottlenecks to be removed for further improvement of the lysine producing strains reported in this chapter. In this regard, integration of "omics" such as metabolic profiling, transcriptomics and proteomics with metabolic engineering approaches will be highly useful.

Chapter 8

Characterisation of *S. cerevisiae* production strains

8.1 Introduction

In a previous experiment carried out in this project (section 7.3), two-fold increase in flux towards lysine production was demonstrated by *S. cerevisiae* double mutant Δkgd1Δalt1 (M1), while both *S. cerevisiae* double mutants, Δlsc2Δalt1 (M4) and Δalt1Δglt1 (M5) showed four-fold increase in lysine production relative to the control strain. In order to optimise these lysine producing strains further, it was decided to carry out metabolic profiling of the three double mutant strains so as to link the genetic effects of knockouts to the metabolome. The control strain and the remaining two *S. cerevisiae* double mutants, Δkgd2Δalt1 (M2) and Δlsc1Δalt1 (M3) which did not produce an appreciable amounts of lysine than the control strain were included in the metabolic profiling experiment for comparisons.

Hence, the objectives of the work in this chapter are:

1. To investigate the metabolite profiles of the five mutant strains so as to unravel patterns related to the different genetic perturbations in the mutants.

2. To gain understanding of how cellular fluxes in knock-out strains and control strain

are different.

3. To characterise the lysine mutant strains and gain information that may help in their lysine producing capacity.

8.2 Materials and Methods

8.2.1 Experiment design

A statistical test is unlikely to detect a true difference when the sample size is too small in comparison with the magnitude of the difference. The probability of rejecting a false hypothesis is power, avoiding the Type II error or a false negative decision. False negative rate (β) decreases with increase in power, and hence power is equal to 1 - β (i.e., sensitivity). Power analysis can be used to calculate the minimum sample size required so that a specified difference can be detected. It is generally accepted to use $\beta = 0.8$ or 0.9. In this analysis, $\beta = 0.9$ was used. Factors influencing power are: (1) The statistical significance criterion, (2) The magnitude of the effect of interest in the population and (3) Sample size used to detect the effect.

The Problem statement was formulated as follows:

Power analysis for two-sample t test was considered in this experiment design to determine the sample size (replicate samples), that with a power of 90%, using a two-sided test at the level of 0.05%, can detect a difference in mean concentration of metabolite (m) in a distribution with a 2 standard deviation produced by yeast mutant and control strains. Since *S. cerevisiae* mutants are expected to produce more or less of metabolites than the control strain, a two-sided test at the level of 0.05 was considered.

Null hypothesis: $H_0 : m = 0$ (no change in mean metabolite concentration)

Alternative hypothesis: $H_0 : m \neq 0$ (there is change in mean metabolite concentration).

Power analysis was carried out using the R statistical package to determine the sample size.

8.2.2 Metabolite profiling

8.2.2.1 Data pre-processing

Chemical identification of chromatographic peaks was performed by comparison of the retention index and mass spectrum of each chromatographic peak to those present in mass spectral libraries. A definitive identification was compassigned when the retention index (+/-10) and mass spectrum ($match > 70\%$) of the chromatographic peak were matched to those present in the Manchester Metabolomics Database (MMD) electron impact mass spectral library (Brown et al., 2009). A putative identification was assigned if the mass spectrum ($match > 70\%$) matched that of a mass spectrum in the Golm metabolome Database (Kopka et al., 2005) or the NIST/EPA/NIH05 EI mass spectral library. An assignment of unidentified was provided if the mass spectrum was not matched to any mass spectrum in any of the three defined libraries above with a $match > 70\%$.

Figure 8.1 depicts the steps involved in the metabolic profiling of the five *S. cerevisiae* double mutants.

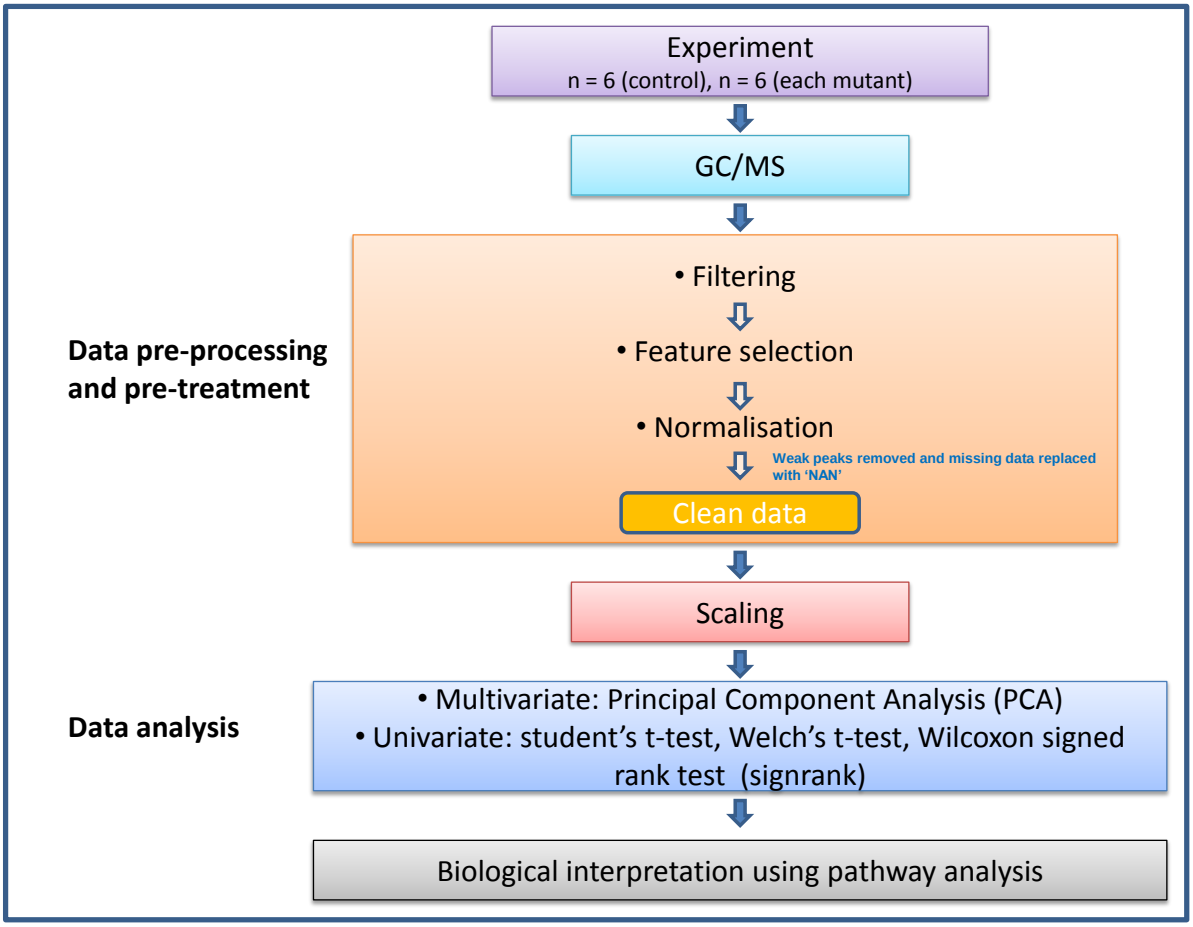

Figure 8.1: **Steps involved in the metabolic profiling**. Figure depicts the steps involved in the metabolic profiling of strains. Biological experiments yield extracts which are analysed by GC/MS. Next, data from GC/MS are pre-processed and pre-treated. Clean and scaled data were then analysed using a combination of multivariate and univariate statistical analysis techniques. Finally, biological meanings are inferred from the analysed data using pathway analysis.

8.2.2.2 Data pre-treatment

Data pretreatment was carried out in Excel and Matlab. The data were presented as transposed data matrices **X** with i rows of experiments and j columns of metabolites, and hence element x_{ij} is the measurement of metabolite j in experiment i. Several sets of data matrices were prepared from the raw data and the normalised (normalised to internal standard) data with one or all of the following treatments:

1. Zeros in data replaced with NAN - this was for univariate analysis.

2. zeros in data replaced with very small number (0.000001).

3. Column data (metabolite) with too many weak points removed (clean data for further pre-treatment).

4. Data matrix scaled with either centering scaling, auto scaling or pareto scaling method.

Since some of the column data contained many missing values, fourteen columns were removed, thereby reducing the number of metabolites from 144 to 130. Hence, the clean data was a pre-treated (steps 1 - 3 above) and reduced dataset consisting of 5 experiments (6 replicates each) and 130 variables (metabolites). However, as the entire dataset also included 12 replicates of quality control (QC) samples, the clean data consisted of 42 rows of experiment and 130 columns of variables (metabolites). Comparisons of four scaling methods (Center, Pareto, Autoscaling and Range: unsupervised cluster analysis was continued with pareto scaling) were carried out to find out the most suitable scaling method. To identify the source of greatest variation within the combined and individual groups of data for all samples multivariate analysis was employed.

8.2.2.3 Data analysis

Unsupervised exploratory data analysis: PCA was employed to discover any natural groups within the data and also used for discovering any outliers before preprocessing. Matlab (using prcomp function) was used to analyse the preprocessed data. PCA decomposes the variation in matrix X into scores S, loadings L. S is a I X A matrix containing the scores and P is a J X A matrix containing the A selected loadings. The PCs to investigate further were determined by inspecting the scree plots of scores from the initial PCA and the scores were then plotted in two-dimensions. Loadings plot of the dominant PC were examined to determine which metabolites were extreme, that is far away from zero based on the cut off of 0.05. From the loadings plot, the variable peaks on the extremes were mainly responsible

for the separation exhibited in the scores plot whereas those close to the origin had little or no contribution to such separation. Examination of the loading plot revealed peaks with the highest variability. This was used to compare relative metabolite concentrations in samples and whether peaks (metabolite IDs) from loading plots correspond with those from univariate analysis.

Loadings were labelled with "variable ID" numbers 1 - 130, and the variable ID number of any metabolite with relative concentrations greater than +0.05 or lesser than -0.05 were displayed on the plot.

PCA was carried out on scaled and unscaled data matrices (described in section 8.2.2.2 according to the following steps:

1. PCA was carried out on sample data matrices.
2. Scree plot of components was inspected in order to determine the number of PCs (eigenvalues) to investigate further.
3. Two-dimensional plots of PCs with maximum amount of variance was performed.
4. Occurrence of tight clustering of QCs in the two-dimensional plots of PCs were noted.
5. QCs were removed from sample data matrices, and the first three steps were repeated.
6. Clustering of samples was in the two-dimensional plots of PCs that captured maximal variance was inspected.
7. Outliers, if any, were removed and PCA was performed again. This was also done during preprocessing steps.
8. Where good clustering occurred in the samples (in score plots e.g PC1 vs. PC2), loadings of PCs with the most variances were plotted .

Univariate hypothesis testing: For univariate hypothesis testing, Matlab functions were used to perform student's t-test and signed rank test on the clean data matrices. A paired, two-sided student's t-test under the assumption of equal population variances at a significance level of 5% and a paired, two-sided signed rank test at a significance level of 5% were implemented.

Boxplots: Boxplots of CS against mutants M2, M3, M4 and M5 were also created on clean data to display the relative concentrations of the statistically significant variables (metabolites) according to signrank test. The boxplot display was checked to see if the statistically significant metabolites match the metabolites with major contributions to variances as determined by PCA.

8.3 Results for metabolic profiling

8.3.1 GC-MS results

Metabolic profiling of intracellular metabolomes of four mutant samples (M2, M3, M4 and M5) and CS carried out by GC-MS provided a raw data comprising of 144 metabolite peaks for each of the 42 experiments, out of which 130 metabolites were kept as clean data. Figures 8.2 and 8.3 show examples of total ion current (TIC) in for QC samples 4 - 8 and experiment samples 8 - 12.

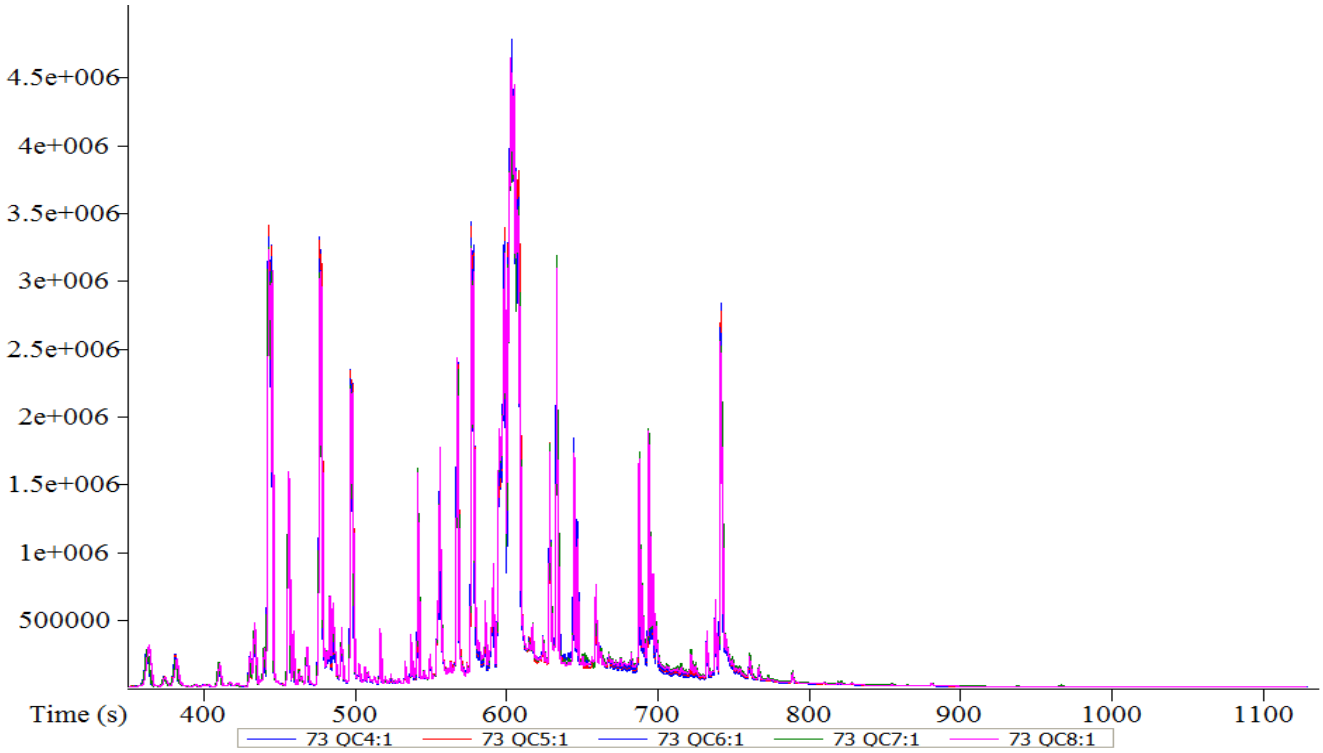

Figure 8.2: **TIC for QC samples 4 - 8**. Figure shows TIC for QC samples 4 - 8, indicating good reproducibility.

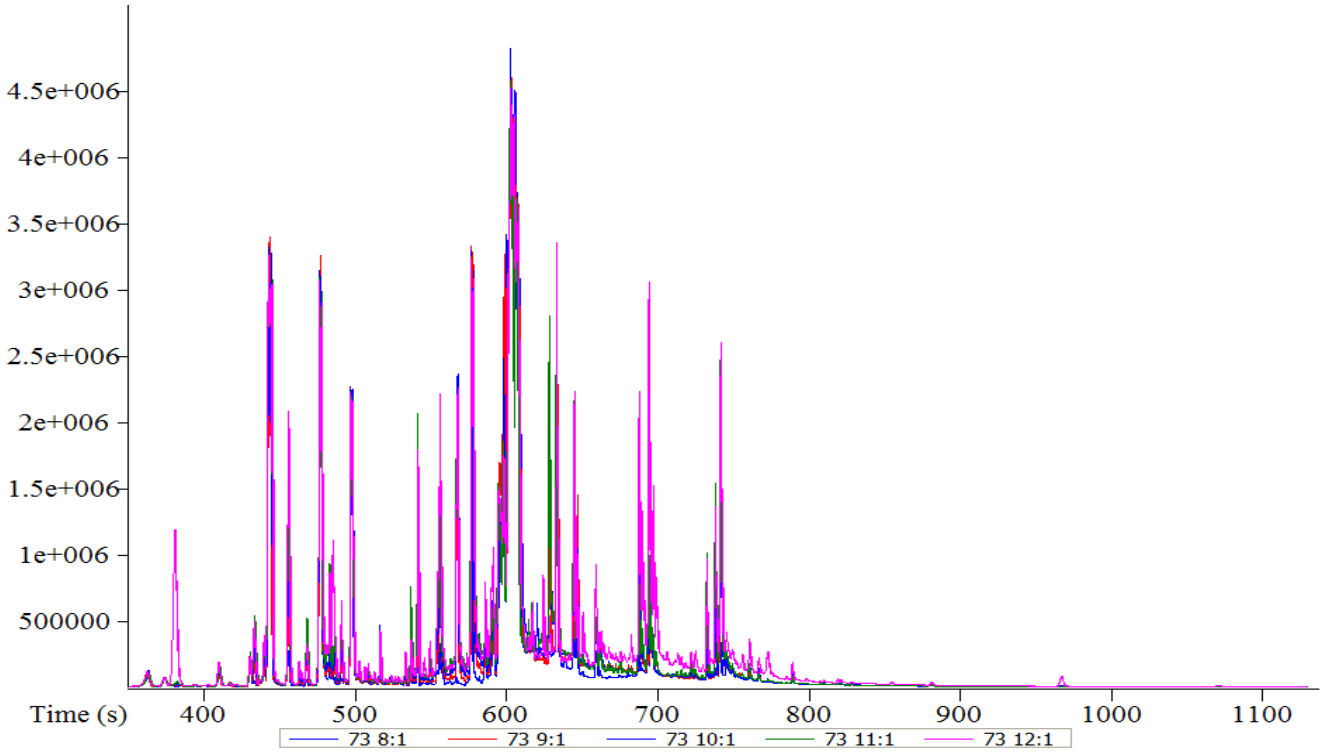

Figure 8.3: **TIC for Samples 8-12**. Figure shows a complete mass 73 chromatogram for Samples 8-12

Tables 8.1, 8.2 and 8.3 show the all the 81 identified metabolite peaks (out of a total of 130 identifed and unidentified metabolite peaks in the clean data) from GC-MS and the biochemical pathways with which that they are associated.

