Metabolic Engineering for Substrate Co-utilization

by

Pratish Gawand

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Chemical Engineering and Applied Chemistry
University of Toronto

© Copyright 2014 by Pratish Gawand
Abstract

Metabolic Engineering for Substrate Co-utilization

Pratish Gawand

Doctor of Philosophy

Graduate Department of Chemical Engineering and Applied Chemistry

University of Toronto

2014

Production of biofuels and bio-based chemicals is being increasingly pursued by chemical industry to reduce its dependence on petroleum. Lignocellulosic biomass (LCB) is an abundant source of sugars that can be used for producing biofuels and bio-based chemicals using fermentation. Hydrolysis of LCB results in a mixture of sugars mainly composed of glucose and xylose. Fermentation of such a sugar mixture presents multiple technical challenges at industrial scale. Most industrial microorganisms utilize sugars in a sequential manner due to the regulatory phenomenon of carbon catabolite repression (CCR). Due to sequential utilization of sugars, the LCB-based fermentation processes suffer low productivities and complicated operation. Performance of fermentation processes can be improved by metabolic engineering of microorganisms to obtain superior characteristics such as high product yield. With increased computational power and availability of complete genomes of microorganisms, use of model-based metabolic engineering is now a common practice. The problem of sequential sugar utilization, however, is a regulatory problem, and metabolic models have never been used to solve such regulatory problems. The focus of this thesis is to use model-guided metabolic engineering to construct industrial strains capable of co-utilizing sugars. First, we develop a novel bilevel optimization algorithm SimUp, that uses metabolic models to identify reaction deletion strategies to force co-utilization of two sugars. We then use SimUp to identify reaction deletion strategies to force glucose-xylose co-utilization in Escherichia coli. To validate SimUp predictions, we construct three mutants with multiple gene knockouts and test them for glucose-xylose utilization characteristics. Two mutants, designated as LMSE2 and LMSE5, are shown to co-utilize glucose and xylose in agreement with SimUp predictions. To understand the molecular mechanism involved in glucose-xylose co-utilization of the mutant LMSE2, the mutant is subjected to targeted and
whole genome sequencing. Finally, we use the mutant LMSE$_2$ to produce D-ribose from a mixture of glucose and xylose by overexpressing an endogenous phosphatase. The methods developed in this thesis are anticipated to provide a novel approach to solve sugar co-utilization problem in industrial microorganisms, and provide insights into microbial response to forced co-utilization of sugars.
I would like to express my sincere gratitude towards everyone who has been a part of my Doctoral program at the University of Toronto. First and foremost, my supervisor Professor Mahadevan has been a source of constant support and guidance. His excitement towards scientific inquiry and deep insights in the field of metabolic engineering have always been an inspiration to me. I also thank the members of my reading committee. Professor Master kindled my curiosity towards the fundamental aspects of my project. Professor Saville motivated me to constantly think about the applications of my research to the real world problems. I have tried my best to abide by the committee’s suggestions and I hope it is reflected in my final thesis.

I also thank my colleagues in the Laboratory of Metabolic and Systems Engineering and BioZone. Laurance Yang’s help with the modeling studies, and Patrick Hyland’s help with the experimental studies were critical for my graduate research. I would also like to thank Siavash Partow for his help in sequencing studies and Cleo Ho for discussions on bioinformatics analysis. Fahimeh Salimi and Sarat Chandra with their frank opinions have been valuable friends throughout my PhD. Endang Susilawati from BioZone has been ever helpful in smooth running of my experiments. I would also like to thank all the undergraduate students who contributed to my thesis with their help in experiments. In particular, Echo Da Zhang, Amanda Giang, Jeffrey Chu, and Fatumina Saidabukar made solid contributions to my thesis. The whole graduate research experience would not have been as pleasant without the company of my friends Victor Balderas, Nikolaos Anesiadis, Nadeera Jayasinghe, Kai Zhuang, Kevin Corriea, Vik Pandit, Naveen Venayak, Kayla Nemr, Chris Gowen, Srinath Garg, and Jeong Chan Joo.

Our collaborators outside the University of Toronto, Professor Vincent Martin and Andrew Ekins at Concordia University, Montreal, taught me the basics of molecular biology, and introduced me to the excitement of genetic engineering. I would always be thankful to them for allowing me to spend time in their lab. I would also like to thank my funding agencies NSERC Bioconversion Network and ABIP for the financial support for my graduate work.
Finally, I would like to thank my family and friends in India who have always shown tremendous confidence in me.
## Contents

1 Introduction 1

1.1 Motivation 1

1.2 Objectives 3

1.3 Contributions 4

1.3.1 SimUp: a bilevel optimization algorithm 4

1.3.2 Co-utilizing mutants 5

1.3.3 Sequencing studies 6

1.3.4 D-Ribose production 6

1.3.5 Additional contributions 7

1.4 Publication summaries 8

2 Literature Review 13

2.1 Sugar transport in industrial microorganisms 13

2.1.1 Sugar transport in *E. coli* and carbon catabolite repression 15

2.1.2 Sugar transport in *Saccharomyces cerevisiae* 20

2.1.3 Sugar transport and regulation in other microorganisms 22

2.1.4 Glucose-xylose co-utilization in *E. coli* 24

2.1.5 Opportunities for advancement 27

2.2 Mathematical modeling in metabolic engineering 29

2.2.1 Constraint-based modeling basics 29

2.2.2 Alternatives to FBA 32

2.2.3 Constraints for improved CBM predictions 34
2.3 Model-based strain design ........................................... 35
  2.3.1 Opportunities for advancement ................................... 37
2.4 Summary of the literature review .................................. 39
  2.4.1 Sugar utilization .................................................. 39
  2.4.2 Metabolic modeling for strain engineering ..................... 40
2.5 Hypotheses, objectives, and organization of the thesis .......... 41
2.6 Unifying theme of the thesis ....................................... 43

3 SimUp: a Bilevel Optimization Algorithm .......................... 45
  3.1 Abstract .......................................................... 45
  3.2 Introduction ....................................................... 46
  3.3 Logic of SimUp ..................................................... 48
  3.4 Materials and Methods ............................................ 48
    3.4.1 SimUp algorithm formulation ................................ 48
    3.4.2 Solution to the SimUp algorithm ............................ 51
    3.4.3 Stoichiometric feasibility analysis ......................... 57
  3.5 Results .......................................................... 58
    3.5.1 Strategy A and the mutant LMSE_{1} ....................... 60
    3.5.2 Strategy B and the mutant LMSE_{2} ....................... 61
    3.5.3 Strategy C and the mutant LMSE_{5} ....................... 62
    3.5.4 Feasibility analysis of biomass precursors ................. 64
  3.6 Conclusions ..................................................... 65

4 Glucose and Xylose Co-utilizing Mutants of *Escherichia coli* ... 66
  4.1 Abstract .......................................................... 66
  4.2 Introduction ....................................................... 67
  4.3 Materials and Methods ............................................ 69
    4.3.1 Strains, media, and plasmids ................................ 69
    4.3.2 Phenotypic characterization ................................ 69
    4.3.3 Batch cultivation ............................................ 70
    4.3.4 Metabolite analysis ......................................... 71
5 Genome Sequencing of the *Escherichia coli* Mutant LMSE$_2$

5.1 Abstract ................................................................. 79
5.2 Introduction .............................................................. 80
5.3 Materials and Methods .................................................. 82
   5.3.1 Sanger sequencing .................................................. 82
   5.3.2 Whole genome sequencing ........................................... 82
   5.3.3 Complementation studies ......................................... 83
5.4 Results ................................................................. 84
   5.4.1 Sanger sequencing results ......................................... 84
   5.4.2 Whole genome sequencing ........................................... 86
   5.4.3 Characterization of icdA* ........................................ 89
5.5 Conclusions ........................................................... 90

6 D-Ribose Production Using the *Escherichia coli* Mutant LMSE$_2$

6.1 Abstract ................................................................. 91
6.2 Introduction .............................................................. 92
6.3 Materials and Methods .................................................. 95
   6.3.1 FBA and robustness analysis ...................................... 95
   6.3.2 Strains and plasmids ............................................... 96
   6.3.3 Phenotypic tests ................................................... 96
   6.3.4 Media and batch cultivation .................................... 96
   6.3.5 Fed-batch cultivation ............................................ 97
6.4 Results ................................................................. 97
   6.4.1 FBA and robustness analysis ...................................... 97
   6.4.2 Phenotypic characterization ..................................... 98
   6.4.3 Batch studies ..................................................... 99
G Coverage of the Deleted Regions in the Mutant LMSE$_2$
List of Tables

1.1 Global markets for some bio-based chemicals, 2007-2011 (in million lbs) . . . . . . 2

2.1 Hexose transporters in S. cerevisiae .................................................. 21

2.2 Glucose-xylose co-utilizing E. coli mutants. ........................................ 26

2.3 Selected strain design algorithms published. ...................................... 38

3.1 The solutions predicted by the SimUp algorithm. ............................... 59

4.1 List of the mutants constructed in this study. .................................... 70

4.2 Growth characteristics of the constructed mutants. ........................... 77

5.1 Genes of the mutant LMSE\textsubscript{2} sequenced using Sanger sequencing. . . . . . . . . . . . . . . . . . . . 82

5.2 The SNPs identified using whole genome sequencing. .......................... 88

6.1 D-Ribose producing strains. ............................................................... 93

6.2 Phenotypes of the mutants constructed in this study. .......................... 99

6.3 Carbon balance for the mutants RB-003 and RB-006. ........................ 102

6.4 D-Ribose production characteristics by the RB-series mutants. ............ 103

A.1 Growth rates of the glucose-6-phosphate node mutants. .................... 123

A.2 Growth characteristics of \textit{pgi} and \textit{aceA} mutants ........................ 124

A.3 FBA and MOMA simulations for the higher-order mutants. .................. 126

B.1 Files used to solve SimUp. ............................................................. 128

B.2 Parameter values used for SimUp. .................................................. 130

C.1 Primers. ......................................................................................... 131
D.1 Mutants and plasmids used in this thesis. .............................. 134
List of Figures

2.1 Glucose transport system and the CCR phenomenon in *E. coli*. Two major mechanisms involved in CCR, inducer exclusion and global regulatory mechanism, are shown. In *E. coli*, the protein EIIA$^{Glc}$ plays a pivotal role in CCR. In its phosphorylated state, EIIA$^{Glc}$ activates adenylate cyclase (AC) which converts AMP into cAMP. High concentration of cAMP results in formation of cAMP-CRP complex, which binds to the promoters and activates the expression of genes involved in transport and catabolism of secondary sugars such as xylose. This mechanism is known as global regulatory mechanism. In its non-phosphorylated state, EIIA$^{Glc}$ binds to the proteins such as XylE and others involved in transport and catabolism of secondary sugars such as xylose. This mechanism is known as inducer exclusion. The phosphorylation state of EIIA$^{Glc}$ is determined by the intracellular ratio of PEP to pyruvate. The figure has been modified and redrawn from Görke and Stülke (2008a).

2.2 Schematic representation of activation on xyl system in presence of different sugars. a) In presence of xylose and absence of glucose, xylAB is activated by binding of XylR-xylose complex, RNA polymerase (RNAP), and CRP-cAMP complex, b) in presence of arabinose, the AraC-arabinose complex binds to the promoter $P_{xylA}$ thereby inhibiting the binding of XylR-xylose complex and repressing the expression of the operon xylAB, c) in presence of glucose, the intracellular CRP-cAMP levels are low and XylR-xylose complex is absent causing no expression of xylAB operon.
3.1 Schematic of the bilevel optimization problem for SimUp. Three different optimization problems, each simulating growth in presence of different sugar conditions, are nested in an outer problem. The outer problem maximizes the growth on glucose and xylose mixture while penalizing the growth on individual sugars. The inner problems simulate the growth of the organism under three different media conditions: 1) growth on glucose and xylose, 2) growth on glucose alone, and 3) growth on xylose alone. To maximize the outer objective function, growth is desired only under condition 1 (represented in green), and not under conditions 2 and 3 (represented in red).

3.2 a) The solutions belonging to strategy A, and b) the metabolic map of the mutant LMSE$_1$. The highlighted solution (in red) was used to construct the mutant LMSE$_1$.

3.3 a) The solutions belonging to strategy B, and b) the metabolic map of the mutant LMSE$_2$. The highlighted solution (in red) was used to construct the mutant LMSE$_2$.

3.4 a) The solutions belonging to strategy C, and b) the metabolic map of the mutant LMSE$_5$. The highlighted solution (in red) was used to construct the mutant LMSE$_5$.

3.5 Stoichiometric feasibilities for the mutants LMSE$_2$ and LMSE$_5$. G – glucose as carbon source; X – xylose as carbon source; GX – glucose and xylose as carbon source. a) Feasibilities of biomass precursors for the mutant LMSE$_2$, b) feasibilities of biomass precursors for the mutant LMSE$_5$, c) feasibilities for all the biomass precursors with additional secretion reaction for F6P in the metabolic network of the mutant LMSE$_5$. F6P – fructose-6-phosphate, G3P – glyceraldehyde-3-phosphate, 3PG – 3-phosphoglycerate, PEP – phosphoenolpyruvate, PYR – pyruvate, R5P – ribose-5-phosphate, E4P – erythrose-4-phosphate, OA – oxaloacetate, AKG – α-ketoglutarate.
4.1 Phenotypic characterization of the LMSE mutants. a) Shows the growth characteristics on seven different media combinations including LB and minimal medium, b) shows the limited growth phenotype of the mutant LMSE_5.

4.2 Batch cultivation studies of the LMSE mutants on glucose and xylose. a) Growth characteristics of the wild-type *E. coli*, b) growth characteristics of the mutant LMSE_1, c) growth characteristics of the mutant LMSE_2, d) growth characteristics of the mutant LMSE_5. The symbol, $\bullet$ represents biomass expressed as dry cell weight, $\Diamond$ represents glucose concentration, and $\blacktriangle$ represents xylose concentration. The error bars represent standard deviation of three replicates.