Table 8.1: **Identified chromatographic peaks and pathways: table 1 of 3**. Table shows thirty of eighty-one metabolite peaks identified and the biochemical pathways they are associated with. A "definitive" identification was a match of the chromatographic peak to the Manchester Metabolomics Database (MMD) electron impact mass spectral library and a "putative" identification was a match of mass spectrum to that of a mass spectrum in the Golm metabolome Database or the MIST/EPA/NIH05 EI mass spectral library.

Metabolite	ChEBI ID	Definitive/Putative	Biochemical Pathway
Histidine	CHEBI:27570	Definitive	Amino acid biosyhnthesis
Tryptophan	CHEBI:27897	Definitive	
Isoleucine	CHEBI:24898	Definitive	
Leucine	CHEBI:25017	Definitive	
Leucine	CHEBI:25017	Definitive	
Alanine	CHEBI:16449	Definitive	
Isoleucine	CHEBI:24898	Definitive	
Proline	CHEBI:26271	Definitive	
Glycine	CHEBI:15428	Definitive	
Serine	CHEBI:17822	Definitive	
Proline	CHEBI:26271	Definitive	
Threonine	CHEBI:26986	Definitive	
Glycine	CHEBI:15428	Definitive	
Serine	CHEBI:17822	Definitive	
Threonine	CHEBI:26986	Definitive	
Homoserine	CHEBI:30653	Putative	
Valine	CHEBI:27266	Definitive	
Homoserine	CHEBI:30653	Putative	
Aspartic acid	CHEBI:22660	Definitive	
Aspartic acid	CHEBI:22660	Definitive	
Glutamic acid	CHEBI:18237	Definitive	
Methionine	CHEBI:16811	Definitive	
Glutamine	CHEBI:28300	Definitive	
Cysteine	CHEBI:15356	Definitive	
Methionine	CHEBI:16811	Definitive	
N-formylmethionine or methionine	CHEBI:16552	Definitive	
N-formylmethionine or methionine	CHEBI:16552	Putative	
Pyroglutamic acid and/or glutamic acid	CHEBI:16010	Putative	
Homocysteine	CHEBI:17230	Definitive	
Phenylalanine	CHEBI:28044	Definitive	

Table 8.2: **Identified chromatographic peaks and pathways: table 2 of 3**. Table shows thirty-one of eighty-one metabolite peaks identified and the biochemical pathways they are associated with. A "definitive" identification was a match of the chromatographic peak to the Manchester Metabolomics Database (MMD) electron impact mass spectral library and a "putative" identification was a match of mass spectrum to that of a mass spectrum in the Golm metabolome Database or the MIST/EPA/NIH05 EI mass spectral library.

Metabolite	ChEBI ID	Definitive/Putative	Biochemical Pathway
Asparagine	CHEBI:22653	Definitive	Amino acid biosyhnthesis
Valine	CHEBI:27266	Definitive	
Phenylalanine	CHEBI:28044	Definitive	
Glutamine	CHEBI:28300	Definitive	
Valine	CHEBI:27266	Definitive	
Lysine	CHEBI:25094	Definitive	
Sugar	CHEBI:16646	Putative	Carbohydrate metabolism
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Trehalose	CHEBI:27082	Definitive	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Sugar	CHEBI:16646	Putative	
Hexadecanoic acid	CHEBI:15756	Definitive	Fatty Acid Metabolism
Hexadecenoic acid	CHEBI:24548	Definitive	
Octadecanoic acid	CHEBI:28842	Definitive	
Octadecenoic acid	CHEBI:25634	Definitive	
Octadecenoic acid	CHEBI:25634	Definitive	

Table 8.3: **Identified chromatographic peaks and pathways: table 3 of 3**. Table shows thirty-one of eighty-one metabolite peaks identified and the biochemical pathways they are associated with. A "definitive" identification was a match of the chromatographic peak to the Manchester Metabolomics Database (MMD) electron impact mass spectral library and a "putative" identification was a match of mass spectrum to that of a mass spectrum in the Golm metabolome Database or the MIST/EPA/NIH05 EI mass spectral library.

Metabolite	ChEBI ID	Definitive/Putative	Biochemical Pathway
Butyrolactone	CHEBI:42639	Putative	Fatty Acid Metabolism
Glycerol	CHEBI:17754	Definitive	Glycerolipid Metabolism
Glycerol	CHEBI:17754	Definitive	
Glycerol	CHEBI:17754	Definitive	
Fructose-6-phosphate	CHEBI:15946	Definitive	Glycolysis pathway
Fructose-6-phosphate	CHEBI:15946	Definitive	
Glucose-6-phosphate	CHEBI:17719	Definitive	
Glucose-6-phosphate	CHEBI:17719	Definitive	
Glycerol-3-phosphate	CHEBI:15978	Definitive	
Lactic acid	CHEBI:28358	Definitive	Fermentation pathway
AMP	CHEBI:16027	Putative	Metabolism of Cofactors and Vitamins
Phosphate, monmethyl ester	CHEBI:340824	Putative	
Phosphate	CHEBI:18367	Definitive	
Phosphate	CHEBI:18367	Definitive	
Nicotinamide	CHEBI:17154	Definitive	
Cystathionine	CHEBI:17755	Definitive	Amino acid metabolism
Cystathionine	CHEBI:17755	Definitive	
Cystathionine	CHEBI:17755	Definitive	
4-hydroxyproline	CHEBI:20392	Definitive	
2-aminobutanoic acid	CHEBI:35621	Putative	
Glutamine	CHEBI:28300	Definitive	
Pipecolic acid	CHEBI:17964	Putative	
Fumaric acid	CHEBI:18012	Definitive	Tricarboxylic acid cycle
Malic acid	CHEBI:6650	Definitive	
Citric acid	CHEBI:30769	Definitive	
Citrulline	CHEBI:18211	Definitive	Urea cycle
Ornithine	CHEBI:18257	definitive	
Citrulline	CHEBI:18211	Definitive	
Orotic acid	CHEBI:16742	Definitive	Purine and pyrimidine biosynthetic pathways
Uracil	CHEBI:17568	Definitive	
Adenine	CHEBI:16708	Definitive	

8.3.2 Multivariate analysis

8.3.2.1 Pretreatment effects of scaling methods

A scree plot of clean data is presented in Figure 8.4, indicating that the first 10 PCs contain 94% of all variances, and the first three PCs (23.0%, 17.7%, 16.9% respectively) account for 57.6% of all variances.

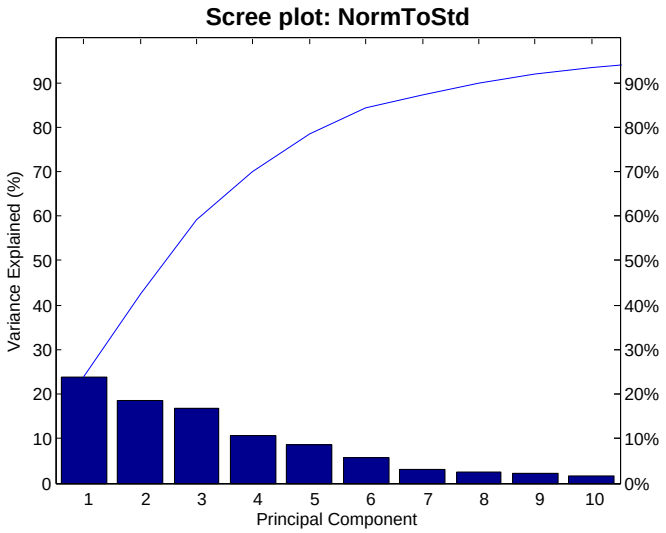

Figure 8.4: **Scree plot of clean data (including QCs)**. Figure depicts a scree plot of principal component scores from PCA analysis of clean data (including QCs)

When the two-dimensional plot of PC1 and PC2 was carried out (Figure 8.5),ten out of the twelve QCs clustered together while two the QCs were outliers. Five samples of M2, M3 and M4 are close together with one outlier, while four of sample CS are close together and two are outliers. In the case of M5, four of the samples are located in different positions.

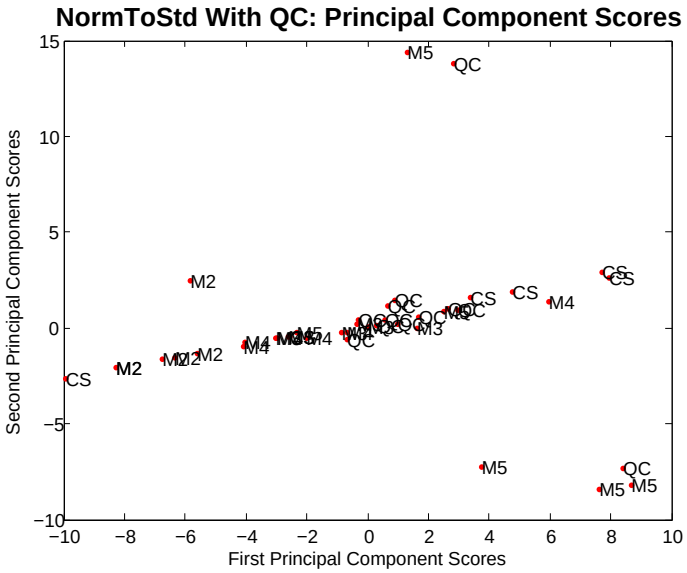

Figure 8.5: **PC1 against PC2: clean data with QCs**. Figure depicts a two-dimensional plot of PC1 (23.0%) against PC2 (17.7%) of clean data (including QCs)

The Pareto scaling method performed best out of the 4 scaling methods tested, and so was used for scaling purposes in the rest of the study. Figure 8.6 is a scree plot of scores from PCA analysis of Pareto scaled clean sample data (including QCs). As the figure shows, the first 10 PCs contain 87% of all variances, and the first four PCs (31.8205%, 12.8579%, 8.5743%, 8.0162% respectively)account for 61.3% of all variances.

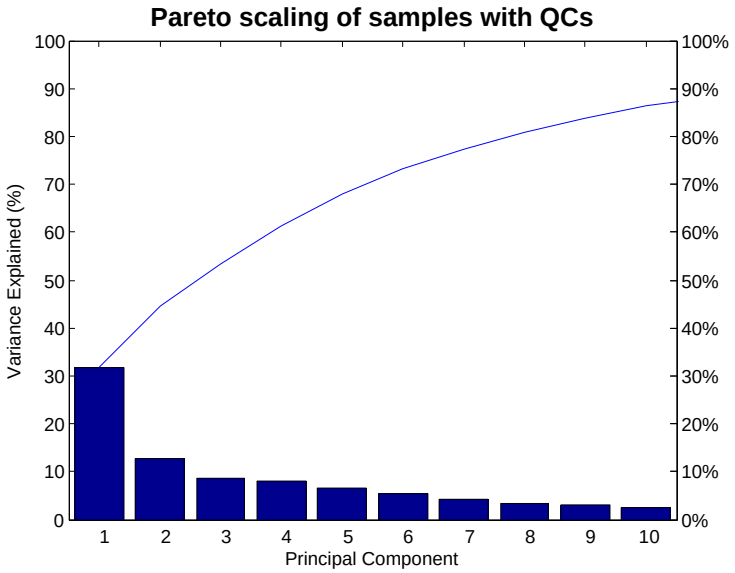

Figure 8.6: **Pareto scaled samples with QCs**. Figure depicts a scree plot of principal component scores from PCA analysis of Pareto scaled samples (including QCs)

The two-dimensional plot of first two PCs (PC1 and PC2) indicate fairly good clustering of all QC samples and fairly good separation for all samples except for four outliers: two samples of CS outliers, one sample of M5 and one sample of M4 (Figure 8.7).

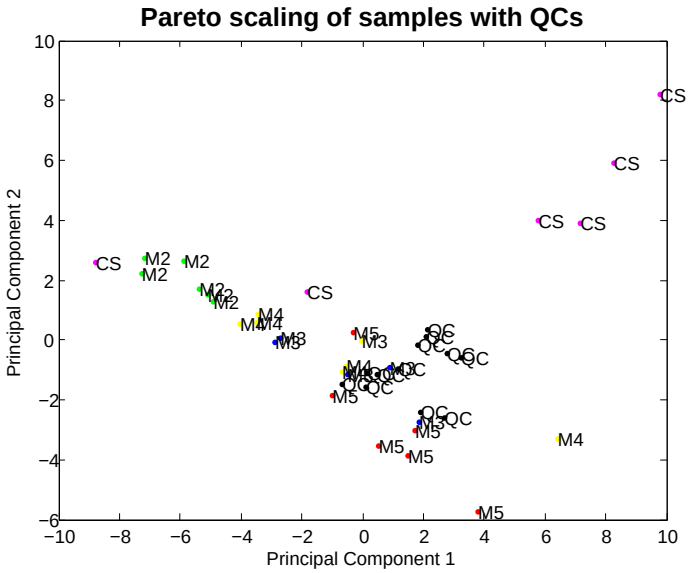

Figure 8.7: **PC1 against PC2: Pareto scaled samples with QCs**. Figure depicts a a two-dimensional plot of PC1 against PC2 of Pareto scaled samples (including QCs)

8.3.2.2 PCA analysis of scaled sample data (without QCs)

PCA of scaled sample data (without QCs) for CS and M5 revealed that six PCs accounted for 97% of all variances in the data and the first four PCs contained 90.9% of all variances (70.1%, 10.1%, 5.6% and 5.0% respectively). These results prompted further investigations of the first three dominant components, leading to plots of PC1 against PC2, PC1 against PC3, and PC2 against PC3. PC4 was not investigated further since it contains about 5% of all variance.

The 2-dimensional plots of PC1 against PC2 and PC1 against PC3 for CS and M2 (Figures 8.8 and 8.9 respectively) indicated a good separation in PC1 between the two samples. The plot of PC2 against PC3 for CS and M2, however, did not reveal any separation between CS and M5 either in PC2 or PC3.

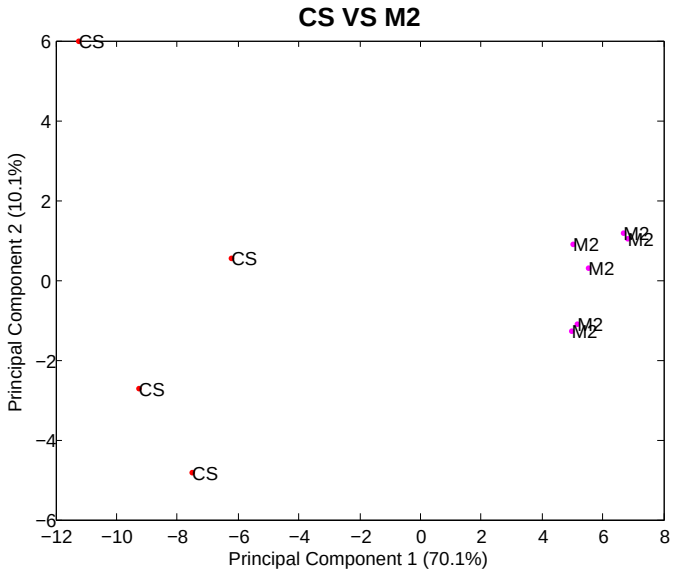

Figure 8.8: **A plot of PC1 against PC2 for CS and M2**. Figure shows a two-dimensional plot of PC1 against PC2 for CS vs M2

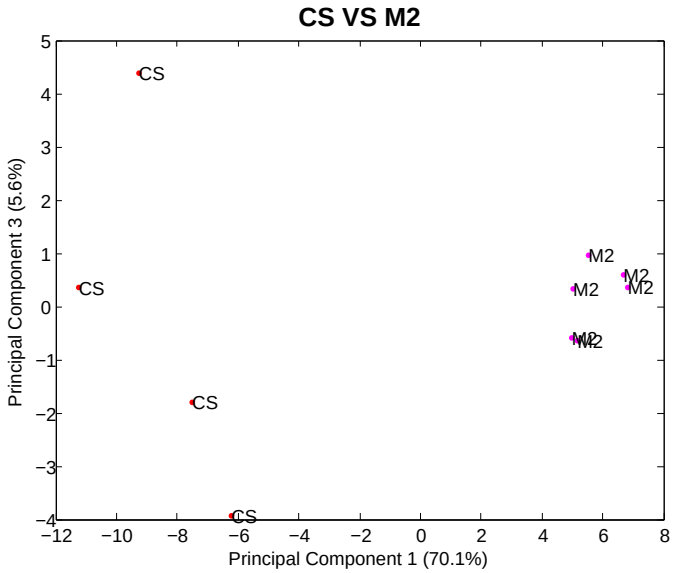

Figure 8.9: **A plot of PC1 against PC2 for CS and M2**. Figure shows a two-dimensional plot of PC1 against PC3 for CS and M2

As indicated in Figure 8.10 and Table 8.4, a plot of loadings from PC1 for CS and M2 indicated that fifty-four metabolites had major contributions to the variance. Three of these showed very positive loadings (above +0.05) and fifty-one had very negative

loadings (below -0.05). Furthermore, the three very positive loadings consisted of contributions from one identified and two unknown metabolites, while the very negative loadings are contributions from thirty-six identified and fifteen unknown metabolites.

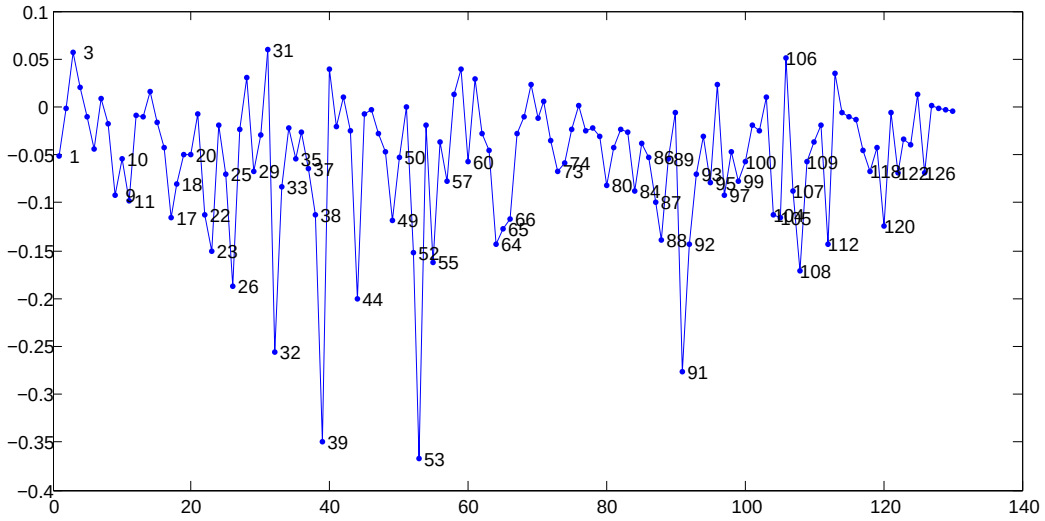

Figure 8.10: **A plot of loadings from PC1 for CS and M2**. Figure shows a plot of loadings from PC1 for CS and M2. Extreme metabolite relative concentrations outside the threshholds of +/- 0.05 are numbered.

Table 8.4: **A summary of loadings plots**. Table shows a summary of loadings plots.

	Contribution to variance		
Strain comparison	**Major**	**Positive**	**Negative**
CS versus M2	54	3	51
CS versus M3	49	6	43
CS versus M4	51	3	48
CS versus M5	50	14	36

For CS and M3, plots of PC1 against PC2 and PC1 against PC3 for (Figures C.4 and C.5, Appendix C) indicated a good separation in PC1 between the two samples. No separation was found between CS and M3 either in PC2 or PC3.

The loadings plot of PC1 for CS and M3 show that forty-nine metabolites had major contributions to the variances (Table 8.4). Six of these showed very positive

loadings (above +0.05) and forty-three had very negative loadings (below -0.05). The six very positive loadings consisted of contributions from four identified and two unknown metabolites, while the very negative loadings are contributions from twenty-nine identified and fourteen unknown metabolites (Table 8.4).

A good separation of CS and M4 was observed in PC1 the 2-dimensional plots of PC1 against PC2 and PC1 against PC3 as shown in Figures C.6 (Appendix C) and C.7 (Appendix C). However, no separation was found between CS and M5 either in PC2 or PC4.

In the case of PC1 for CS and M4, loadings plot (Table 8.4) revealed that fifty-one metabolites had major contributions to the variance. Three of which showed very positive loadings (above +0.05) and forty-eight had very negative loadings (below -0.05). Upon further analysis, the three very positive loadings consisted of contributions from two identified and one unknown metabolites, and the very negative loadings are contributions from thirty-two identified and sixteen unknown metabolites.

When the 2-dimensional plots of PC1 against PC2 and PC1 against PC3 for CS and M5 were carried out, a good separation along PC1 between the two samples was demonstrated in each case(Figures C.8 and C.9, Appendix C). However, the plot of PC2 against PC3 did not reveal any separation between CS and M5 either along PC2 or PC3.

As indicated in (Table 8.4), a plot of loadings from PC1 for CS and M5 indicated that fifty metabolites had major contributions to the variance. Fourteen of these showed very positive loadings (above +0.05) and thirty-six had very negative loadings (below -0.05). Furthermore, the fourteen very positive loadings consisted of contributions from eight identified and six unknown metabolites, and the very negative loadings are contributions from twenty-eight identified and eight unknown metabolites.

8.3.3 Results for hypothesis testing

Tables 8.5, 8.6 and 8.7 show the the p values of hypothesis testing of CS against each of M2, M3, M4 and M5 using Welch's t-test for each of the identified eighty-one metabolite peaks. Significance testing was carried out at the level of 5%.

Table 8.5: **Results of hypothesis testing: table 1 of 3**. Table shows results of hypothesis testing of CS against each of M2, M3, M4 and M5 using Welch's t-test for each of the identified eighty-one metabolite peaks. Null hypothesis was rejected at $p \leq 0.05$. The p-values in red colour indicate significant down regulation of specific metabolites.

		***P*-values from Welch's t-test**			
ID	**Metabolite (ChEBI ID)**	**CS and M2**	**CS and M3**	**CS and M4**	**CS and M5**
1	Sugar (CHEBI:16646)	0.75843	0.44033	0.62491	0.42626
2	Sugar (CHEBI:16646)	0.23960	0.58736	0.45167	0.30609
3	Sugar (CHEBI:16646)	0.12741	0.58061	0.11137	0.52908
4	Sugar (CHEBI:16646)	0.03920	0.84670	0.50578	0.57939
5	Sugar (CHEBI:16646)	0.52373	0.33198	0.41051	0.51620
6	Sugar (CHEBI:16646)	0.18388	0.31575	0.61955	0.23224
7	Histidine (CHEBI:27570)	0.02802	0.05455	0.04490	0.06603
8	Glycerol (CHEBI:17754)	0.10569	0.38231	0.26860	0.51453
9	Hexadecanoic acid (CHEBI:15756)	0.51293	0.63773	0.79726	0.77953
10	Hexadecenoic acid (CHEBI:24548)	0.13374	0.12743	0.14691	0.18915
11	Adenine (CHEBI:16708)	0.01192	0.05650	0.09060	0.04812
12	Sugar (CHEBI:16646)	0.01215	0.29894	0.49966	0.99606
13	Cystathionine (CHEBI:17755)	0.01606	0.04775	0.04589	0.05386
14	Fructose-6-phosphate (CHEBI:15946)	0.31181	0.80076	0.85234	0.72509
15	Fructose-6-phosphate (CHEBI:15946)	0.00730	0.05739	0.09664	0.05452
16	Glucose-6-phosphate (CHEBI:17719)	0.00835	0.02888	0.03307	0.03511
17	Octadecanoic acid (CHEBI:28842)	0.04323	0.51162	0.23884	0.33627
18	Octadecenoic acid (CHEBI:25634)	0.13043	0.35900	0.50397	0.67855
19	Glucose-6-phosphate (CHEBI:17719)	0.00947	0.02620	0.03104	0.03060
20	Cystathionine (CHEBI:17755)	0.00679	0.02235	0.02112	0.02487
21	Cystathionine (CHEBI:17755)	0.00689	0.01975	0.02103	0.02123
22	Tryptophan (CHEBI:27897)	0.02136	0.07710	0.07686	0.11606
23	Trehalose (CHEBI:27082)	0.01700	0.41813	0.23358	0.54493
24	Sugar (CHEBI:16646)	0.07380	0.09325	0.13093	0.52442
25	Glycerol (CHEBI:17754)	0.96624	0.41431	0.34627	0.25125
26	Sugar (CHEBI:16646)	0.17737	0.27284	0.66852	0.99990
27	Sugar (CHEBI:16646)	0.02570	0.08308	0.14814	0.10888
28	Sugar (CHEBI:16646)	0.95504	0.38871	0.87068	0.23085
29	AMP (CHEBI:16027)	0.06367	0.13912	0.11066	0.08169
30	Isoleucine (CHEBI:24898)	0.07660	0.53714	0.46361	0.97981

Table 8.6: **Results of hypothesis testing: table 2 of 3**. Table shows results of hypothesis testing of CS against each of M2, M3, M4 and M5 using Welch's t-test for each of the identified eighty-three metabolite peaks. Null hypothesis was rejected at $p \leq 0.05$. The p-values in red colour indicate significant down regulation of specific metabolites while the p values in green represent significant up regulation of specific metabolites.

		***P*-values from Welch's t-test**			
ID	**Metabolite (ChEBI ID)**	**CS and M2**	**CS and M3**	**CS and M4**	**CS and M5**
31	Leucine (CHEBI:25017)	0.01316	0.12358	0.11335	0.31773
32	Leucine (CHEBI:25017)	0.00177	0.22341	0.48075	0.28199
33	Phosphate, monmethyl ester (CHEBI:340824)	0.02055	0.44910	0.41732	0.54983
34	Alanine (CHEBI:28044)	0.01660	0.01710	0.01682	0.01725
35	Isoleucine (CHEBI:24898)	0.00040	0.00988	0.03658	0.05855
36	Proline (CHEBI:26271)	0.00338	0.82690	0.78559	0.17695
37	Glycine (CHEBI:15428)	0.04485	0.54777	0.54688	0.13375
38	Phosphate (CHEBI:18367)	0.01436	0.41227	0.20452	0.43020
39	Serine (CHEBI:17822)	0.01494	0.90885	0.96027	0.49771
40	Phosphate (CHEBI:18367)	0.00650	0.41360	0.44891	0.64110
41	Threonine (CHEBI:26986)	0.02038	0.98645	0.95545	0.38711
42	Glycine (CHEBI:15428)	0.05613	0.26878	0.34684	0.43021
43	Serine (CHEBI:17822)	0.01492	0.15412	0.19925	0.44873
44	Threonine (CHEBI:26986)	0.01474	0.15453	0.20901	0.45364
45	Fumaric acid (CHEBI:18012)	0.50507	0.81594	0.90912	0.85000
46	Homoserine (CHEBI:30653)	0.07682	0.92771	0.82011	0.64727
47	Valine (CHEBI:27266)	0.57050	0.32687	0.29346	0.23632
48	Homoserine (CHEBI:30653)	0.01769	0.10049	0.12202	0.24472
49	Uracil (CHEBI:17568)	0.42557	0.44836	0.71323	0.32800
50	4-hydroxyproline (CHEBI:20392)	0.49892	0.39288	0.72940	0.97363
51	Malic acid (CHEBI:6650)	0.03682	0.07956	0.14195	0.12790
52	Aspartic acid (CHEBI:22660)	0.25834	0.05422	0.17275	0.02480
53	2-aminobutanoic acid (CHEBI:35621)	0.01694	0.41656	0.14236	0.75372
54	Aspartic acid (CHEBI:22660)	0.02471	0.88932	0.92945	0.52912
55	Glutamic acid (CHEBI:18237)	0.60270	0.24804	0.25519	0.07916
56	Methionine (CHEBI:16811)	0.02130	0.18375	0.10750	0.36251
57	Cysteine (CHEBI:15356)	0.01010	0.64118	0.53814	0.85908
58	Methionine (CHEBI:16811)	0.04809	0.19975	0.23895	0.41136
59	Butyrolactone (CHEBI:42639)	0.05599	0.01892	0.06873	0.08287
60	Citrulline (CHEBI:18211)	0.24203	0.40603	0.38407	0.33803

Table 8.7: **Results of hypothesis testing: table 3 of 3**. Table shows results of hypothesis testing of CS against each of M2, M3, M4 and M5 using Welch's t-test for each of the identified eighty-three metabolite peaks. Null hypothesis was rejected at $p \leq 0.05$. The p-values in red colour indicate significant down regulation of specific metabolites while the p values in green represent significant up regulation of specific metabolites.

		***P*-values from Welch's t-test**			
ID	**Metabolite (ChEBI ID)**	**CS and M2**	**CS and M3**	**CS and M4**	**CS and M5**
61	Glutamine (CHEBI:28300)	0.02654	0.42981	0.95861	0.78049
62	Glutamine (CHEBI:28300)	0.01221	0.78888	0.75612	0.78607
63	Pyroglutamic acid and/or glutamic acid (CHEBI:16010)	0.30708	0.21004	0.97287	0.39827
64	Homocysteine (CHEBI:17230)	0.01181	0.03964	0.04383	0.06004
65	Phenylalanine (CHEBI:28044)	0.03067	0.24075	0.24550	0.44887
66	Nicotinamide (CHEBI:17154)	0.05031	0.15123	0.14197	0.16473
67	Sugar (CHEBI:16646)	0.76202	0.60660	0.78113	0.26007
68	Sugar (CHEBI:16646)	0.32091	0.62252	0.63772	0.75111
69	Ornithine (CHEBI:18257)	0.01040	0.11261	0.13396	0.02358
70	Glycerol-3-phosphate (CHEBI:15978)	0.29215	0.82260	0.54049	0.36542
71	Sugar (CHEBI:16646)	0.38373	0.76727	0.52562	0.31651
72	Pipecolic acid (CHEBI:17964)	0.67160	0.37422	0.26948	0.09050
73	Citric acid (CHEBI:30769)	0.38525	0.71677	0.71546	0.60617
74	Glutamine (CHEBI:28300)	0.04072	0.44177	0.73059	0.66974
75	Valine (CHEBI:27266)	0.02429	0.25878	0.26872	0.82197
76	Lysine (CHEBI:25094)	0.18507	0.56099	0.45140	0.98338
77	Citrulline (CHEBI:18211)	0.03801	0.43781	0.35775	0.84937
78	Sugar (CHEBI:16646)	0.95536	0.28025	0.59128	0.25478
79	Sugar (CHEBI:16646)	0.44553	0.28310	0.99648	0.01619
80	Sugar (CHEBI:16646)	0.03445	0.03995	0.03091	0.04813
81	Sugar (CHEBI:16646)	0.64929	0.33216	0.63729	0.32811

8.3.4 Comparing PCA with Univariate analysis

Matching the statistically significant metabolites (at 5% level using Welch's t-test) with the metabolites showing major contributions to variance in the PCA loadings, there were between 69 and 80% agreements between the results of CS against M2, CS against M3, CS against M4 and CS against M5, respectively. Using signrank p values in place of Welch's t-test for similar comaparisons, there were only 11 to 20% agreements between the results of CS against M2, CS against M3, CS against M4 and

CS against M5, respectively. Hence, p values for Welch's t-test were used for further investigations. Tables 8.8, 8.9, 8.10 and 8.11 show the lists of metabolites found to have major contributions to variance in PC1 loadings of CS against M2 (see section 8.3.2) and are also statistically significant (Welch's t-test) comparing CS against M2 (see section 8.3.3).

In the comparision of strains CS and M2, eleven metabolites out of the twenty-four which are statistically significant at 5% level matched the eleven major metabolites with extreme values identifed by PCA (Table 8.8). Similarly, for the comparisons of CS against M4 and CS against M5, twenty-three out of thirty-eight significant metabolites (Welch's t-test) matching the twenty-three metabolites with extreme values (PCA) and twenty-one out of forty-two significant metabolites (Welch's t-test) matching the twenty-one metabolites with extreme values (PCA), respectively (Tables 8.10 and 8.11)

Table 8.8: **Welch's *t*-tests against PCA results: CS and M2**. Table shows list of twenty-nine identified metabolite peaks found to have major contributions to variance in PC1 loadings of CS against M2 and are also statistically significant at 5% level using Welch's *t*-test to compare CS and M2.

ID	Metabolite
10	Histidine
18	Adenine
20	Sugar
23	Cystathionine
25	Fructose-6-phosphate
29	Glucose-6-phosphate
32	Cystathionine
33	Cystathionine
35	Tryptophan
44	Sugar
49	Leucine
50	Leucine
52	Phosphate, monmethyl ester
55	Proline
60	Serine
65	Serine
66	Threonine
74	Homoserine
80	Malic acid
84	2-aminobutanoic acid
91	Methionine
92	Cysteine
93	Methionine
97	Glutamine
105	Homocysteine
107	Phenylalanine
112	Ornithine
120	Valine
126	Sugar

Table 8.9: **Welch's t-tests against PCA results: CS and M3**. Table shows list of seven identified metabolite peaks found to have major contributions to variance in PC1 loadings of CS against M3 and are also statistically significant at 5% level using Welch's *t*-test to compare CS and M3.

ID	Metabolite
23	Cystathionine
26	Glucose-6-phosphate
29	Glucose-6-phosphate
33	Cystathionine
53	Alanine
105	Homocysteine
126	Sugar

Table 8.10: **Welch's t-tests against PCA results: CS and M4**. Table shows list of nine identified metabolite peaks found to have major contributions to variance in PC1 loadings of CS against M4 and are also statistically significant at 5% level using Welch's *t*-test to compare CS and M4.

ID	Metabolite
10	Histidine
23	Cystathionine
26	Glucose-6-phosphate
29	Glucose-6-phosphate
32	Cystathionine
33	Cystathionine
53	Alanine
105	Homocysteine
126	Sugar

Table 8.11: **Welch's t-tests against PCA results: CS and M5**. Table shows list of seven identified metabolite peaks found to have major contributions to variance in PC1 loadings of CS against M5 and are also statistically significant at 5% level using Welch's t-test to compare CS and M5.

ID	Metabolite
26	Glucose-6-phosphate
29	Glucose-6-phosphate
32	Cystathionine
33	Cystathionine
81	Aspartic acid
112	Ornithine
126	Sugar

Figures 8.11, 8.12, 8.13 and 8.14 are the box plots of CS versus the mutants strains M2, M3, M4 and M5, respectively, displaying the relative concentrations of the matching metabolites identified by both PCA and Welch's t-test.