4.3 Batch cultivation studies on the mutants LMSE_2 with different ratios of glucose and xylose. a) Mutant LMSE_2 on 6.7 g/L glucose and 3.3 g/L xylose, b) mutant LMSE_2 on 3.3 g/L glucose and 6.7 g/L xylose. The symbol, $\bullet$ represents biomass, $\Diamond$ represents glucose concentration, and $\blacktriangle$ represents xylose concentration. The error bars represent standard deviation of two replicates.

5.1 a) Schematic showing the position of the SNP in the gene *ptsG* of the mutant LMSE_2, b) growth characteristics of the mutant *E. coli ΔptsG*, c) growth characteristics of the mutant SQ-002. The symbol, $\bullet$ represents biomass, $\Diamond$ represents glucose concentration, and $\blacktriangle$ represents xylose concentration.

5.2 Complementation studies on the mutant SQ-003 for characterization of the mutant gene *icdA*$. The symbol, $\bullet$ represents biomass, $\Diamond$ represents glucose concentration, and $\blacktriangle$ represents xylose concentration.
6.1 a) Metabolic network showing modifications in the mutant LMSE$_2$ for production of D-ribose. Additional reactions, ybiV or yidA, have been added to the network showing phosphatase activity required for D-ribose production. The values beside the reaction names are the flux values expressed in mmol/gDW/h. The top values on the arrow (in black) are the wild-type fluxes and the bottom values (in grey) are the mutant LMSE$_2$ fluxes. Note that the directions of the arrows in the metabolic map represent the directions of the fluxes and not irreversibility of the reactions. The names of all the reactions and the metabolites in the figure can be found in the list of abbreviations provided in the beginning of this thesis. b) Robustness analysis of D-ribose production flux with respect to glucose exchange flux in the mutant LMSE$_2$, and c) robustness analysis of D-ribose production flux with respect to xylose exchange flux in the mutant LMSE$_2$.

6.2 Growth plots of RB-series mutants, a) RB-002, b) RB-003, c) RB-005, and 6) RB-006. The symbol, ● represents biomass, ◆ represents glucose concentration, ▲ represents xylose concentration, ■ represents D-ribose concentration, and ‡ represents acetate concentration. Error bars represents standard deviation of three replicates.

6.3 Growth profile for fed-batch culture of the mutant RB-006. The symbol, ● represents biomass, ◆ represents glucose concentration, ▲ represents xylose concentration, ■ represents D-ribose concentration, ‡ represents acetate concentration, and † represents xylulose concentration. Error bars represent standard deviation of three replicates.

A.1 Metabolic pathways of E. coli, a) under normal growth conditions, and b) after deletion of the gene pgi. The major flux during normal growth is through glycolysis with no flux through either the Entner-Doudoroff pathway or the glyoxylate shunt. Both these pathways get activated immediately after deletion of pgi, and major flux diversion from glycolysis to the PPP takes place.
A.2 Growth profiles of different single and double gene knockout mutants constructed in the this study. a) Glyoxylate shunt mutants, b) Entner-Doudoroff pathway mutants. The error bars represent standard deviation of three replicates. 122

A.3 Growth plots of two double mutants with different order of gene deletions, a) E. coli ∆pgi1 ∆aceA2, and b) E. coli ∆aceA1 ∆pgi2. The symbol, ● represents biomass, ◆ represents glucose concentration, and ■ represents acetate concentration. The error bars represent standard deviation of three replicates. 125

G.1 Coverage of the region encoding the gene pgi in the mutant LMSE2 and the wild-type. 142

G.2 Coverage of the region encoding the gene rpe in the mutant LMSE2 and the wild-type. 143
Chapter 1

Introduction

1.1 Motivation

Over the past century, chemical industry has evolved to depend heavily on petroleum-based feedstocks for production of fuels as well as commodity chemicals. However, recent socio-economic and environmental concerns, such as volatility of petroleum prices, and increase in the greenhouse gas emissions, are driving chemical industry to adopt renewable, bio-based feedstocks. Last decade witnessed an intensive research and development effort in the area of biofuels and bio-based chemicals. Though the development in the area of biofuels, mainly cellulosic ethanol, is considered to have somewhat stagnated in the recent years (Hayes, 2013), bio-based chemicals are, in contrast, being aggressively pursued. Several recent success stories have demonstrated the production of natural and non-natural chemicals using bio-based routes (Atsumi et al., 2008; Zhang et al., 2009; Yim et al., 2011; Kaur et al., 2012). Table 1.1 shows the recent trends and market projections for some bio-based chemicals that have already been integrated with the current petroleum-based supply-chain.

Despite the laboratory-scale success of bio-based processes, commercialization of these processes almost entirely depends on the cost-parity with their existing petroleum-based counterparts. However, bio-based processes are typically known to have higher capital and operating costs, and complex operation compared to the petroleum-based processes. For scale-up and commercialization of bio-based processes, three performance metrics are of prime importance: yield, titre, and productivity (Stephanopoulos, 2007; Van Dien, 2013). Titre of the product
Table 1.1: Global markets for some bio-based chemicals, 2007-2011 (in million lbs).

<table>
<thead>
<tr>
<th>Chemicals</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>CAGR(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>butanol</td>
<td>110.3</td>
<td>275.6</td>
<td>11.0</td>
<td>136.0</td>
<td>95.2</td>
<td>-3.6%</td>
</tr>
<tr>
<td>epichlorohydrin</td>
<td>2.6</td>
<td>12.5</td>
<td>29.5</td>
<td>40.5</td>
<td>52.5</td>
<td>112.0%</td>
</tr>
<tr>
<td>methanol</td>
<td>2.0</td>
<td>11.0</td>
<td>57.2</td>
<td>110.0</td>
<td>220.0</td>
<td>223.9%</td>
</tr>
<tr>
<td>ethylene glycol</td>
<td>296.5</td>
<td>179.5</td>
<td>142.1</td>
<td>245.5</td>
<td>433.8</td>
<td>10%</td>
</tr>
<tr>
<td>propanediol</td>
<td>22.1</td>
<td>49.6</td>
<td>74.4</td>
<td>84.3</td>
<td>97.2</td>
<td>44.8%</td>
</tr>
<tr>
<td>propylene glycol</td>
<td>18.2</td>
<td>5.7</td>
<td>39.7</td>
<td>92.6</td>
<td>135.1</td>
<td>65.1%</td>
</tr>
<tr>
<td>succinic acid</td>
<td>1.1</td>
<td>1.8</td>
<td>2.2</td>
<td>4.0</td>
<td>7.0</td>
<td>58.8%</td>
</tr>
<tr>
<td>others</td>
<td>3.0</td>
<td>5.0</td>
<td>7.0</td>
<td>9.0</td>
<td>10.0</td>
<td>35.1%</td>
</tr>
<tr>
<td>total</td>
<td>455.8</td>
<td>540.7</td>
<td>363.1</td>
<td>721.9</td>
<td>1050.8</td>
<td>23.2%</td>
</tr>
</tbody>
</table>

\(^a\) Compound Annual Growth Rate


plays a critical role in the downstream processing, yield is critical in determining the final cost of the product, and productivity is the principal determinant of the capital cost. Additional considerations such as strain robustness and by-product profile also play a major role in the feasibility of bio-based processes (Van Dien, 2013).

Metabolic engineering, the major theme of this thesis, is commonly used for improving the characteristics of the industrial microorganisms. Metabolic engineering differs from genetic engineering in its holistic approach towards metabolism in contrast to focus on individual genes (Alper and Stephanopoulos, 2009). Recent advances in the allied fields including high-throughput omics technologies, DNA synthesis, synthetic biology, bioinformatics tools, and protein engineering have greatly accelerated the development of the field of metabolic engineering (Lee et al., 2012). Using these technologies, bio-based production of non-natural chemicals such as 1,4-butanediol (Yim et al., 2011) and advanced biofuels such as farnesene (Peralta-Yahya et al., 2012) have already been achieved at commercial scale.

Model-based approaches can be used to aid metabolic engineering of industrial microorganisms. Algorithms that can suggest manipulations to the metabolic network, for production of
desired chemicals, have been formulated and successfully implemented to increase the yields of bio-based processes (Burgard et al., 2003). However, model-based approaches have limited scalability to genome-scale models. Additionally, comprehensive models that account for phenomena such as transcriptional regulation, enzyme kinetics, and allosteric regulation are not commonly used for devising engineering strategies, limiting the utility of the modeling-based approaches only to metabolic problems.

In this thesis, we extend the application of metabolic models to solve an important regulatory problem pertinent to the use of lignocellulosic biomass (LCB) as feedstock. Carbon catabolite repression (CCR) is a regulatory phenomenon found in many important industrial microorganisms (Yao and Shimizu, 2013). CCR is responsible for sequential utilization of sugars from mixture of sugars such as glucose and xylose, typically obtained from hydrolysis of biomass. The sequential utilization of sugars results in complex fermentation processes that are difficult to control, and also cause low overall productivity (Kim et al., 2010). Microorganisms without CCR, i.e., with a capability to simultaneously utilize multiple sugars, can benefit the bio-based chemical production by simplifying the fermentation processes, and by potentially increasing the productivity by reducing the batch-times. In this thesis, we demonstrate a novel application of metabolic models to solve the regulatory problem of CCR in the industrial microorganisms.

1.2 Objectives

The main challenge this thesis addresses is the regulatory problem of sequential sugar utilization caused due to presence of CCR. Though the composition of LCB hydrolysate varies based on the process used for production of the hydrolysate, two major components of LCB hydrolysate are glucose and xylose. Fermentation of such a mixture of sugars is a complex process due to the presence of CCR in industrial microorganisms. Metabolic engineering of microorganisms can be used to eliminate CCR, and force the organisms to consume two sugars simultaneously. However, engineering of microorganisms for elimination of CCR involves manipulations to the regulatory network controlling the sugar utilization. Previous attempts to achieve co-utilization phenotypes in industrial microorganisms have been largely
ad hoc, and have met with limited success. One of the basic objectives of this thesis is to develop a rational approach, based on genome-scale metabolic models, to design industrial microorganisms for simultaneous sugar utilization. Very few examples exist in the literature that report experimental validation of strains designed using model-guided approaches. Thus, this thesis aims to test the accuracy of model predictions by constructing and characterizing the model-predicted mutants.

The main premise of this thesis is that metabolic models can be used to engineer regulatory networks, without directly altering the regulatory genes. However, to alter its regulation, the organism may accumulate mutations in its regulatory network. Investigations on how the organism adapts to the forced phenotype (such as co-utilization of sugars) can provide insights into the fundamental response behaviour of the organism. With the objective of investigating the molecular mechanisms involved in achieving co-utilization phenotype, one of the mutants constructed in this study was intensively analyzed using whole-genome sequencing.

Finally, with an objective to demonstrate the utility of the sugar co-utilizing mutants, one of the glucose-xylose co-utilizing mutants was engineered to produce a valuable compound, D-ribose. This mutant was tested for D-ribose production on different sugar concentrations, and different fermentation modes such as batch and fed-batch.

1.3 Contributions

As outlined above, this thesis is concerned with metabolic engineering for sugar co-utilization with the objective of simplifying the fermentation processes based on LCB hydrolysates. Further engineering of a sugar co-utilizing strain for production of a valuable chemical, D-ribose, has also been explored. The following section describes the contributions of each chapter of this thesis.

1.3.1 SimUp: a bilevel optimization algorithm

SimUp is a bilevel optimization algorithm that can be used to identify metabolic gene knockout strategies to force organisms to co-utilize two sugars. SimUp algorithm’s novelty lies in two different aspects that differentiate it from previously published bilevel optimization algorithms: i)
SimUp solves the regulatory problem of sugar co-utilization using the knowledge of metabolism alone (without using any information on the regulatory network), and ii) SimUp employs simulation of different media conditions in a single optimization problem. The development SimUp algorithm is described in Chapter 3, and its experimental validation is described in Chapter 4 of this thesis. The work related to the SimUp algorithm and its validation has been published or presented in journals and conferences as listed below:


### 1.3.2 Co-utilizing mutants

A significant portion of the research in this thesis was focused on validating the SimUp algorithm and investigating whether a regulatory problem could be solved by manipulations to the metabolic network. Many multiple gene knockout mutants were constructed and characterized in this study based on SimUp predictions. Two mutants, designated as LMSE$_2$ and LMSE$_5$, showed co-utilization of glucose and xylose. These mutants were novel as they had no manipulations to their regulatory network, yet could co-utilize glucose and xylose. Additionally, the mutant LMSE$_5$ was of particular significance as this mutant was based on a strategy that could not be intuitively designed from the metabolic network connectivity. No such mutant had previously been reported. These studies are summarized in Chapter 4 of this thesis. The work on the LMSE-series mutants has been presented in conferences listed below:

- Gawand, P. and Mahadevan, R. Metabolic engineering strategies for substrate co-utilization. 2$^{nd}$ Annual General Meeting, NSERC Bioconversion Network, Toronto, June 6-8, 2011 (Oral presentation and poster presentation. Focus on the mutant LMSE$_5$).
Chapter 1. Introduction


1.3.3 Sequencing studies

The LMSE-series mutants constructed in this study had no manipulations to the regulatory genes involved in CCR. However, to utilize xylose along with glucose, the regulatory process of CCR had to be circumvented. The possible mechanism of glucose-xylose co-utilization by the mutant LMSE$_2$ was investigated using sequencing studies. Using Sanger sequencing of the genes involved in CCR, it was established that none of these genes was responsible for the co-utilization phenotype. Results from the Sanger sequencing studies were included in the manuscript submitted to Metabolic Engineering. Additionally, whole-genome sequencing of the mutant LMSE$_2$ was carried out to identify the genes (not involved in CCR), if any, that had possible roles in the co-utilization phenotype. The studies on whole-genome sequencing are unpublished as of yet. The sequencing results on the mutant LMSE$_2$ are summarized in Chapter 5 and have been presented as follows:

- Gawand, P. and Mahadevan, R. Sequencing of a glucose-xylose co-utilizing mutant of *Escherichia coli*. 3$^{rd}$ Annual General Meeting, NSERC Bioconversion Network, Vancouver, June 6-8, 2012 (Oral presentation and poster presentation. Focus on Sanger sequencing studies).