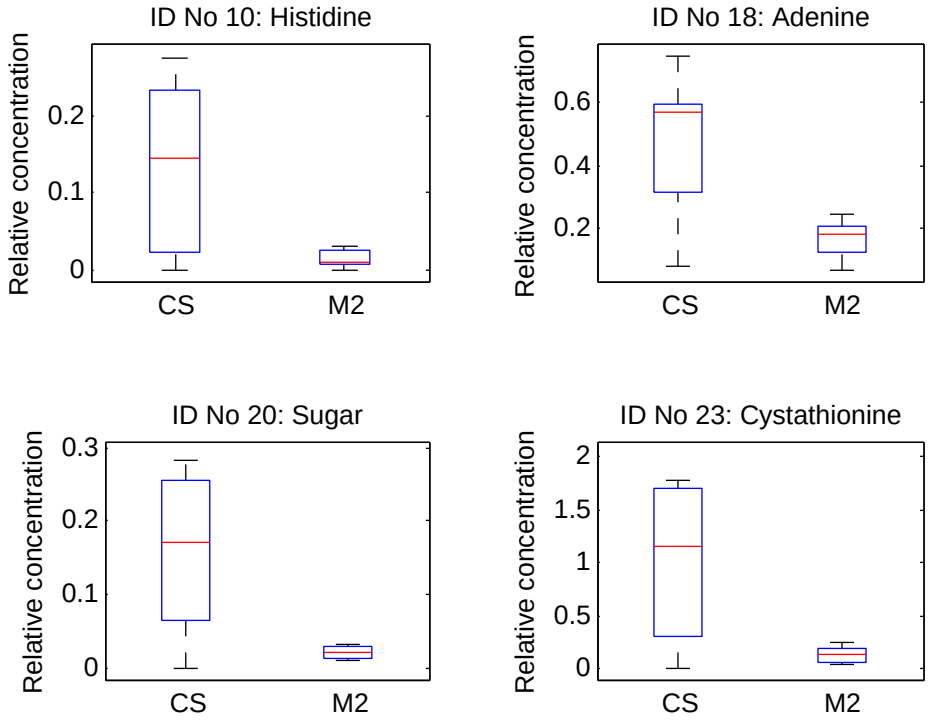

Figure 8.11: **Box plot metabolites: CS against M2**. Figure shows box plots of CS against M2 displaying the relative concentrations of four of the matching metabolites: histidine, adenine, sugar and cystathionine.

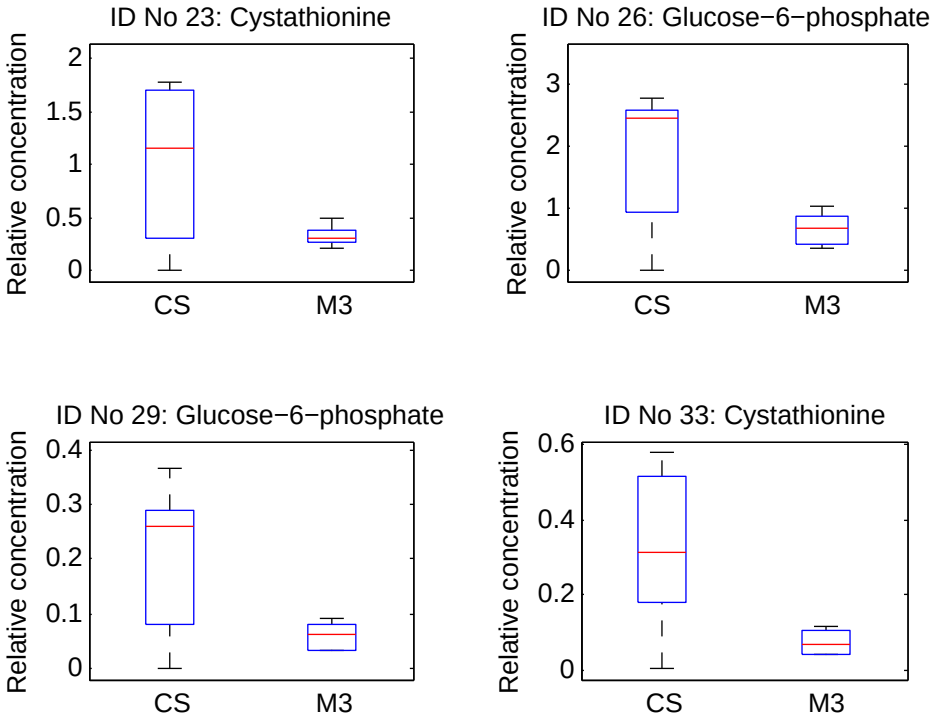

Figure 8.12: **Box plot metabolites: CS against M3**. Figure shows box plots of CS against M3 displaying the relative concentrations of four of the matching metabolites: cystathionine, Glucose-6-phosphate.

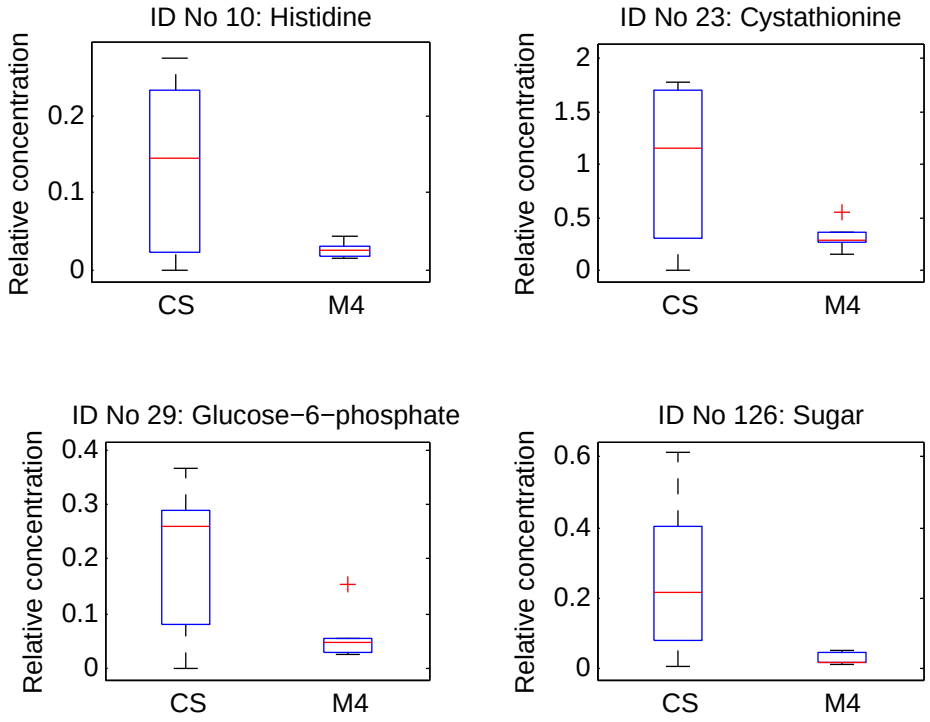

Figure 8.13: **Box plot metabolites: CS against M4**. Figure shows box plots of CS against M4 displaying the relative concentrations of four of the matching metabolites: histidine, cystathionine, glucose-6-phosphate,and sugar.

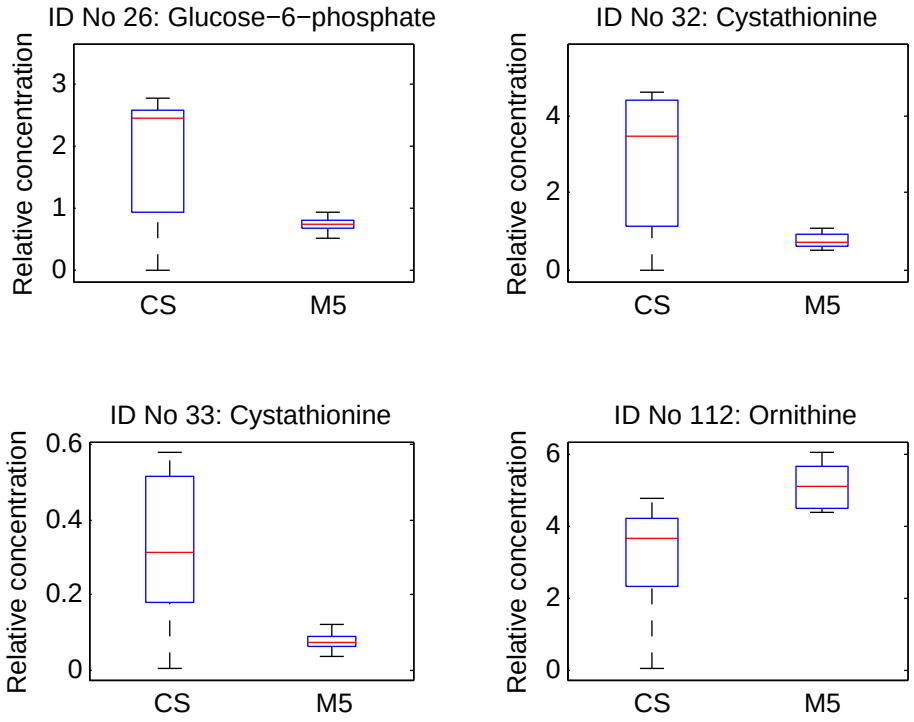

Figure 8.14: **Box plot metabolites: CS against M5**. Figure shows box plots of CS against M5 for the relative concentrations of four of the matching metabolites: Fructose-6-phosphate, glucose-6-phosphate, malic acid and aspartic acid.

8.4 Discussion

Further optimisation of the double mutants, especially the lysine overproducing producing mutants (M4 and M5) required a global assessment of the effects of genetic perturbations in the metabolic pathways of the strains using a systems biology approach. A systems biology tool, metabolomics, is most suited to the identification of bottlenecks for metabolic engineering as the level of the metabolome is close to that of the phenotype. Hence, metabolic profiling analysis carried out allowed for the comparative analysis of CS against each of double mutants, M2, M3, M4 and M5, except M1 which showed poor growth in the aerobic minimal media condition and was

excluded from further studies. The GC-TOF-MS results of the relative quantification of metabolites in the metabolome extracts of strains CS, MM2, M3, M4 and M5 for selected pathways showed both identified and unidentified metabolite peaks. The 81 identified metabolite peaks were involved in eleven metabolic pathways and product classes, namely "Amino acid biosyhnthesis", "Carbohydrate metabolism", "Fatty Acid Metabolism", "Glycerolipid Metabolism", "Glycolysis pathway", "Fermentation pathway","Metabolism of Cofactors and Vitamins", "Amino acid metabolism", "Tricarboxylic acid cycle", "Urea cycle", and "Purine and pyrimidine biosynthetic pathways" (see Tables 8.1, 8.2 and 8.3). In GC-MS analysis, the process of derivatisation sometimes leads to different derivatives of the same metabolite. Table 8.5, shows two metabolite peaks (IDs 14 and 15) identified as fructose-6-phosphate (CHEBI:15946). However, the results of the Welch's t-test in the comparison of CS against M2 strain, indicated that the fructose-6-phosphate with ID number 15 was a significantly changing metabolite ($p = 0.00730$) and which was not the case with the fructose-6-phosphate with ID number 14 ($p = 0.31181$). The most probable reason for this difference is that the concentrations detected were close to the methodological limit of detection, resulting in variation introduced at these low signal-to-noise ratios.

8.4.1 Effects of scaling

It was expected that the effects of the gene knockouts in the mutant strains would be reflected in the GC-MS metabolomic data sets as induced biological variation. Possibly too, uninduced biological variation such as large fluctuations in metabolite concentrations in the data may occur. In addition, heteroscedasticity may be reflected in GC-MS results due to the total uninduced variation as a result of biology, sampling procedures and analytical measurements. Hence, extraction of relevant biological information from metabolomics data sets require appropriate pretreatment of the data sets before data analysis. The pretreatment methods such as scaling and transformation improve the interpretability of data by removing the noise that hinders biological interpretation.

The two-dimensional plot of first two PCs (PC1 and PC2) of Pareto scaled data indicate fairly good clustering of all QC samples and fairly good separation for all samples except for four outliers: two samples of CS outliers, one sample of M5 and one sample of M4 (Figure 8.7). These results represented improvements over the results of PCA on unscaled clean data (Figures 8.4 and 8.5) which did not show clustering of all twelve QCs and and also showed more problems with closeness in the samples for M2, M3, M4, M5 and CS. A similar plot of PC1 and PC2 for clean data scaled by either center scaling, range scaling or autoscaling method did not show good clustering of QC samples as the pareto scaled data. Pareto scaling reduces the relative importance of large values, but keeps data structure partially intact and also corrects for heteroscedasticity (van den Berg et al., 2006). In addition, Pareto scaled data stays closer to the original measurement than autoscaling as shown in Figure C.3 (Appendix C) where the autoscaled clean sample data resulted in a fairly good separation for all samples except for two CS outliers and one M4, and in addition QCs separated into two groups. Although centering method allowed a fairly good separation for all samples, it only showed good clustering of ten out of twelve QC samples, with two outliers (see Figure C.1 in Appendix C), possibly due to the inability of center scaling approach to sufficiently scale the heteroscedastic data. A fairly good separation of samples and a separation of QCs into two groups were obtained when range scaling was applied on the clean data (see Figure C.2 in Appendix C), which renders this method a less robust pretreatment method for the data sets used in this study.

Pareto emerged as the method of choice for scaling the clean sample data sets used in this study because all the twelve replicate QC samples were closer together in the plot of PC1 against PC2 of pareto scaled data than in similar plots for data sets scaled by Center scaling, autoscaling and Range scaling methods (Appendix C, Figures C.1, C.3 and C.2). Since Pareto also corrects for heteroscedasticity (van den Berg et al., 2006) transformation (using logarithm or power) of clean data sets was not considered. Pareto scaled clean data improved over the unscaled clean data in

the amount of variance captured in PC1 from 23% (Figure 8.4) to 31% (Figure 8.6).

8.4.2 Analysis of results

Multivariate data analysis using PCA was used as the exploratory tool in the analysis of metabolic profiling dataset because it is an established method for reducing the dimension of data. PCA exposes the internal structure that best explains the variance in data. The univariate data analysis (hypothesis testing using Welch's t-test at 5% significance level) identified more metabolites with significantly different relative concentrations between CS and the mutant strains (M1, M2, M3 and M4) than PCA. In order to ensure that biological interpretation was based on true biological changes in relative concentrations of metabolites, it was decided to select the metabolites whose relative concentrations have been found by both Welch's t-test and PCA to be significantly different between the control strain CS and a specific mutant strain.

The twenty-nine metabolites with significant relative changes between the CS and M2 (Table 8.8) map to 7 different pathways: Purine and pyrimidine biosynthetic pathways (adenine), amino acid metabolism (cystathionine and 2-aminobutanoic acid), amino acid biosynthetic pathways (histidine, proline, leucine, serine, threonine, Homoserine, glutamine, valine, methionine, cysteine and phenylalanine), metabolism of cofactors and vitamins (phosphate, monmethyl ester), glycolysis (fructose-6-phosphate, glucose-6-phosphate and sugar), urea cycle (ornithine) and tricarboxylic acid cycle (malic acid). Similarly, for CS and M3, the seven significantly different metabolites (Table 8.9) are found in glycolysis (sugar), amino acid metabolism (cystathionine) and amino acid biosynthetic pathways (alanine and Homocysteine). The nine metabolites with significant relative changes between the CS and M4 (Table 8.10) map to three different pathways: glycolysis (glucose-6-phosphate and sugar), amino acid metabolism (cystathionine) and amino acid biosynthetic pathways (histidine,alanine and homocysteine). The seven changing metabolites in the comparison of CS and M5 (Table 8.11)indicated four affected pathways as follows: glycolysis (glucose-6-phosphate and

sugar), amino acid metabolism (cystathionine), amino acid biosynthetic pathways (aspartic acid) and urea cycle (ornithine).

The trends depicted in the boxplots of relative concentrations of CS against each of the other four mutants (examples in Figures 8.11, 8.12, 8.13 and 8.14) indicate that most of the significantly different metabolites were downregulated in endometabolome snapshot under study. Only aspartic acid and ornithine showed higher concentrations in M5 than CS, which matched the result of PCA as upregulated metabolites.

8.4.3 Biological significance of results

Mutant strains M2, M3, M4 and M5 contain the deletions, Δkgd2Δalt1, Δlsc1Δalt1, Δlsc2Δalt1 and Δalt1Δglt1, respectively, meant for redirection of flux to lysine pathway as shown in the metabolic and genetic flux map (Figure 8.15).