1.3.4 D-Ribose production

Previous attempts to detect D-ribose production in metabolic mutants of *E. coli* have been unsuccessful. The engineered mutant, RB-006, based on the mutant LMSE$_2$, is the first *E. coli* mutant that was shown to produce D-ribose. The work carried out on D-ribose production is
summarized in Chapter 6 of this thesis. The work related to D-ribose production using the mutant LMSE$_2$ has been published and presented as listed below:


### 1.3.5 Additional contributions

In addition to work listed above, studies were carried out during the course of this thesis that characterized a range of multiple-gene knockout mutants (higher-order mutants). Higher-order mutants show metabolic and regulatory responses that are not observable in single gene knockout mutants. Characterization of higher-order mutants can provide new insights into the metabolic behaviour, and the data can potentially be used to improve metabolic-models.

Appendix A describes the characterization and model simulations of some of the higher-order mutants constructed during this study. Most notably, we investigated the phenomenon of latent pathway activation and the effect of order of gene deletions on the phenotypes of the mutants. No studies have been previously reported that systematically investigate these phenomena experimentally. The results obtained from this form the basis of a manuscript. These results have also been presented at the conference listed below:

- Gawand, P. and Mahadevan, R. Higher order gene knockout mutants give insights into activation of latent reactions. Annual Meeting, AICHE, Minneapolis, September 17-21, 2011 (Oral presentation).

During the course of this thesis, the author has also contributed to a book chapter and a review article on the topics related to metabolic engineering which have been published as listed below:

Chapter 1. Introduction


Finally, in addition to the aforementioned contributions, this thesis also poses some open-ended questions such as the exact molecular mechanism of information relay in the cells which need to shift from single sugar utilization state to multiple sugar utilization state, and possibility of further engineering of the cells for utilization of different sugar ratios. All these future studies are summarized in Chapter 8 of this thesis.

1.4 Publication summaries

Summaries of all the manuscripts, published and in preparation, are provided in the following section.

1. **Original research article on SimUp and LMSE-series mutants (published)**


**Summary:** Economical production of lignocellulosic biofuels requires efficient utilization of two sugars, glucose and xylose. Microorganisms that are capable of co-utilizing glucose and xylose are of considerable interest to the biofuels industry. However, most industrial microorganisms preferentially utilize glucose over xylose owing to the regulatory phenomenon of carbon catabolite repression (CCR). Elimination of CCR in microorganisms is challenging due to the multiple coordinating mechanisms involved. This publication reports a novel algorithm, SimUp, which finds metabolic engineering strategies to force co-utilization of glucose and xylose, without targeting the regulatory pathways of CCR. Mutants of *E. coli* based on SimUp predictions showed predicted growth phenotypes and co-utilized glucose and xylose. Some of the solutions identified by the algorithm were
based on stoichiometric imbalance and were not obvious from the network topology.

2. Original research article on D-ribose production using the mutant LMSE$_2$ (published)


**Summary**: D-Ribose is a commercially important functional sugar used as a nutritional supplement and as a starting compound in synthesis of antiviral drugs. This publication reports engineered *E. coli* mutants that can produce D-ribose from glucose and xylose. Two endogenous haloacid dehalogenase-like (HAD) phosphatases from *E. coli*, HAD12 and HAD13, encoded by the genes *ybiV* and *yidA*, respectively, were expressed in *E. coli* wild-type and the glucose-xylose co-utilizing mutant LMSE$_2$. All the mutants constructed in this study produced D-ribose. The mutant RB-006 (LMSE$_2$ expressing *ybiV*) showed the highest D-ribose titre of 1.16 g/L from 5 g/L each of glucose and xylose. Using xylose feeding, D-ribose titre was improved to 3.36 g/L. Xylulose and acetate were formed as the major by-products in the fed-batch study. This study represents the first example of engineered *E. coli* for production of D-ribose. It also demonstrates reengineering of a glucose-xylose co-utilizing mutant for production of a valuable chemical.

3. Book chapter on metabolic model refinement (published)


**Summary**: Phenotypic microarrays are standardized high-throughput technology for profiling phenotypes of microorganisms, which allows for characterization of around 2,000
different media conditions. The data generated using phenotypic microarrays can be incorporated into genome-scale metabolic models to improve their predictive capability. In addition, a comparison of phenotype profiles of wild-type and gene knockout mutants can give essential information about the functions of unknown genes and can be used for gene annotation. In this book chapter, a protocol to refine preconstructed metabolic models using the phenotypic microarray data has been described. Both manual refinement and algorithmic approaches for integrating the phenotypic microarray data have been discussed.

4. Review article on strain improvement using high-throughput technologies (published)


**Summary:** To serve as the biocatalyst of choice for a viable lignocellulose-based bioethanol industry, *Saccharomyces cerevisiae* requires extensive metabolic reprogramming for enhanced capabilities. These capabilities include increased tolerance to fermentation inhibitors found in lignocellulosic hydrolysates, pentose fermentation pathways, and potential expression of cellulase activity while maintaining the industrial productivity levels. Engineering of these complex traits can be facilitated by an in-depth understanding of the *S. cerevisiae* cellular system as a whole. Such knowledge is being generated by taking an integrated systems approach using current and emerging ‘omics’ tools. These technologies have already enhanced the understanding of *S. cerevisiae* physiology, identified novel targets for engineering of *S. cerevisiae* strains, and provided the data necessary for elaborate metabolic modeling. This review discusses the high-throughput ‘omics’ technologies that are anticipated to greatly aid the future strain development for lignocellulosic biomass to bioethanol process.
5. **Original research article on molecular mechanism of sugar co-utilization in the mutant LMSE$_2$ (in preparation)**

Gawand, P., Partow, S., Bettenbrock, K., and Mahadevan, R. Mechanism of glucose-xylose co-utilization without alteration to the regulatory network of carbon catabolite repression.

**Summary:** The *E. coli* mutant LMSE$_2$ co-utilizes glucose and xylose due to the deletion of the three metabolic genes: *pgi*, *rpe*, and *eda*. The regulatory network responsible for carbon catabolite repression (CCR) has not been externally altered in the mutant. This publication reports the investigation of CCR mechanism of the mutant LMSE$_2$ using whole genome sequencing. Next-generation Illumina Sequencing was used for sequencing the mutant’s genome. The mutated genes were further characterized by complementation studies. The tested genes were not found responsible for the glucose-xylose co-utilization phenotype of the mutant LMSE$_2$. The results suggest that the mutant most likely co-utilizes the two sugars by altering its enzyme expression levels or phosphorylation state of the regulatory proteins, without accumulating any significant mutations in the CCR network. The studies on phosphorylation state of the protein EIIA$_{Glc}$ will be included in this manuscript.

**Note:** This publication will contain results from the experiments not discussed in this thesis. These results include the studies on phosphorylation state of the protein EIIA$_{Glc}$.

6. **Original research article on effect of different order of gene deletion (in preparation)**

**Summary:** Metabolic networks are characterized by multiple redundant pathways that do not have a clear biological function. The redundancies in the metabolic networks are implicated in adaptation to random mutations and/or survival under different environmental conditions. Characterization of multiple gene knockout mutants (higher-order mutants) can uncover metabolic phenomena that are not observed in single gene knockout mutants, and can potentially aid in the identification of the roles of redundant metabolic pathways. This study reports the effects of deletion of the glyoxylate-shunt gene aceA on the growth-rate of the mutant *E. coli* ∆pgi. This publication will report the investigation of the effect of order of gene deletions on the sub-optimal growth phase by detailed characterization of two mutants: *E. coli* ∆pgi\textsuperscript{1} ∆aceA\textsuperscript{2} and *E. coli* ∆aceA\textsuperscript{1} ∆pgi\textsuperscript{2}. Additional, characterization of the mutants using RNA-Seq will be included in this manuscript.

**Note:** This research article will contain experimental results from the experiments not discussed in this thesis. These results include RNA-Seq characterization studies of the higher order mutants.
Chapter 2

Literature Review

This chapter summarizes the current literature on two topics relevant to this thesis: 1) sugar transport in industrial microorganisms with emphasis on Escherichia coli, and 2) use of constraint-based modeling methods in metabolic engineering. At the end of the literature review, hypotheses for the thesis are developed and the unifying theme of the thesis is discussed.

2.1 Sugar transport in industrial microorganisms

Lignocellulosic biomass (LCB) is an abundant and renewable source of sugars that can be used for production of bio-based fuels and chemicals. However, despite over a decade of intensive research, LCB has not replaced corn and sugarcane as the primary feedstocks in the biofuels industry; and contrary to all predictions, cellulosic ethanol still remains a commercial challenge (Hayes, 2013). The sluggish technology transfer of LCB-based processes is mainly due to their capital intensive nature and complex operation. For example, hydrolysis of LCB is significantly more complicated than hydrolysis of corn-based starch due to the presence of five different sugars and lignin (Mabee and Saddler, 2010). In addition, the by-products formed during the pretreatment stage of LCB hydrolysis inhibit the fermentation process resulting in reduced yields and productivities. Current research on LCB-based processes focuses on multiple objectives such as eliminating the need for detoxification of hydrolysates, engineering robust strains with high tolerance to inhibitors, capability to ferment glucose and xylose simultaneously, reducing water usage, and increasing product yields and titres (Geddes et al., 2011). Among
other requirements, process simplification is an essential criterion for commercial success of LCB-based processes.

Glucose and xylose are the two major constituent sugars of the LCB hydrolysate, irrespective of the technology used for hydrolysis (Xiros et al., 2013). The importance of efficient utilization of glucose and xylose has been emphasized consistently in the literature (Buschke et al., 2013; Yao and Shimizu, 2013; Kim et al., 2012a; Stephanopoulos, 2007). Sugar utilization patterns influence the productivity and ease of operation of the LCB-based fermentation process (Kim et al., 2012a). To achieve the best yields and high productivities, complete and efficient fermentation of all the sugars is required. Some industrial microorganisms such as *E. coli* are capable of utilizing all the sugars obtained from LCB, however, these sugars are utilized in a sequential manner. Such sequential utilization of sugars causes multiple complications in the fermentation process.

The challenges commonly observed in mixed sugar fermentations are as follows:

1. sequential utilization of sugar reduces productivities of the target products (Nichols et al., 2001),

2. secondary sugars such as xylose accumulate till the end of the batch and inhibitory product accumulation at this stage reduces the utilization rates of these sugars (Kim et al., 2010),

3. sequential sugar utilization causes frequent changes in growth rates, which makes fermentation process difficult to predict and control,

4. secondary sugars such as xylose may not be recognized as fermentable sugars by microorganisms resulting in reduced yields of the fermentation products (Jin et al., 2004), and

5. sequential sugar utilization complicates the fed-batch mode of fermentation with mixed sugar feeds.

Due to the above reasons, organisms capable of co-utilizing sugars, especially glucose and xylose, are very valuable in industrial fermentation processes. Metabolic engineering techniques
are commonly employed to improve the performance of industrial microorganisms. Engineering efforts have been made towards improving sugar utilization patterns in industrial organisms such as *E. coli* and *S. cerevisiae* (Yao and Shimizu, 2013; Ikeda, 2012; Madhavan et al., 2012).

### 2.1.1 Sugar transport in *E. coli* and carbon catabolite repression

*E. coli* is one of the most well-studied model organisms and finds extensive application as an industrial host (Huffer et al., 2012; Vickers et al., 2012). One of the reasons for choice of *E. coli* as the host organism is its capability to utilize a wide range of sugars and sugar alcohols as a carbon source (Huffer et al., 2012). However, like most other microorganisms, *E. coli* shows a hierarchical utilization of sugars. Glucose is *E. coli*’s most preferred carbon source. In presence of glucose, transport and metabolism of any other sugar is suppressed. *E. coli* utilizes the secondary sugars only after glucose is completely exhausted. This phenomenon of sequential utilization of sugars is known as glucose repression, or more generally, as carbon catabolite repression (CCR). Görke and Stülke (2008a) define CCR as, “a regulatory phenomenon by which the expression of functions for the use of secondary carbon sources and the activities of the corresponding enzymes are reduced in the presence of a preferred carbon source”. Though the precise physiological role of CCR in microorganisms remains unclear, it has been thought to provide significant advantages to microorganisms during growth under changing environmental conditions, and also has a role in increasing virulence of microorganisms (Görke and Stülke, 2008a). A study by Beg et al. (2007) traced the origin of sequential utilization of sugars in *E. coli* to the physical limitations of the cell’s volume, which cannot accommodate all the enzymes required to utilize different sugars simultaneously. Thus, CCR should be viewed as a mechanism that enables most efficient carbon source utilization, and not as a cause of sequential sugar utilization. At molecular level, CCR in *E. coli* is achieved through two distinct mechanisms: a global regulatory mechanism involving cAMP-CRP complex, and multiple operon specific mechanisms collectively known as inducer exclusion (Fig. 2.1). The glucose uptake system in *E. coli*, known as phosphoenolpyruvate-carbohydrate phosphotransferase system (PTS), has a major role to play in both global regulatory mechanism and inducer exclusion mechanism of CCR.
Glucose transport

PTS is an active glucose transport system unique to bacterial species (Görke and Stülke, 2008a). It consists of multiple proteins which participate in a phosphorelay system, eventually leading to phosphorylation and simultaneous transport of glucose across the cytoplasmic membrane. The *E. coli* PTS consists of three distinct, soluble proteins: enzyme I (EI, encoded by *ptsI*), histidine protein (HPr, encoded by *ptsH*), and enzyme II (EIIAGlc, encoded by *crr*). Glucose transport in *E. coli* is tightly coupled to its metabolism, and depends on the intracellular ratio of pyruvate to phosphoenolpyruvate (PEP) (Fig. 2.1). At high extracellular glucose concentrations, a large glycolytic flux results in high intracellular PEP levels. Under these conditions, the phosphorelay is initiated with autophosphorylation of EI which accepts a phosphate group from PEP. The phosphate group is then transferred to HPr, which donates the phosphate group to EIIAGlc. EIIAGlc transfers the phosphate group to the cytoplasmic EIIB domain of the glucose transporter EIIBC<sub>Glc</sub> (encoded by *ptsG*). The phosphate group is finally transferred to glucose while it is transported across the membrane by the transmembrane EIIC domain of EIIBC<sub>Glc</sub>. A schematic of the PTS system in *E. coli* is shown in Figure 2.1.