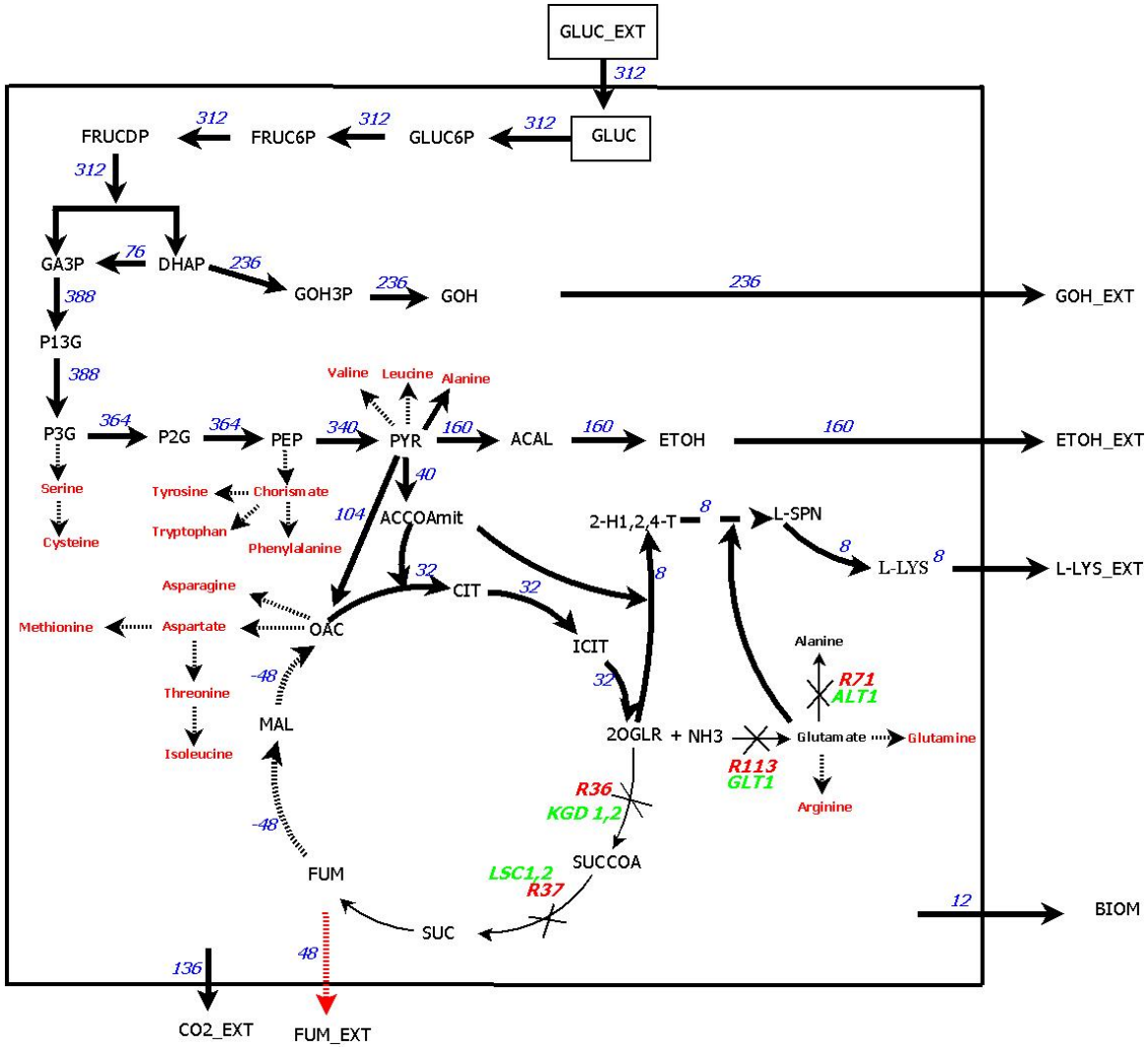

Figure 8.15: **Metabolic and genetic flux map**. Figure shows the genetic points of intervention in a network of reactions (including amino acids) for redirecting carbon flux towards increasing the yield of lysine. R36 is reaction number 36 representing the genes for the enzymes KGD1 and KGD2; R37 is reaction number 37 representing the genes for the enzymes LSC1 and LSC2; R71 is reaction number 71 representing the gene for the enzymes ALT1 and R113 is reaction number 113 representing the gene for the enzyme GLT1. Each of the numbers in blue represents the flux value of a reaction.

In a previous quantitative metabolomic experiment (section 7.3), M4 and M5 produced four times the amount of lysine produced by CS, while M2 and M3 did not produce an appreciable amounts of lysine than the control strain. The numbers of pathways and metabolites affected by changes in relative concentrations seems to reflect that genetic pertubations were more pronounced in the strain M2 than in strains M3, M4 and M5. Although, the results of this study concerning the metabolic profiling of CS against the mutants did not show lysine as one of the significant metabolites.

Evidence of overproduction of lysine in M4 and M5 might have been swamped by the relative concentrations of large-aboundant metabolites, and hence undetectable either as a major contributor to variance (PCA) or to be statistically significant by hypothesis testing. However, there are evidences of metabolic changes due to the induced genetic perturbations in the mutant strains. The general pattern of genetic effects in all mutants were disruption to the TCA cycle, glycolysis and branches of amino acid biosynthetic pathways.

8.4.3.1 Downregulated metabolites in mutant M2

In the case of M2, deletions of kgd2 and alt1 genes (Figure 8.15) might have disrupted the TCA cycle partially (since kgd1 gene must still have been active) and a also a disruption to production of alanine. The partial disruption in the TCA cycle at this point might have led to low level of malic acid and oxaloacetate which is the main precursor for the production of aspartate family of amino acids (aspartate, asparagine, isoleucine, threonine and methionine). Cystathionine, a product of amino acid metabolism was downregulated in M2. Notably too, the relative concentrations of fructose-6-phosphate, glucose-6-phosphate and certain unknown Sugars were lower than in CS. Possibly, disruption to the TCA cycle led to the reduced relative concentrations of the glycolytic metabolites.

8.4.3.2 Downregulated metabolites in mutant M3

Deletions of lsc1 and alt1 in M3 had the lesser effect than deletions of kgd2 and alt1 genes in strain M3 regrading the significantly changing metabolites between CS and M3 as only 7 identified metabolites from 3 different pathways were affected. However, M3 showed some similar effects in glycolysis, while the 2 amino acids affected are from 2 different amino acid families. As expected, deletion of alt1 gene led to low concentration of alanine.

8.4.3.3 Downregulated metabolites in mutant M4

Deletions in lsc2 and alt1 in M4 had the most similar effects to lsc2 and alt deletions in M3 in terms of the number of changing metabolites. One of the differences is the downregulated histidine level in M4. Metabolites in glycolysis and the TCA cycle are affected as mentioned for M2.

8.4.3.4 Downregulated metabolites in mutant M5

The genetic effects of knocking out alt1 and glt1 in M5 are similar to the perturbations in glycolysis and Cystathionine metabolism of strains M3 and M4. Although M5 shared one deletion (alt1) with M4, metabolic perturbations observed for glycolysis were similar. However, only M5 demonstrated aspartate and ornithine as upregulated and significantly changing amino acids. High flux of aspartate is beneficial to realising a high flux to lysine biosynthetic pathway. This may an indication and supporting evidence that strain M5 overproduces lysine. Higher concentration of ornithine in M5 than in CS certainly reflect the high aspartate flux, as aspartate is the link between aspartate biosynthetic pathway and the urea cycle.

8.5 Conclusion

Metabolic profiling analysis of four *S. cerevisiae* mutant strains, M2, M3, M4 and M5 revealed informative profiles of significantly changing metabolites when compared with a control strain (CS) under aerobic growth conditions in mimimal media. Although, direct evidence of flux redistribution towards overproduction of lysine was not apparent in the results obtained in this study, evidences of supportive genetic effects were demonstrated. The metabolites shown to be downregulated and upregulated in these strains indicate are a direct evidence of genetic perturbations in the mutant strains. Henceforth, it is important that results presented here is supported by metabolic flux analysis to reveal the nature of flux redistributions in the metabolic

pathways affected. Understanding the flux redistributions will enable the understanding of which enzymes to overexpress, especially in the glycolytic pathway, which will benefit the flux in the lysine biosynthetic pathway.

Chapter 9

General Discussion

9.1 Summary of findings

This thesis describes how systems biology approaches can be applied to rational metabolic engineering of *S. cerevisiae* for enhanced production of lysine and other commercially valuable products. Wild type *S. cerevisiae* is naturally poor in lysine, and lysine is one of the most sought-after essential amino acids due to its wide spectrum of applications, especially as food additives, components of therapeutic products and cosmetics.

The findings reported in chapter 6 show that the general modelling approach adopted here was successful. Computation of EFMs based on the stoichiometric models from the network of reactions from *S. cerevisiae* revealed the metabolic capabilities of *S. cerevisiae* for production of lysine, glutamate, ethanol, trehalose and fumaric acid. EFM analysis based on stoichiometric models revealed metabolic capabilities of *S. cerevisiae* grown on glucose under aerobic condition for production of lysine and other products. However, as the large number of EFMs obtained did not permit easy interpretation for metabolic engineering purposes, a novel methodology was developed to decipher the EFM data. The first part of the methodology were computational steps for reducing the dimensionality of the EFM data. The effectiveness of of this

was reduction up to 60.4%, was demonstrated in section 6.3.1. The second part of the methodology involved clustering analysis. The PAM clustering method, the best out of eight tested (see section 6.2.3.4), partitioned the EFMs into homogeneous groups that were neither too big nor too small, and had easy biological interpretation. In essence, the new method for decomposing EFM data into manageable subsets of EFMs, allowed for fast detection of alternative biochemical routes in the metabolic network for the development of improved *S. cerevisiae* production strains.

Chapter 6 also reported the successful exploitation of the EFM solution space for *in silico* design of single, double and triple mutants for improved lysine production in *S. cerevisiae*.

The successful construction of 5 double mutants (M1, M2, M3, M4 and M5) and one triple mutant, using a PCR-based gene deletion method was demonstrated in chapter 7. Quantitative GC-MS results reported in chapter 7 showed that out of the seven single mutants tested for lysine overproduction, the strains with single deletions, *lsc2* and *glt1*, excreted into the medium five times the amount of lysine than the control strain. In addition, two-fold increase in flux towards lysine production was demonstrated by *S. cerevisiae* double mutant M1, while both *S. cerevisiae* double mutants M4 and M5 showed four-fold increase in lysine production more than the control strain. However, the growth of *S. cerevisiae* triple mutant could not be sustained in minimal SD (3XALL). Possibly, the triple mutant was not viable due to the synthetic lethal effect of the deletion of *ZWF1* gene on top of the other two gene deletions. The *S. cerevisiae* with deleted *ZWF1* gene is known to be viable. The *in silico* triple mutant was expected to be more effective in eliminating the competing pathways to lysine biosynthetic pathway and was expected to produce more lysine. Hence, another possibility for the lack of viability of the triple mutant in 3XALL medium may be due to a toxic effect of high level of lysine and intermediates of lysine pathway in the medium.

Chapter 5 reports FBA-based *in silico* strain design. Ethanol and lysine overproducing *in silico* strains were also developed by OptKnock and GDLS. It was interesting

to find that some of the gene (reactions) target knockouts predicted by OptKnock and GDLS for lysine overproduction were from the TCA cycle as it was the case with the EFM-based *in silico* modelling. Furthermore, the knockout of Oxoglutarate Dehydrogenase (catalyses a crucial step in the TCA cycle) was predicted by OptKnock, GDLS and EFM analysis as a point of intervention for lysine overproduction in *S. cerevisiae*.

Chapter 8 examines the metabolite profiles of CS versus each of the 5 double mutants (M1, M2, M3, M4 and M5) in an attempt to optimise the demonstrated lysine over-producing mutants (M4 and M5). The final GC results presented were only for mutants M2, M3, M4 and M5 since mutant M1 did not grow well and was eliminated before GC-MS analysis. Metabolite profiling results indicate that all the four double mutants tested had lower relative metabolite concentrations than the control strain.

9.2 Future perspectives

In future it will be necessary to optimise the two lysine over-producing double mutant strains, M4 and M5. The bottleneck in strain optimisation is usually due to metabolic regulation. There are two possible different complementary approaches for achieving the optimisation of strains M4 and M5 in future, discussed as follows:

Tackling the known regulations in the lysine biosynthetic pathway:

The two homocitrate synthase isoenzymes which catalyse the first step of the lysine biosynthetic pathway are inhibited by lysine (Feller et al., 1999) and two intermediates in the lysine pathway. Removal of the lysine inhibition would be beneficial to increasing the intracellular accumulation of lysine as demonstrated by Feller *et al* (1999). However, it will be important to tackle the issue of lysine excretion which will be highly beneficial from commercial point of view. Excretion of lysine into the medium means less effort and cost for the industry in its purification. In yeast, there is a known vacuolar transporter of lysine and presumably transporters also export

lysine to the medium. Hence, in future, overexpression of the transporter and exporter genes in addition to the removal of lysine inhibition to the two homocitrate synthase isoenzymes would be necessary. In this way, strains M4 and M5 will be able to produce and excrete large amounts of lysine.

Integration of transcriptomics, proteomics and metabolomics:

Another improvement would be to take into account the transcriptional, translational and post-translational levels of regulation. Improvement of mutants M4 and M5 would indeed benefit from such a complete analysis. In this second approach, the combined approaches of transcriptomics, proteomics, metabolomics, bioinformatics and data analyses will be used to identify and monitor key the metabolic pathways in mutants M4 and M5. Integration of multi-omics studies will help unravel the multiple layers of control that superimpose the flux network, and also provide complementary coverage of metabolism. Biological insights arising from these studies will form the testable hypotheses for optimal design of lysine producing M4 and M5 strains.

Since the level of the metabolome is the closest to phenotype, it will be rewarding to carry out metabolite profiling of the two double mutant strains so as to link the genetic effects of knockouts to the metabolome. Transcriptomics and proteomics data will provide information about the locations of the affected genes in the metabolic pathways, and also help indicate which genes (for instance in the central metabolism) to overexpress. Increased flux in the central metabolism will be beneficial to the lysine biosynthetic pathway.

The above-outlined future perspectives concerns only the mutant strains developed by EFM analysis. The in silico mutant strains developed for lysine and ethanol using FBA methods will also require validation in the laboratory.

9.3 Conclusion

The results presented in this thesis validate the strategy of using EFM, FBA and the metabolomic measurements for the creation of the lysine overproducing mutants. Two of the strains developed here, M4 and M5, are very promising, showing 4X and 5X higher levels of lysine than the original strain. These are suitable platforms for further developments into higher *S. cerevisiae* lysine producers.

Bibliography

Akinterinwa, O., Khankal, R., and Cirino, P. C. (2008). Metabolic engineering for bioproduction of sugar alcohols. *Curr Opin Biotechnol*, 19(5):461–467.

Albertyn, J., Hohmann, S., Thevelein, J. M., and Prior, B. A. (1994). GPD1, which encodes glycerol-3-phosphate dehydrogenase, is essential for growth under osmotic stress in *Saccharomyces cerevisiae*, and its expression is regulated by the high-osmolarity glycerol response pathway. *Mol Cell Biol*, 14(6):4135–4144.

Alper, H., Miyaoku, K., and Stephanopoulos, G. (2005). Construction of lycopene-overproducing *Escherichia coli* strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol*, 23(5):612–616.

Altmeyer, P. J., Matthes, U., Pawlak, F., Hoffmann, K., Frosch, P. J., Ruppert, P., Wassilew, S. W., Horn, T., Kreysel, H. W., Lutz, G., and et al. (1994). Antipsoriatic effect of fumaric acid derivatives. Results of a multicenter double-blind study in 100 patients. *J Am Acad Dermatol*, 30(6):977–981.

Aristidou, A. and Penttila, M. (2000). Metabolic engineering applications to renewable resource utilization. *Curr Opin Biotechnol*, 11(2):187–198.

Askenazi, M., Driggers, E. M., Holtzman, D. A., Norman, T. C., Iverson, S., Zimmer, D. P., Boers, M. E., Blomquist, P. R., Martinez, E. J., Monreal, A. W., Feibelman, T. P., Mayorga, M. E., Maxon, M. E., Sykes, K., Tobin, J. V., Cordero, E., Salama, S. R., Trueheart, J., Royer, J. C., and Madden, K. T. (2003). Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat Biotechnol*, 21(2):150–156.

Azuma, S., Tsunekawa, H., Okabe, M., Okamoto, R., and Aiba, S. (1993). Hyperproduction of L-tryptophan via fermentation with crystallization. *Appl Microbiol Biotechnol*, 39:471–476.

Backman, K., O'Connor, M. J., Maruya, A., Rudd, E., McKay, D., Balakrishnan, R., Radjai, M., DiPasquantonio, V., Shoda, D., and Hatch, R. (1990). Genetic engineering of metabolic pathways applied to the production of phenylalanine. *Ann N Y Acad Sci*, 589:16–24.

Barrett, C. L., Kim, T. Y., Kim, H. U., Palsson, B. O., and Lee, S. Y. (2006). Systems biology as a foundation for genome-scale synthetic biology. *Curr Opin Biotechnol*, 17(5):488–492.

Basi, G., Schmid, E., and Maundrell, K. (1993). TATA box mutations in the *Schizosaccharomyces pombe* nmt1 promoter affect transcription efficiency but not the transcription start point or thiamine repressibility. *Gene*, 123(1):131–136.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. (1993). A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 21(14):3329–3330.

Ben-Amotz, A. and Avron, M. (1979). Osmoregulation in the halophilic algae *Dunaliella* and *Asteromonas*. *Basic Life Sci*, 14:91–99.

Berry, A. (1996). Improving production of aromatic compounds in *Escherichia coli* by metabolic engineering. *Trends Biotechnol*, 14(7):250–256.

Biebl, H. (1991). Glycerol Fermentation Of 1,3-Propanediol By *Clostridium butyricum* Measurement Of Product Inhibition By Use Of A pH-Auxostat. *Appl Microbiol Biotechnol*, 35:701–705.

Biebl, H., Marten, S., Hippe, H., and Deckwer, W. D. (1992). Glycerol conversion to 1,3-propanediol by newly isolated *clostridia*. *Appl Microbiol Biotechnol*, 36:592–597.

Blombach, B., Schreiner, M. E., Bartek, T., Oldiges, M., and Eikmanns, B. J. (2008). *Corynebacterium glutamicum* tailored for high-yield L-valine production. *Appl Microbiol Biotechnol*, 79(3):471–479.

Blombach, B., Schreiner, M. E., Holatko, J., Bartek, T., Oldiges, M., and Eikmanns, B. J. (2007). L-valine production with pyruvate dehydrogenase complex-deficient *Corynebacterium glutamicum*. *Appl Environ Microbiol*, 73(7):2079–2084.

Boenigk, R., Bowien, S., and Gottschalk, G. (1993). Fermentation of glycerol to 1,3-propanediol in continuous cultures of *Citrobacter freundii*. *Appl Microbiol Biotechnol*, 38:453–457.

Boghigian, B. A., Shi, H., Lee, K., and Pfeifer, B. A. (2010). Utilizing elementary mode analysis, pathway thermodynamics, and a genetic algorithm for metabolic flux determination and optimal metabolic network design. *BMC Syst Biol*, 4:49.

Branduardi, P., Fossati, T., Sauer, M., Pagani, R., Mattanovich, D., and Porro, D. (2007). Biosynthesis of vitamin C by yeast leads to increased stress resistance. *PLoS One*, 2(10):e1092.