In addition to glucose transport, PTS also regulates sugar uptake in *E. coli*. Among the proteins involved in PTS, EIIAGlc plays a pivotal role in CCR by participating in both global regulatory mechanism and inducer exclusion mechanism (explained further). In presence of high extracellular concentrations of glucose, when *E. coli* shows a rapid utilization of the sugar and a high flux through glycolysis, EIIAGlc remains predominantly non-phosphorylated by continuously donating the phosphate group to glucose. In its non-phosphorylated state, EIIAGlc binds and inactivates the metabolic enzymes and transporters of secondary carbon sources (Huffer et al., 2012; Görke and Stülke, 2008a) (Fig. 2.1). This mechanism of inhibition of secondary sugar uptake is known as inducer exclusion. In the absence of extracellular glucose, EIIAGlc is present in phosphorylated state due to lower flux of phosphorylation and transport of glucose. In its phosphorylated state, EIIAGlc binds and activates the membrane bound enzyme adenylate cyclase (AC), which catalyzes the conversion of AMP to cyclic AMP (cAMP). At high concentrations, cAMP binds to CRP (cAMP receptor protein), forming cAMP-CRP complex. The cAMP-CRP complex is responsible for activation of transcription of the genes responsible
for transport and catabolism of secondary carbon sources such as lactose, xylose, and glycerol. The cAMP-CRP-based regulation of genes responsible for secondary sugar utilization is known as global regulatory mechanism.
\textit{E. coli} is also capable of transporting glucose using non-PTS transporters for glucose, however, these modes of glucose transport are used only under drastic genetic perturbations such as deletion of PTS. PTS\(^{-}\) strains of \textit{E. coli} are known to recruit the galactose transporters GalP (galactose permease) and MglBAC (galactose ABC transporter) to transport glucose across the cell membrane (Flores et al., 2005).

**Xylose and arabinose transport**

Xylose transport in \textit{E. coli} is primarily achieved through an ATP-dependent high-affinity ABC (ATP Binding Cassette superfamily) transporter (encoded by \textit{xylFGH}), and to a lower extent by an alternative low affinity xylose symporter (encoded by \textit{xylE}) (Desai and Rao, 2010). The ABC transporter of xylose consists of the ATP-binding protein XylG, the periplasmic substrate-binding protein XylF, and the membrane component of the ABC transport system XylH. In presence of xylose, the xylose transcription factor XylR, binds to the promoter regions of the operons \textit{xylAB} and \textit{xylFGHR}, thereby activating their transcription. The presence of high concentrations of cAMP-CRP complex activates XylR (Fig. 2.2a). In the absence of glucose, the intracellular concentration of cAMP-CRP is high, which induces xylose utilization and metabolism through expression of \textit{xylAB} and \textit{xylFGHR}. The genes \textit{xylA} and \textit{xylB} encode for xylose isomerase and xylulokinase, respectively, which convert xylose into xylulose-5-phosphate (X5P).

In addition to cAMP-CRP mediated global regulatory mechanism, xylose utilization is also regulated by the transcription factor AraC involved in arabinose uptake and metabolism (Desai and Rao, 2010). Between xylose and arabinose, \textit{E. coli} preferably consumes arabinose (Desai and Rao, 2010; Hernández-Montalvo et al., 2001; Kang et al., 1998). The molecular mechanism of arabinose repression of xylose occurs at transcriptional level by binding of AraC transcriptional factor to the promoter regions of \textit{xyl} system. AraC binds to the promoter region \(P_{xylA}\) thereby avoiding the binding of XylR to the promoter region for activation of xylose metabolism (Fig. 2.2b). Regulation of arabinose operon in \textit{E. coli} has been subject of extensive research as a model system for positive gene regulation, and has led to the discovery of the phenomenon known as DNA looping (simultaneous binding of a protein or protein complex to two different sites of DNA causing the intervening DNA to form a loop), which is widely
Figure 2.2: Schematic representation of activation on xyl system in presence of different sugars. a) In presence of xylose and absence of glucose, xylAB is activated by binding of XylR-xylose complex, RNA polymerase (RNAP), and CRP-cAMP complex, b) in presence of arabinose, the AraC-arabinose complex binds to the promoter P_{xylA} thereby inhibiting the binding of XylR-xylose complex and repressing the expression of the operon xylAB, c) in presence of glucose, the intracellular CRP-cAMP levels are low and XylR-xylose complex is absent causing no expression of xylAB operon.

used by organisms for gene regulation (Schleif, 2010). Arabinose uptake in E. coli is brought about by an ABC transporter encoded by araFGH and a proton symporter encoded by araE. Arabinose is metabolized by the products of the operon araBAD which convert arabinose into X5P. X5P enters metabolism through the pentose phosphate pathway. In presence of arabinose, the protein AraC activates the transcription of the ara system and enables arabinose uptake
and metabolism. The protein also regulates expression of its own synthesis, and as discussed above, represses the expression of genes from \textit{xyl} system.

Understanding the utilization mechanisms of glucose, xylose, and arabinose in \textit{E. coli} is central to the utilization of LCB as the second generation feedstock. The strains of \textit{E. coli} devoid of CCR are especially important in the biofuels and renewable chemicals industry. Multiple studies have reported construction of strains that can co-utilize the major sugars of lignocellulosic hydrolysate.

\subsection*{2.1.2 Sugar transport in \textit{Saccharomyces cerevisiae}}

Glucose uptake system of eukaryotic organisms is distinct from bacterial PTS-based active transport. As the work reported in this thesis is focused on bacterial system, the eukaryotic system (\textit{S. cerevisiae}) is only briefly described. Glucose transport in \textit{S. cerevisiae} is brought about by facilitated diffusion by a family of 17 sugar transporters. The sugar transporters in \textit{S. cerevisiae} are designated as HXT1-17 (Hexose Transporter). The galactose transporter GAL2 is also known to transport glucose in \textit{S. cerevisiae}, but is activated only in presence of galactose. The kinetic properties of HXTs vary considerably allowing \textit{S. cerevisiae} to uptake glucose in a large concentration range. Based on their kinetic properties, the HXTs are divided into three different groups: 1) low affinity transporters ($K_M = 50 - 100$ mM), 2) moderately low affinity transporters ($K_M = 10$ mM), and 3) high affinity transporters ($K_M = 1 - 2$ mM). A summary of the hexose transporters of \textit{S. cerevisiae} is provided in Table 2.1.

Various regulatory mechanisms control the expression of appropriate \textit{HXT} genes based upon the extracellular glucose concentrations. For example, in presence of low concentration of glucose, high affinity transporters (such as HXT6 and HXT7) are expressed, whereas in presence of high concentration of glucose, low affinity transporters (such as HXT1) are expressed. For sensing the extracellular glucose concentration, \textit{S. cerevisiae} uses sensor proteins, SNF3 and RGT2. SNF3 and RGT2 sense glucose and mediate subsequent signaling at low and high glucose concentrations, respectively.