Brown, M., Dunn, W. B., Dobson, P., Patel, Y., Winder, C. L., Francis-McIntyre, S., Begley, P., Carroll, K., Broadhurst, D., Tseng, A., Swainston, N., Spasic, I., Goodacre, R., and Kell, D. B. (2009). Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134(7):1322–1332.

Bruggeman, F. J. and Westerhoff, H. V. (2007). The nature of systems biology. *Trends Microbiol*, 15(1):45–50.

Brynildsen, M. P. and Liao, J. C. (2009). An integrated network approach identifies the isobutanol response network of *Escherichia coli*. *Mol Syst Biol*, 5:277.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–657.

Cakir, T., Kirdar, B., and Ulgen, K. O. (2004). Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnol Bioeng*, 86(3):251–260.

Caplice, E. and Fitzgerald, G. F. (1999). Food fermentations: role of microorganisms in food production and preservation. *Int J Food Microbiol*, 50(1-2):131–149.

Carpinelli, J., Kramer, R., and Agosin, E. (2006). Metabolic engineering of *Corynebacterium glutamicum* for trehalose overproduction: role of the TreYZ trehalose biosynthetic pathway. *Appl Environ Microbiol*, 72(3):1949–1955.

Chibata, I., Tosa, T., and Takata, I. (1983). Continuous production of L-malic acid by mmobilized cells. *Trends Biotechnol*, 1:9–11.

Choi, D. K., Ryu, W. S., Choi, C. Y., and Park, Y. H. (1996). Production of L-ornithine by arginine auxotrophic mutants of *Brevibacterium ketoglutamicum* in dual substratelimited continuous culture. *J Ferment Bioeng*, 81:216–219.

Chotani, G., Dodge, T., Hsu, A., Kumar, M., LaDuca, R., Trimbur, D., Weyler, W., and Sanford, K. (2000). The commercial production of chemicals using pathway engineering. *Biochimica et Biophysica Acta*, 1543:434–435.

Cordier, H., Mendes, F., Vasconcelos, I., and Francois, J. M. (2007). A metabolic and genomic study of engineered *Saccharomyces cerevisiae* strains for high glycerol production. *Metab Eng*, 9(4):364–378.

Dale, B. E. (2003). "Greening" the chemical industry: research and development priorities for biobased industrial products. *J Chem Technol Biotechnol*, 78:1093–1103.

De Paepe, A. and Van der Straeten, D. (2005). Ethylene biosynthesis and signaling: an overview. *Vitam Horm*, 72:399–430.

Deng, X. X. and Ho, N. W. (1990). Xylulokinase activity in various yeasts including *Saccharomyces cerevisiae* containing the cloned xylulokinase gene. Scientific note. *Appl Biochem Biotechnol*, 24-25:193–199.

Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. (2006). Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat Genet*, 38(9):993–998.

Dobson, P. D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., Lu, C., Swainston, N., Dunn, W. B., Fisher, P., Hull, D., Brown, M., Oshota, O., Stanford, N. J., Kell, D. B., King, R. D., Oliver, S. G., Stevens, R. D., and Mendes, P. (2010). Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol*, 4:145.

Draths, K. M. and Frost, J. W. (1991). Conversion of D-glucose into catechol: the not-so-common pathway of aromatic biosynthesis. *J Am Chem Soc*, 113:9361–9363.

Draths, K. M. and Frost, J. W. (1992). Biocatalysis and nineteenth century organic chemistry: conversion of D-glucose into quinoid organics. *J Am Chem Soc*, 114:9725–9726.

Draths, K. M. and Frost, J. W. (1994). Environmentally compatible synthesis of adipic acid from D-glucose. *J Am Chem Soc*, 114:9725–9726.

Duarte, N. C., Becker, S. A., Jamshidi, N., Thiele, I., Mo, M. L., Vo, T. D., Srivas, R., and Palsson, B. O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6):1777–1782.

Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004). Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309.

Ecker, J. R. (1995). The ethylene signal transduction pathway in plants. *Science*, 268(5211):667–675.

Edwards, J. S., Ramakrishna, R., and Palsson, B. O. (2002). Characterizing the

metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng*, 77(1):27–36.

Eliasson, A., Christensson, C., Wahlbom, C. F., and Hahn-Hagerdal, B. (2000). Anaerobic xylose fermentation by recombinant *Saccharomyces cerevisiae* carrying XYL1, XYL2, and XKS1 in mineral medium chemostat cultures. *Appl Environ Microbiol*, 66(8):3381–3386.

Ensley, B. D., Ratzkin, B. J., Osslund, T. D., Simon, M. J., Wackett, L. P., and Gibson, D. T. (1983). Expression of naphthalene oxidation genes in *Escherichia coli* results in the biosynthesis of indigo. *Science*, 222(4620):167–169.

EuropaBio (2004). EuropaBio, White Biotechnology: Gateway to a More Sustainable Future, 2004, 14. Technical report.

Farfan, M. and Calderon, I. L. (2000). Enrichment of threonine content in *Saccharomyces cerevisiae* by pathway engineering. *Enzyme Microb Technol*, 26(9-10):763–770.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. O. (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, A. M. and Palsson, B. O. (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol*, 26(6):659–667.

Fell, D. A. (1992). Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem J*, 286 (Pt 2):313–330.

Feller, A., Ramos, F., Pierard, A., and Dubois, E. (1999). In *Saccharomyces cerevisiae*, feedback inhibition of homocitrate synthase isoenzymes by lysine modulates the activation of LYS gene expression by Lys14p. *Eur J Biochem*, 261(1):163–170.

Fiehn, O. (2002). Metabolomics–the link between genotypes and phenotypes. *Plant Mol Biol*, 48(1-2):155–171.

Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., and Palsson, B. O. (2005). *in silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng*, 91(5):643–648.

Fossati, T., Solinas, N., Porro, D., and Branduardi, P. (2010). L-ascorbic acid producing yeasts learn from plants how to recycle it. *Metab Eng*, 13(2):177–185.

Fukuda, H., Ogawa, T., and Tanase, S. (1993). Ethylene production by microorganisms. *Adv Microb Physiol*, 35:275–306.

Fukuda, H., Ogawa, T., Tazaki, M., Nagahama, K., Fujii, T., Tanase, S., and Morino, Y. (1992). Two reactions are simultaneously catalyzed by a single enzyme: the arginine-dependent simultaneous formation of two products, ethylene and succinate, from 2-oxoglutarate by an enzyme from *Pseudomonas syringae*. *Biochem Biophys Res Commun*, 188(2):483–489.

Gianchandani, E. P., Chavali, A. K., and Papin, J. A. (2010). The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med*, 2(3):372–382.

Gietz, R. D. and Woods, R. A. (2001). Genetic transformation of yeast. *Biotechniques*, 30(4):816–820, 822–826, 828 passim.

Gietz, R. D. and Woods, R. A. (2006). Yeast transformation by the LiAc/SS Carrier DNA/PEG method. *Methods Mol Biol*, 313:107–120.

Goldberg, I. (2006). Review Organic acids: old metabolites, new themes . *J Chem Technol Biotechnol*, 81:1601–1611.

Gombert, A. K., Moreira dos Santos, M., Christensen, B., and Nielsen, J. (2001). Network identification and flux quantification in the central metabolism of *Saccharomyces cerevisiae* under different conditions of glucose repression. *J Bacteriol*, 183(4):1441–1451.

Goto, M., Ishida, Y., Takikawa, Y., and Hyodo, H. (1985). Ethylene production by the kudzu strains of *Pseudomonas syringae* pv. phaeseolicola causing halo blight in Pueraria lobata (Willd) Ohwi. *Plant Cell Physiol*, 26(1):141–150.

Grafahrend-Belau, E., Schreiber, F., Heiner, M., Sackmann, A., Junker, B. H., Grunwald, S., Speer, A., Winder, K., and Koch, I. (2008). Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics*, 9:90.

Granstrom, T. B., Izumori, K., and Leisola, M. (2007). A rare sugar xylitol. Part II: biotechnological production and future applications of xylitol. *Appl Microbiol Biotechnol*, 74(2):273–276.

Griffin, J. L. (2004). Metabolic profiles to define the genome: can we hear the phenotypes? *Philos Trans R Soc Lond B Biol Sci*, 359(1446):857–871.

Gueldener, U., Heinisch, J., Koehler, G. J., Voss, D., and Hegemann, J. H. (2002). A second set of loxP marker cassettes for Cre-mediated multiple gene knockouts in budding yeast. *Nucleic Acids Res*, 30(6):e23.

Guldener, U., Heck, S., Fielder, T., Beinhauer, J., and Hegemann, J. H. (1996). A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic Acids Res*, 24(13):2519–2524.

Hahn-Hagerdal, B., Karhumaa, K., Fonseca, C., Spencer-Martins, I., and Gorwa-Grauslund, M. F. (2007). Towards industrial pentose-fermenting yeast strains. *Appl Microbiol Biotechnol*, 74(5):937–953.

Hanai, T., Atsumi, S., and Liao, J. C. (2007). Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Appl Environ Microbiol*, 73(24):7814–7818.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212.

Hatzimanikatis, V. and Bailey, J. E. (1996). MCA has more to say. *J Theor Biol*, 182(3):233–242.

Hatzimanikatis, V. and Bailey, J. E. (1997). Effects of spatiotemporal variations on metabolic control: approximate analysis using (log)linear kinetic models. *Biotechnol Bioeng*, 54(2):91–104.

Hayn, M., Steiner, W., Klinger, R., Steinmuller, H., Sinner, M., and Esterbauer, H. (1993). Basic research and pilot studies on the enzymatic conversion of lignocellulosics. Bioconversion of Forest and Agricultural Plant Residues, In Saddler J.N.(Eds), *Bioconversion of forest and agricultural plant residues, CAB International, Wallingford, U.K,* page 33-72.

Hirche, C. (2006). Trend Report No. 16: Industrial Biotechnology - White Biotechnology: A promising tool for the chemical industry. In *ACHEMA 2006 28th International Exhibition-Congress on Chemical Engineering, Environmental Protection and Biotechnology*, volume 16, pages 1–7.

Ho, N. W., Chen, Z., and Brainard, A. P. (1998). Genetically engineered *Saccharomyces* yeast capable of effective cofermentation of glucose and xylose. *Appl Environ Microbiol*, 64(5):1852–1859.

Hoefnagel, M. H., Starrenburg, M. J., Martens, D. E., Hugenholtz, J., Kleerebezem, M., Van, Swam, I., Bongers, R., Westerhoff, H. V., and Snoep, J. L. (2002). Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis. *Microbiology*, 148(Pt 4):1003–1013.

Hong, S. (2007). Systems Approaches to Succinic Acid-producing Microorganisms. *Biotechnol Bioproc Eng*, 12:73–79.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., and Kummer, U. (2006). COPASI–a COmplex PAthway SImulator. *Bioinformatics*, 22(24):3067–3074.

Hugenholtz, J. (2008). The lactic acid bacterium as a cell factory for food ingredient production. *Int Dairy J*, 18:466–475.

Huh, W. K., Lee, B. H., Kim, S. T., Kim, Y. R., Rhie, G. E., Baek, Y. W., Hwang, C. S., Lee, J. S., and Kang, S. O. (1998). D-Erythroascorbic acid is an important antioxidant molecule in *Saccharomyces cerevisiae*. *Mol Microbiol*, 30(4):895–903.

Ikeda, M. (2003). Amino acid production processes. *Adv Biochem Eng Biotechnol*, 79:1–35.

Ikeda, M. and Katsumata, R. (1999). Hyperproduction of tryptophan by *Corynebacterium glutamicum* with the modified pentose phosphate pathway. *Appl Environ Microbiol*, 65(6):2497–2502.

Ikeda, M., Nakanishi, K., Kino, K., and Katsumata, R. (1994). Fermentative production of trptophan by a stable recombinant strain of *Corynebacterium glutamicum* with a modified serine-biosyn-thetic pathway. *Biosci Biotech Biochem*, 58:674–676.

Ingram, L. O., Aldrich, H. C., Borges, A. C., Causey, T. B., Martinez, A., Morales, F., Saleh, A., Underwood, S. A., Yomano, L. P., York, S. W., Zaldivar, J., and Zhou, S. (1999). Enteric bacterial catalysts for fuel ethanol production. *Biotechnol Prog*, 15(5):855–866.

Iwatani, S., Yamada, Y., and Usuda, Y. (2008). Metabolic flux analysis in biotechnology processes. *Biotechnol Lett*, 30(5):791–799.

Janke, C., Magiera, M. M., Rathfelder, N., Taxis, C., Reber, S., Maekawa, H., Moreno-Borchart, A., Doenges, G., Schwob, E., Schiebel, E., and Knop, M. (2004). A versatile toolbox for PCR-based tagging of yeast genes: new fluorescent proteins, more markers and promoter substitution cassettes. *Yeast*, 21(11):947–962.

Jeffries, T. W. and Jin, Y. S. (2004). Metabolic engineering for improved fermentation of pentoses by yeasts. *Appl Microbiol Biotechnol*, 63(5):495–509.

Jin, Y. S., Cruz, J., and Jeffries, T. W. (2005). Xylitol production by a *Pichia stipitis* D-xylulokinase mutant. *Appl Microbiol Biotechnol*, 68(1):42–45.

Jojima, T., Inui, M., and Yukawa, H. (2008). Production of isopropanol by metabolically engineered *Escherichia coli*. *Appl Microbiol Biotechnol*, 77(6):1219–1224.

Kadam, K. L., Chin, C. Y., and Brown, L. W. (2008). Flexible biorefinery for producing fermentation sugars, lignin and pulp from corn stover. *J Ind Microbiol Biotechnol*, 35(5):331–341.

Katsumata, R. and Ikeda, M. (1993). Hyperproduction of tryptophan in *Corynebacterium glutamicum* by pathway engineering. *Nature Biotechnology*, 11:921–925.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Keasling, J. D. and Chou, H. (2008). Metabolic engineering delivers next-generation biofuels. *Nat Biotechnol*, 26(3):298–299.

Kern, A., Tilley, E., Hunter, I. S., Legisa, M., and Glieder, A. (2007). Engineering primary metabolic pathways of industrial micro-organisms. *J Biotechnol*, 129(1):6–29.

Kim, H. U., Kim, T. Y., and Lee, S. Y. (2008). Metabolic flux analysis and metabolic engineering of microorganisms. *Mol Biosyst*, 4(2):113–120.

Kim, J. and Reed, J. L. (2010). OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol*, 4:53.

Klamt, S., Saez-Rodriguez, J., and Gilles, E. D. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol*, 1:2.

Klamt, S. and Stelling, J. (2002). Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236.

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21(2):64–69.

Klamt, S. and Stelling, J. (2006). *Stoichiometric and constraint-based modeling.* In: Z. Szallasi, J. Stelling, V. Periwal, ed. 2006. System modeling in cellular biology: From concepts to nuts and bolts. The MIT Press, pp. 73-96, London.

Kohlstedt, M., Becker, J., and Wittmann, C. (2010). Metabolic fluxes and beyond-systems biology understanding and engineering of microbial metabolism. *Appl Microbiol Biotechnol*, 88(5):1065–1075.

Koizumi, S., Endo, T., Tabata, K., and Ozaki, A. (1998). Large-scale production of UDP-galactose and globotriose by coupling metabolically engineered bacteria. *Nat Biotechnol*, 16(9):847–850.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R., and Steinhauser, D. (2005). GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, 21(8):1635–1638.

Kotter, P. and Ciriacy, M. (1993). Xylose fermentation by *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol*, 38:776–783.

Kramer, R. (2005). Production of amino acids: physiological and genetic approaches. *Food Biotechnol*, 18(2):171–216.

Krishnan, M. S., Ho, N. W., and Tsao, G. T. (1999). Fermentation kinetics of ethanol production from glucose and xylose by recombinant Saccharomyces 1400(pLNH33). *Appl Biochem Biotechnol*, 77-79:373–388.

Kromer, J. O., Wittmann, C., Schroder, H., and Heinzle, E. (2006). Metabolic pathway analysis for rational design of L-methionine production by *Escherichia coli* and *Corynebacterium glutamicum*. *Metab Eng*, 8(4):353–369.

Kuhn, A., van Zyl, C., van Tonder, A., and Prior, B. A. (1995). Purification and partial characterization of an aldo-keto reductase from *Saccharomyces cerevisiae*. *Appl Environ Microbiol*, 61(4):1580–1585.

Kurtzman, C. P. (1994). Molecular taxonomy of the yeasts. *Yeast*, 10(13):1727–1740.

Kwon, D. H., Kim, M. D., Lee, T. H., Oh, Y. J., Ryu, Y. W., and Seo, J. H. (2006). Elevation of glucose 6-phosphate dehydrogenase activity increases xylitol production in recombinant *Saccharomyces cerevisiae*. *J Mol Catal B: Enzym*, 43:86–89.

Lee, J. K., Song, J. Y., and Kim, S. Y. (2003). Controlling substrate concentration in fed-batch candida magnoliae culture increases mannitol production. *Biotechnol Prog*, 19(3):768–775.

Lee, J. M., Gianchandani, E. P., and Papin, J. A. (2006a). Flux balance analysis in the era of metabolomics. *Brief Bioinform*, 7(2):140–150.

Lee, J. W., Lee, S. Y., Song, H., and Yoo, J. S. (2006b). The proteome of Mannheimia succiniciproducens, a capnophilic rumen bacterium. *Proteomics*, 6(12):3550–3566.

Lee, K. H., Park, J. H., Kim, T. Y., Kim, H. U., and Lee, S. Y. (2007). Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol*, 3:149.

Lee, S., Phalakornkule, C., Domach, M. M., and Grossmann, I. E. (2000). Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comp Chem Eng*, 24:711–716.

Lee, S. J., Lee, D. Y., Kim, T. Y., Kim, B. H., Lee, J., and Lee, S. Y. (2005a). Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. *Appl Environ Microbiol*, 71(12):7880–7887.

Lee, S. K., Chou, H., Ham, T. S., Lee, T. S., and Keasling, J. D. (2008). Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr Opin Biotechnol*, 19(6):556–563.

Lee, S. Y. and Park, J. H. (2010). Integration of systems biology with bioprocess engineering: L: -threonine production by systems metabolic engineering of *Escherichia coli*. *Adv Biochem Eng Biotechnol*, 120:1–19.

Lee, S. Y., Woo, H. M., Lee, L., Choi, H. S., Kim, T. Y., and Yun, H. (2005b). Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol Bioprocess Eng*, 10:425–431.

Liebermeister, W. and Klipp, E. (2006). Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor Biol Med Model*, 3:41.

Ling, M., Merante, F., and Robinson, B. H. (1995). A rapid and reliable DNA preparation method for screening a large number of yeast clones by polymerase chain reaction. *Nucleic Acids Res*, 23(23):4924–4925.

Liu, D. X., Fan, C. S., Tao, J. H., Liang, G. X., Gao, S. E., Wang, H. J., Li, X., and Song, D. X. (2004). Integration of *Escherichia coli* aroG-pheA tandem genes into *Corynebacterium glutamicum* tyrA locus and its effect on L-phenylalanine biosynthesis. *World J Gastroenterol*, 10(24):3683–3687.

Lun, D. S., Rockwell, G., Guido, N. J., Baym, M., Kelner, J. A., Berger, B., Galagan, J. E., and Church, G. M. (2009). Large-scale identification of genetic design strategies using local search. *Mol Syst Biol*, 5:296.

Maertens, J. and Vanrolleghem, P. A. (2010). Modeling with a view to target identification in metabolic engineering: a critical evaluation of the available tools. *Biotechnol Prog*, 26(2):313–331.

Mahadevan, R. and Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng*, 5(4):264–276.

McCoy, M., Reisch, M., Tullo, A. H., Short, P. L., Tremblay, J. F., and Storck, W. J. (2006). Production: growth is the norm. *Chem Eng News*, News 84:59–68.

McKinlay, J. B., Vieille, C., and Zeikus, J. G. (2007). Prospects for a bio-based succinate industry. *Appl Microbiol Biotechnol*, 76(4):727–740.

Mecking, S. (2004). Nature or petrochemistry?-biologically degradable materials. *Angew Chem Int Ed Engl*, 43(9):1078–1085.

Mendes, P. and Kell, D. (1998). Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883.

Menzel, K. and Zeng, A. P. (1997). High concentration and productivity of 1,3-propanediol from continuous fermentation of glycerol by *Klebsiella pneumoniae*. *Enzyme Microb Technol*, 20(2):82–86.

Misawa, N. and Shimada, H. (1997). Metabolic engineering for the production of carotenoids in non-carotenogenic bacteria and yeasts. *J Biotechnol*, 59(3):169–181.

Mitsuhashi, S., Hayashi, M., Ohnishi, J., and Ikeda, M. (2006). Disruption of malate:quinone oxidoreductase increases L-lysine production by *Corynebacterium glutamicum*. *Biosci Biotechnol Biochem*, 70(11):2803–2806.

Mo, M. L., Palsson, B. O., and Herrgard, M. J. (2009). Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol*, 3:37.

Moon, S. Y., Hong, S. H., Kima, T. Y., and Lee, S. Y. (2008). Metabolic engineering of *Escherichia coli* for the production of malic acid. *Biochemical Engineering Journal*, 40:312–320.

Morbach, S., Sahm, H., and Eggeling, L. (1995). Use of Feedback-Resistant Threonine Dehydratases of *Corynebacterium glutamicum* To Increase Carbon Flux towards l-Isoleucine. *Appl Environ Microbiol*, 61(12):4315–4320.

Morbach, S., Sahm, H., and Eggeling, L. (1996). l-Isoleucine Production with *Corynebacterium glutamicum*: Further Flux Increase and Limitation of Export. *Appl Environ Microbiol*, 62(12):4345–4351.

Moreno-Sanchez, R., Saavedra, E., Rodriguez-Enriquez, S., and Olin-Sandoval, V. (2008). Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways. *J Biomed Biotechnol*, 2008:597–913.

Muller, U., van Assema, F., Gunsior, M., Orf, S., Kremer, S., Schipper, D., Wagemans, A., Townsend, C. A., Sonke, T., Bovenberg, R., and Wubbolts, M. (2006). Metabolic engineering of the *Escherichia coli* L-phenylalanine pathway for the production of D-phenylglycine (D-Phg). *Metab Eng*, 8(3):196–208.

Nagahama, K., Yoshino, K., Matsuoka, M., Sato, M., Tanase, S., Ogawa, T., and Fukuda, H. (1994). Ethylene production by strains of the plant-pathogenic bacterium *Pseudomonas syringae* depends upon the presence of indigenous plasmids carrying homologous genes for the ethylene-forming enzyme. *Microbiology*, 140 (Pt 9):2309–2313.

Nakamura, C. E. and Whited, G. M. (2003). Metabolic engineering for the microbial production of 1,3-propanediol. *Curr Opin Biotechnol*, 14(5):454–459.

Nielsen, J. (2003). It Is All about Metabolic Fluxes. *J Bacteriol*, 185(24):7031–7035.

Nielsen, J. and Jewett, M. C. (2008). Impact of systems biology on metabolic engineering of *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 8(1):122–131.

Oberhardt, M. A., Palsson, B. O., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320.

O'Hagan, S., Dunn, W. B., Brown, M., Knowles, J. D., and Kell, D. B. (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(1):290–303.

Ohnishi, J., Katahira, R., Mitsuhashi, S., Kakita, S., and Ikeda, M. (2005). A novel gnd mutation leading to increased L-lysine production in *Corynebacterium glutamicum*. *FEMS Microbiol Lett*, 242(2):265–274.

Oldiges, M., Lutz, S., Pflug, S., Schroer, K., Stein, N., and Wiendahl, C. (2007). Metabolomics: current state and evolving methodologies and tools. *Appl Microbiol Biotechnol*, 76(3):495–511.

Olivier, B. G., Rohwer, J. M., and Hofmeyr, J. H. (2005). Modelling cellular systems with PySCeS. *Bioinformatics*, 21(4):560–561.

Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248.

Otero, J. M. and Nielsen, J. (2010). Industrial systems biology. *Biotechnol Bioeng*, 105(3):439–460.

Panke, S., Sanchez-Romero, J. M., and de Lorenzo, V. (1998). Engineering of quasi-natural *Pseudomonas putida* strains for toluene metabolism through an ortho-cleavage degradation pathway. *Appl Environ Microbiol*, 64(2):748–751.

Parekh, S., Vinci, V. A., and Strobel, R. J. (2000). Improvement of microbial strains and fermentation processes. *Appl Microbiol Biotechnol*, 54(3):287–301.

Park, J. H., Lee, K. H., Kim, T. Y., and Lee, S. Y. (2007). Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proc Natl Acad Sci U S A*, 104(19):7797–7802.

Park, J. H. and Lee, S. Y. (2008). Towards systems metabolic engineering of microorganisms for amino acid production. *Curr Opin Biotechnol*, 19(5):454–460.

Patil, K. R., Rocha, I., Forster, J., and Nielsen, J. (2005). Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics*, 6:308.

Patnaik, R. and Liao, J. C. (1994). Engineering of *Escherichia coli* central metabolism for aromatic metabolite production with near theoretical yield. *Appl Environ Microbiol*, 60(11):3903–3908.

Peralta-Yahya, P. P. and Keasling, J. D. (2010). Advanced biofuel production in microbes. *Biotechnol J*, 5(2):147–162.

Peres, S., Beurton-Aimar, M., and Mazat, J. P. (2006). Pathway classification of TCA cycle. *Syst Biol (Stevenage)*, 153(5):369–371.

Peres, S., Vallee, F., Beurton-Aimar, M., and Mazat, J. P. (2011). ACoM: A classification method for elementary flux modes based on motif finding. *Biosystems*, 103(3):410–419.

Peters-Wendisch, P., Stolz, M., Etterich, H., Kennerknecht, N., Sahm, H., and Eggeling, L. (2005). Metabolic engineering of *Corynebacterium glutamicum* for L-serine production. *Appl Environ Microbiol*, 71(11):7139–7144.

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2003). Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol Bioeng*, 84(7):887–899.

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res*, 14(11):2367–2376.

Pharkya, P. and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng*, 8(1):1–13.

Pines, O., Shemesh, S., Battat, E., and Goldberg, I. (1997). Overexpression of cytosolic malate dehydrogenase (MDH2) causes overproduction of specific organic acids in *Saccharomyces cerevisiae*. *Appl Microbiol Biotechnol*, 48(2):248–255.

Pool, W. A., Neves, A. R., Kok, J., Santos, H., and Kuipers, O. P. (2006). Natural sweetening of food products by engineering *Lactococcus lactis* for glucose production. *Metab Eng*, 8(5):456–464.

Prasad, R., Niederberger, P., and Hutter, R. (1987). Tryptophan accumulation in *Saccharomyces cerevisiae* under the influence of an artificial yeast TRP gene cluster. *Yeast*, 3(2):95–105.

Pscheidt, B. and Glieder, A. (2008). Yeast cell factories for fine chemical and API production. *Microb Cell Fact*, 7:25.

Pszczola, D. E. (1992). The nutraceutical initiative: a proposal for economic and regulatory reform. *Food Biotechnol*, 46:77–79.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, 24(3):218–229.

Ramakrishna, R., Edwards, J. S., McCulloch, A., and Palsson, B. O. (2001). Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. *Am J Physiol Regul Integr Comp Physiol*, 280(3):R695–704.

Richard, P., Toivari, M., and Penttil, M. (1999). Evidence that the gene YLR070c of *Saccharomyces cerevisiae* encodes a xylitol dehydrogenase. *FEBS Lett*, 457(1):135–138.

Richard, P., Verho, R., Putkonen, M., Londesborough, J., and Penttila, M. (2003). Production of ethanol from L-arabinose by *Saccharomyces cerevisiae* containing a fungal L-arabinose pathway. *FEMS Yeast Res*, 3(2):185–189.

Roa Engel, C. A., Straathof, A. J., Zijlmans, T. W., van Gulik, W. M., and van der Wielen, L. A. (2008). Fumaric acid production by fermentation. *Appl Microbiol Biotechnol*, 78(3):379–389.

Rocha, I., Maia, P., Evangelista, P., Vilaca, P., Soares, S., Pinto, J. P., Nielsen, J., Patil, K. R., Ferreira, E. C., and Rocha, M. (2010). OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC Syst Biol*, 4:45.

Rogers, P. L., Jeon, Y., and Svenson, C. J. (2005). Application of biotechnology to industrial sustainability. *Process Safety Environ Protec*, 83(B6):499–503.

Ross, R. P., Morgan, S., and Hill, C. (2002). Preservation and fermentation: past, present and future. *Int J Food Microbiol*, 79(1-2):3–16.

Rueffer, N., Heidersdorf, U., Kretzers, I., Sprenger, G. A., Raeven, L., and Takors, R. (2004). Fully integrated L-phenylalanine separation and concentration using reactiveextraction with liquid-liquid centrifuges in a fed-batch process with *Escherichia coli*. *Bioproc and Biosy Eng*, 26:239–248.

Ruffing, A. and Chen, R. R. (2006). Metabolic engineering of microbes for oligosaccharide and polysaccharide synthesis. *Microb Cell Fact*, 5:25.

Saito, Y., Ishii, Y., Hayashi, H., Imao, Y., Akashi, T., Yoshikawa, K., Noguchi, Y., Soeda, S., Yoshida, M., Niwa, M., Hosoda, J., and Shimomura, K. (1997). Cloning of genes coding for L-sorbose and L-sorbosone dehydrogenases from *Gluconobacter oxydans* and microbial production of 2-keto-L-gulonate, a precursor of L-ascorbic acid, in a recombinant *Gluconobacter oxydans* strain. *Appl Environ Microbiol*, 63(2):454–460.

Sauer, U. (2001). Evolutionary engineering of industrially important microbial phenotypes. *Adv Biochem Eng Biotechnol*, 73:129–169.

Sauer, U. (2006). Metabolic networks in motion: 13C-based flux analysis. *Mol Syst Biol*, 2:62.

Savageau, M. A. (1976). *Biochemical System Analysis, 1st ed. Reading: Addison-Wesley Publishing Company; 1976.*

Savinell, J. M. and Palsson, B. O. (1992). Network analysis of intermediary metabolism using linear optimization. II. Interpretation of hybridoma cell metabolism. *J Theor Biol*, 154(4):455–473.

Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc*, 6(9):1290–1307.

Schilling, C. H., Edwards, J. S., Letscher, D., and Palsson, B. O. (2001). Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng*, 71(4):286–306.

Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, 15(3):296–303.

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? *Trends Biochem Sci*, 28(6):336–341.

Schultz, C., Niebisch, A., Gebel, L., and Bott, M. (2007). Glutamate production by *Corynebacterium glutamicum*: dependence on the oxoglutarate dehydrogenase inhibitor protein OdhI and protein kinase PknG. *Appl Microbiol Biotechnol*, 76(3):691–700.

Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60.

Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2002). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361.

Schwartz, J. M. and Kanehisa, M. (2005). A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics*, 21 Suppl 2:ii204–205.

Schwarz, R., Musch, P., von Kamp, A., Engels, B., Schirmer, H., Schuster, S., and Dandekar, T. (2005). YANA - a software tool for analyzing flux modes, gene-expression and enzyme activities. *BMC Bioinformatics*, 6:135.

Shimada, H., Kondo, K., Fraser, P. D., Miura, Y., Saito, T., and Misawa, N. (1998). Increased carotenoid production by the food yeast *Candida utilis* through metabolic engineering of the isoprenoid pathway. *Appl Environ Microbiol*, 64(7):2676–2680.

Shlomi, T., Berkman, O., and Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21):7695–7700.

Silveira, M. M. and Jonas, R. (2002). The biotechnological production of sorbitol. *Appl Microbiol Biotechnol*, 59(4-5):400–408.

Soetaert, W. and Vandamme, E. (2006). The impact of industrial biotechnology. *Biotechnol J*, 1(7-8):756–769.

Sonderegger, M., Jeppsson, M., Larsson, C., Gorwa-Grauslund, M. F., Boles, E., Olsson, L., Spencer-Martins, I., Hahn-Hagerdal, B., and Sauer, U. (2004). Fermentation performance of engineered and evolved xylose-fermenting *Saccharomyces cerevisiae* strains. *Biotechnol Bioeng*, 87(1):90–98.

Song, H. S. and Ramkrishna, D. (2009). Reduction of a set of elementary modes using yield analysis. *Biotechnol Bioeng*, 102(2):554–568.

Song, X. Y., Chen, Q. H., Ruan, H., He, G. Q., and Xu, Q. (2006). Synthesis and paste properties of octenyl succinic anhydride modified early Indica rice starch. *J Zhejiang Univ Sci B*, 7(10):800–805.

Sprenger, G. A. (2007). From scratch to value: engineering *Escherichia coli* wild type cells to the production of L-phenylalanine and other fine chemicals derived from chorismate. *Appl Microbiol Biotechnol*, 75(4):739–749.

Steinbuchel, A. and Muller, M. (1986). Glycerol, a metabolic end product of Trichomonas vaginalis and Tritrichomonas foetus. *Mol Biochem Parasitol*, 20(1):45–55.

Stephanopoulos, G. (1999). Metabolic fluxes and metabolic engineering. *Metab Eng*, 1(1):1–11.

Stephanopoulos, G. and Sinskey, A. J. (1993). Metabolic engineering–methodologies and future prospects. *Trends Biotechnol*, 11(9):392–396.

Stolz, M., Peters-Wendisch, P., Etterich, H., Gerharz, T., Faurie, R., Sahm, H., Fersterra, H., and Eggeling, L. (2007). Reduced folate supply as a key to enhanced L-serine production by *Corynebacterium glutamicum*. *Appl Environ Microbiol*, 73(3):750–755.

Tantirungkij, M., Nakashima, N., Seki, T., and Yoshida, T. (1993). Construction of xylose-assimilating *Saccharomyces cerevisiae*. *J Ferment Bioeng*, 75:83–88.

Tatarko, M. and Romeo, T. (2001). Disruption of a global regulatory gene to enhance central carbon flux into phenylalanine biosynthesis in *Escherichia coli*. *Curr Microbiol*, 43(1):26–32.

Team, R. D. C. (2009). A language and Environment for statistical computing.

Terzer, M. and Stelling, J. (2008). Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, 24(19):2229–2235.

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., van der Weijden, C. C., Schepper, M., Walsh, M. C., Bakker, B. M., van Dam, K., Westerhoff, H. V., and Snoep, J. L. (2000). Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem*, 267(17):5313–5329.

Toivari, M. H., Aristidou, A., Ruohonen, L., and Penttila, M. (2001). Conversion of xylose to ethanol by recombinant *Saccharomyces cerevisiae*: importance of xylulokinase (XKS1) and oxygen availability. *Metab Eng*, 3(3):236–249.

Toivari, M. H., Salusjarvi, L., Ruohonen, L., and Penttila, M. (2004). Endogenous xylose pathway in *Saccharomyces cerevisiae*. *Appl Environ Microbiol*, 70(6):3681–3686.

Torres, N. V. and Voit, E. O. (2002). *Pathway Analysis and Optimization in Metabolic Engineering*. Cambridge University Press, New York.

Trinh, C. T., Carlson, R., Wlaschin, A., and Srienc, F. (2006). Design, construction and performance of the most efficient biomass producing *Escherichia coli* bacterium. *Metab Eng*, 8(6):628–638.

Trinh, C. T. and Srienc, F. (2009). Metabolic engineering of *Escherichia coli* for efficient conversion of glycerol to ethanol. *Appl Environ Microbiol*, 75(21):6696–6705.

Trinh, C. T., Unrean, P., and Srienc, F. (2008). Minimal *Escherichia coli* cell for the most efficient production of ethanol from hexoses and pentoses. *Appl Environ Microbiol*, 74(12):3634–3643.

Trinh, C. T., Wlaschin, A., and Srienc, F. (2009). Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl Microbiol Biotechnol*, 81(5):813–826.

Tyo, K. E., Alper, H. S., and Stephanopoulos, G. N. (2007). Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends Biotechnol*, 25(3):132–137.

Tyo, K. E., Kocharin, K., and Nielsen, J. (2010). Toward design-based engineering of industrial microbes. *Curr Opin Microbiol*, 13(3):255–262.

Unrean, P., Trinh, C. T., and Srienc, F. (2010). Rational design and construction of an efficient *Escherichia coli* for production of diapolycopendioic acid. *Metab Eng*, 12(2):112–122.

Utagawa, T. (2004). Production of arginine by fermentation. *J Nutr*, 134(10 Suppl):2854S–2857S; discussion 2895S.

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

Varma, A., Boesch, B. W., and Palsson, B. O. (1993). Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol*, 59(8):2465–2473.

Varma, A. and Palsson, B. O. (1993). Metabolic capabilities of *Escherichia coli*: I. synthesis of biosynthetic precursors and cofactors. *J Theor Biol*, 165(4):477–502.

Varma, A. and Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol*, 60(10):3724–3731.

Visser, D. and Heijnen, J. J. (2002). The mathematics of metabolic control analysis revisited. *Metab Eng*, 4(2):114–123.

Visser, D. and Heijnen, J. J. (2003). Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab Eng*, 5(3):164–176.

Visser, D., Schmid, J. W., Mauch, K., Reuss, M., and Heijnen, J. J. (2004). Optimal re-design of primary metabolism in *Escherichia coli* using linlog kinetics. *Metab Eng*, 6(4):378–390.

von Kamp, A. and Schuster, S. (2006). Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics*, 22(15):1930–1931.

Wahlbom, C. F., van Zyl, W. H., Jonsson, L. J., Hahn-Hagerdal, B., and Otero, R. R. (2003). Generation of the improved recombinant xylose-utilizing *Saccharomyces cerevisiae* TMB 3400 by random mutagenesis and physiological comparison with *Pichia stipitis* CBS 6054. *FEMS Yeast Res*, 3(3):319–326.

Wegman, M., Janssen, M., van Rantwijk, F., and Sheldon, R. (2001). Towards biocatalytic synthesis of b-lactam antibiotics. *Adv Synth Catal*, 343:559–576.

Werpy, T. and Petersen, G. (2004). Top value added chemicals from biomass, vol.I. Results of screening for potential candidates from sugars and synthesis gas. Technical report, U.S. Department of Energy, Washington, DC. http://dx.doi.org/10.2172/15008859.

Wiechert, W. (2002). Modeling and simulation: tools for metabolic engineering. *J Biotechnol*, 94(1):37–63.

Willke, T. and Vorlop, K. D. (2004). Industrial bioconversion of renewable resources as an alternative to conventional chemistry. *Appl Microbiol Biotechnol*, 66(2):131–142.

Wisselink, H. W., Moers, A. P. H. A., Mars, A. E., Hoefnagel, M. H. N., de Vos, W. M., and Hugenholtz, J. (2004). Mannitol-1-phospatase: A key enzyme in mannitol production by *Lactococcus lactis*. *Applied and Environmental Microbiology*, 71:1507–1514.

Wittmann, C., Kiefer, P., and Zelder, O. (2004). Metabolic fluxes in *Corynebacterium glutamicum* during lysine production with sucrose as carbon source. *Appl Environ Microbiol*, 70(12):7277–7287.

Yakandawala, N., Romeo, T., Friesen, A. D., and Madhyastha, S. (2008). Metabolic engineering of *Escherichia coli* to enhance phenylalanine production. *Appl Microbiol Biotechnol*, 78(2):283–291.

Yanofsky, C. (1981). Attenuation in the control of expression of bacterial operons. *Nature*, 289(5800):751–758.

Yu, C., Cao, Y., Zou, H., and Xian, M. (2011). Metabolic engineering of *Escherichia coli* for biotechnological production of high-value organic acids and alcohols. *Appl Microbiol Biotechnol*, 89(3):573–583.

Zeikus, J. G., Jain, M. K., and Elankovan, P. (1999). Biotechnology of succinic acid production and markets for derived industrial products. *Appl Microbiol Biotechnol*, 51:545–552.

Zelle, R. M., de Hulster, E., van Winden, W. A., de Waard, P., Dijkema, C., Winkler, A. A., Geertman, J. M., van Dijken, J. P., Pronk, J. T., and van Maris, A. J. (2008). Malic acid production by *Saccharomyces cerevisiae*: engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. *Appl Environ Microbiol*, 74(9):2766–2777.

Zeng, A. P. and Biebl, H. (2002). Bulk chemicals from biotechnology: the case of 1,3-propanediol production and the new trends. *Adv Biochem Eng Biotechnol*, 74:239–259.

Appendices

Appendix A

The contents of Appendix A are in the media attached to the cover page.

Appendix B

The contents of Appendix B are in the media attached to the cover page.

Appendix C

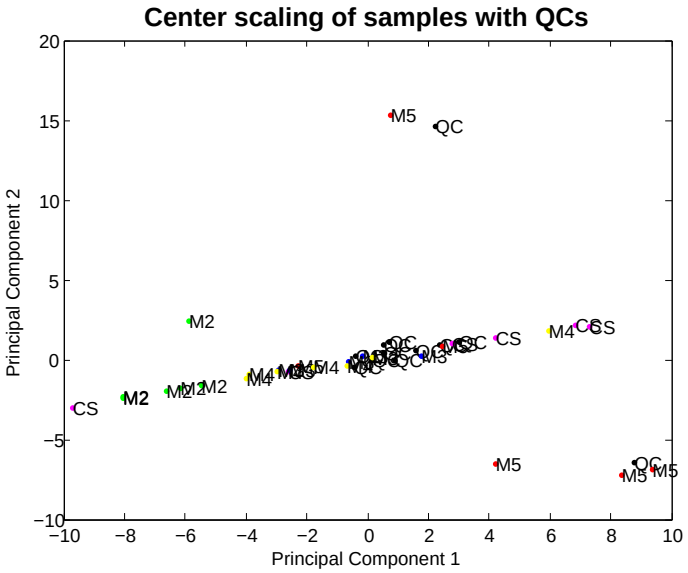

Figure C.1: **PC1 against PC2: Center scaled samples with QCs**. Figure depicts a two-dimensional plot of PC1 against PC2 of Center scaled samples (including QCs)

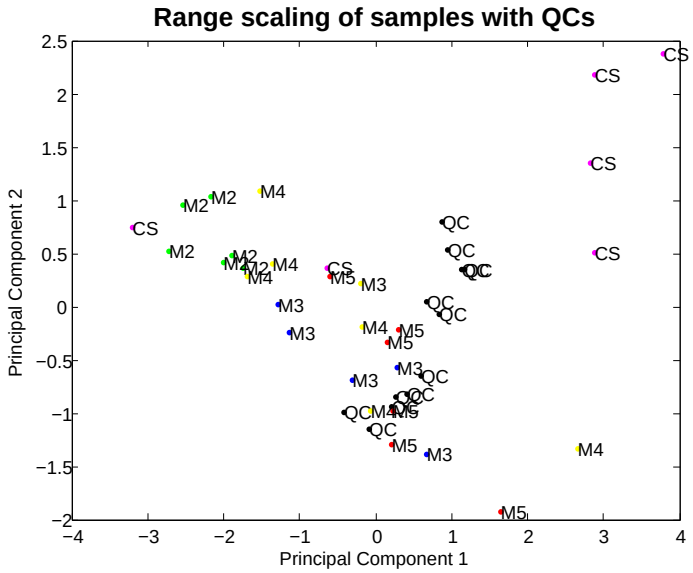

Figure C.2: **PC1 against PC2: Range scaled samples with QCs**. Figure depicts a a two-dimensional plot of PC1 against PC2 of Range scaled samples (including QCs)

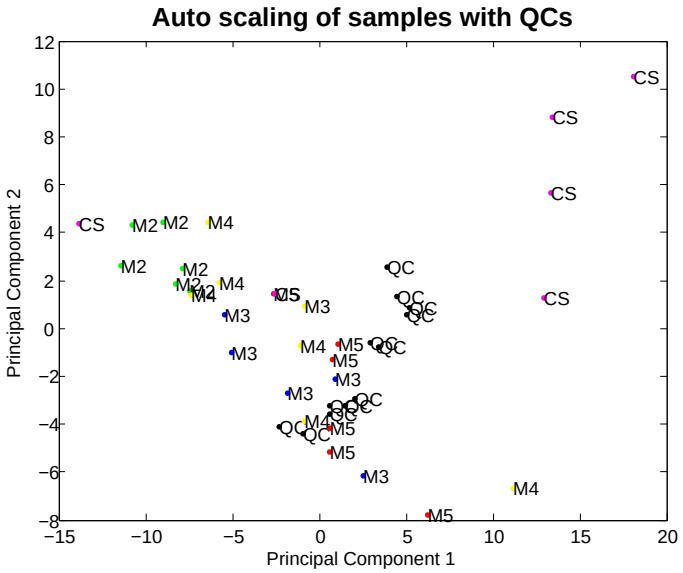

Figure C.3: **PC1 against PC2: Autoscaled samples with QCs**. Figure depicts a two-dimensional plot of PC1 against PC2 of Autoscaled samples (including QCs)

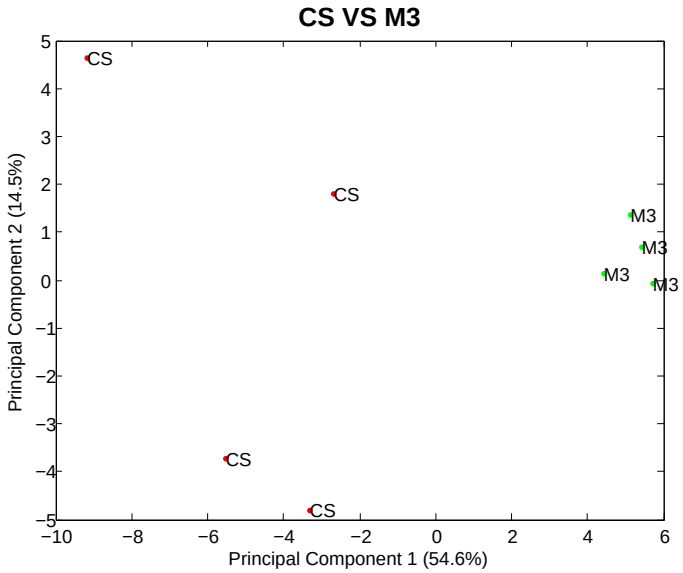

Figure C.4: **A plot of PC1 against PC2 for CS and M3**. Figure shows a two-dimensional plot of PC1 against PC2 for CS vs M3

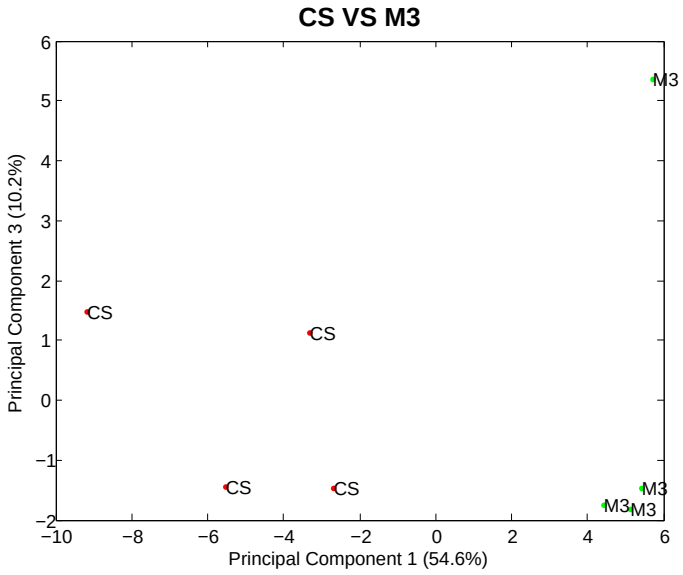

Figure C.5: **A plot of PC1 against PC2 for CS and M3**. Figure shows a two-dimensional plot of PC1 against PC3 for CS and M3

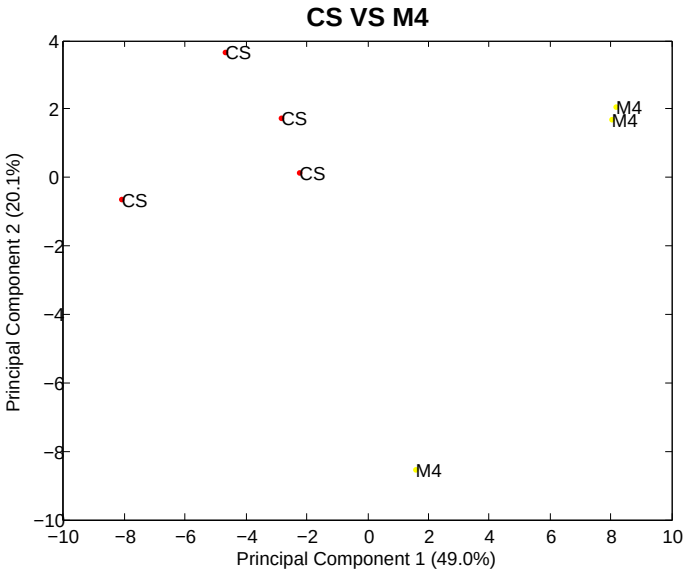

Figure C.6: **A plot of PC1 against PC2 for CS and M4**. Figure shows a two-dimensional plot of PC1 against PC2 for CS vs M4

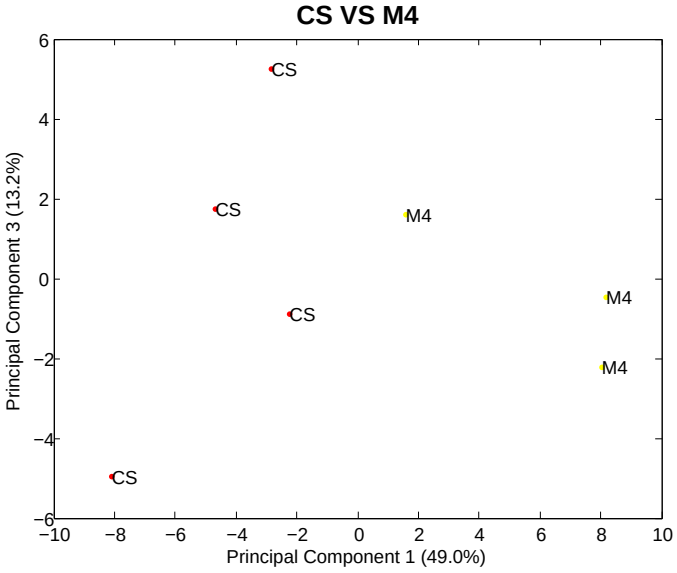

Figure C.7: **A plot of PC1 against PC2 for CS and M4**. Figure shows a two-dimensional plot of PC1 against PC3 for CS and M4

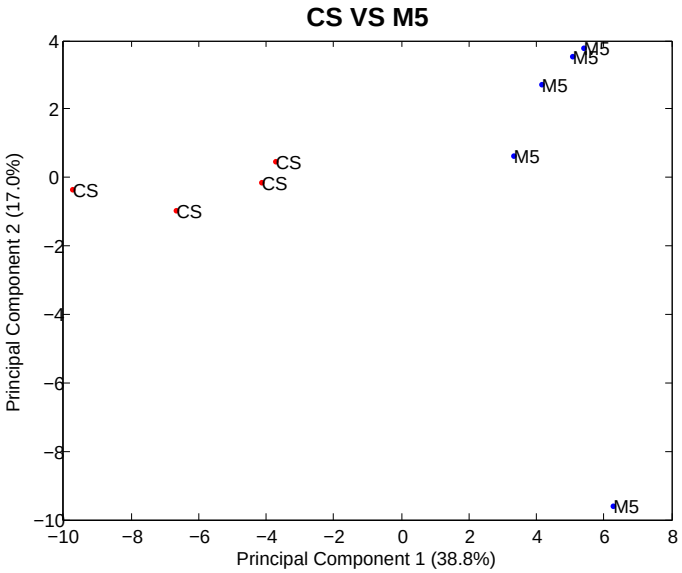

Figure C.8: **A plot of PC1 against PC2 for CS and M5**. Figure shows a two-dimensional plot of PC1 against PC2 for CS VS mutant M5

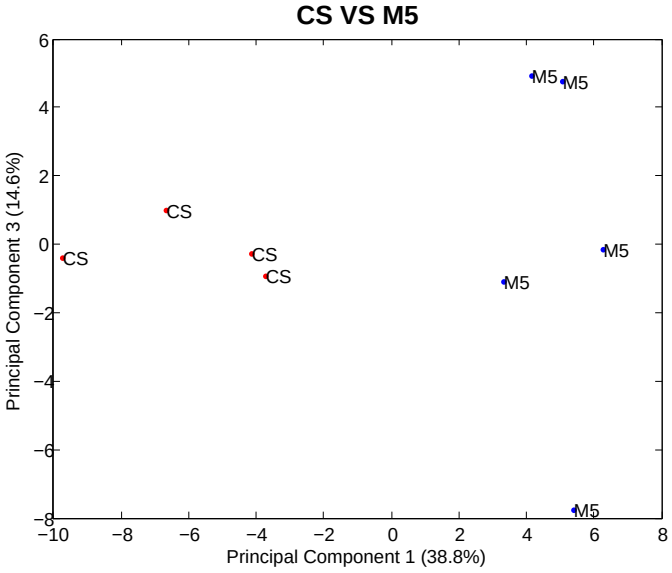

Figure C.9: **A plot of PC1 against PC2 for CS and M5**. Figure shows a two-dimensional plot of PC1 against PC3 for CS and M5