\textit{S. cerevisiae} does not have a dedicated xylose transporter and transports xylose non-specifically using glucose transporters. The transporters HXT1, HXT4, HXT5, HXT7, and GAL2 are known to play important role in xylose transport (Sedlak and Ho, 2004; Saloheimo
### Table 2.1: Hexose transporters in *S. cerevisiae*

<table>
<thead>
<tr>
<th>Transporter</th>
<th>Property</th>
<th>Glucose $K_M^a$ (mM)</th>
<th>Glucose $V_{max}^b$ (g/gDW*h)</th>
<th>Xylose $K_M^c$ (mM)</th>
<th>Xylose $V_{max}^b$ (g/gDW*h)</th>
<th>Role in xylose transport $^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HXT1</td>
<td>low affinity</td>
<td>100</td>
<td>7.45</td>
<td>880</td>
<td>2.34</td>
<td>none</td>
</tr>
<tr>
<td>HXT2</td>
<td>intermediate affinity</td>
<td>1.5</td>
<td>1.90</td>
<td>260</td>
<td>1.8</td>
<td>unknown</td>
</tr>
<tr>
<td>HXT3</td>
<td>low affinity</td>
<td>60</td>
<td>2.70</td>
<td>–</td>
<td>–</td>
<td>unknown</td>
</tr>
<tr>
<td>HXT4</td>
<td>intermediate affinity</td>
<td>9</td>
<td>1.08</td>
<td>170</td>
<td>0.63</td>
<td>supports</td>
</tr>
<tr>
<td>HXT5</td>
<td>intermediate affinity</td>
<td>$10^e$</td>
<td>$7.6^e$</td>
<td>–</td>
<td>–</td>
<td>supports</td>
</tr>
<tr>
<td>HXT6</td>
<td>high affinity</td>
<td>1.5</td>
<td>1.31</td>
<td>–</td>
<td>–</td>
<td>unknown</td>
</tr>
<tr>
<td>HXT7</td>
<td>high affinity</td>
<td>1.5</td>
<td>1.31</td>
<td>130</td>
<td>0.36</td>
<td>supports</td>
</tr>
<tr>
<td>HXT8-11</td>
<td>uncharacterized $^f$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>none</td>
</tr>
<tr>
<td>HXT12</td>
<td>likely a pseudogene $^f$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>unknown</td>
</tr>
<tr>
<td>HXT17</td>
<td>high affinity $^g$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>unknown</td>
</tr>
<tr>
<td>GAL2</td>
<td>high affinity</td>
<td>1.2$^d$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>supports</td>
</tr>
</tbody>
</table>

$^a$ (Boles and Hollenberg, 1997), $^b$ (Bertilsson et al., 2008), $^c$ (Saloheimo et al., 2007), $^d$ (Hamacher et al., 2002), $^e$ (Buziol et al., 2008), $^f$ (Wieczorke et al., 1999), $^g$ (Klockow et al., 2008)

et al., 2007). Additionally, xylose transport in *S. cerevisiae* has been achieved using expression of heterologous xylose transporters (Kotter and Ciriacy, 1993; Kuyper et al., 2003). In presence of glucose, *S. cerevisiae* shows a strong repression of xylose uptake.

Two major regulatory mechanisms mediate glucose repression in *S. cerevisiae*: 1) main glucose repression pathway and 2) glucose induction. The main glucose repression pathway has a dominant role and involves activity of proteins MIG1, SSN6, and TUP1 (Rolland et al., 2002). MIG1 is a DNA binding protein and causes repression of multiple genes by directly binding to the corresponding DNA regions. In presence of glucose, MIG1 is present inside the nucleus and causes repression of genes when bound to SSN6 and TUP1. In absence of glucose, MIG1 is exported out of the nucleus. In absence of MIG1 repression, the genes responsible for uptake and metabolism of secondary sugars are expressed. The detailed mechanisms of glucose
repression in *S. cerevisiae* have been reviewed elsewhere (Rolland et al., 2002; Santangelo, 2006; Kim et al., 2013).

### 2.1.3 Sugar transport and regulation in other microorganisms

Presence of CCR results into similar outcomes in all microorganisms: the genes required for utilization and metabolism of non-preferred substrates are repressed in presence of a preferred substrate such as glucose. The molecular mechanisms responsible for CCR, however, may be entirely different. Whereas most bacterial CCR mechanisms commonly involve the PTS system, eukaryotic CCR mechanisms involve various glucose-dependent signaling pathways. Considerable variation exists even among the PTS-dependent bacterial systems. For example, CCR in *E. coli* is achieved by the prevention of transcriptional activation of secondary substrate genes by inhibition of the inducer uptake; however, CCR in *B. subtilis* is achieved by negative regulation using a repressor protein binding to the promoter regions of secondary substrate genes (Görke and Stülke, 2008a). The key players involved in CCR in *E. coli* are the PTS protein EIIAGlc, CRP, cAMP, and the glycolytic intermediates pyruvate and PEP. Whereas, the main players involved in CCR in *B. subtilis* are the transcription factor CcpA (catabolite control protein A), the PTS protein HPr, HPrK (HPr kinase/phosphorylase), and the glycolytic intermediate fructose-1,6-bisphosphate (FBP) (Titgemeyer and Hillen, 2002; Warner and Lolkema, 2003). Unlike *E. coli*, HPr phosphorylation plays an important role in CCR mechanism of *B. subtilis*. Under high extracellular glucose concentration, intracellular FBP concentration is high causing HPrK to phosphorylate HPr. Phosphorylated HPr binds to CcpA, and the HPr-CcpA complex thus formed binds to the *cre* (catabolite responsive elements) sites, most of which are located in transcription initiation regions of secondary substrate genes. In the absence of extracellular glucose, high intracellular inorganic phosphate concentration activates phosphorylase activity of HPrK and dephosphorylates HPr, releasing the repression of secondary substrate genes. Similar to *B. subtilis*, many other Gram-positive organisms use HPr, HPrK, and CcpA dependent CCR mechanism (Warner and Lolkema, 2003).

Many other variations are known to exist in the bacterial CCR mechanisms. In *Clostridium cellulolyticum*, expression of cellulosome components is repressed in presence of glucose; however, the proteins involved in CCR are not similar to PTS proteins. In *C. cellulolyticum*,
CCR is mediated by CcpA and a protein similar to Crh protein of *B. subtilis* (Abdou et al., 2008). In some organisms such as *Streptomyces coelicolor* CCR is completely independent of PTS and involves glucose kinase as the key regulatory protein (Angell et al., 1994). Similar glucose kinase dependent CCR has also been observed in *Staphylococcus xylosus* and *Bacillus megaterium* (Brückner and Titgemeyer, 2002).

CCR in *Corynebacterium glutamicum* has been extensively studied because of the organism’s industrial relevance as an amino acid producer. It has a PTS system that involves EI, HPr, and four EII permeases that are specific for glucose, fructose, sucrose, and an unknown substrate (Ikeda, 2012). *C. glutamicum* does not show a preference for glucose and co-utilizes multiple carbon sources such as sucrose, fructose, xylose, ribose, pyruvate, and lactate simultaneously with glucose. However, *C. glutamicum* shows diauxic growth on a mixture of glucose and glutamate or a mixture of glucose and acetate. CCR in *C. glutamicum* involves the repressor protein RamB, which is activated in presence of glucose by an unknown mechanism, and binds to the promoter regions of the secondary substrate genes (Gerstmeir et al., 2004).

Using a distinctive molecular mechanism, *Pseudomonas putida* shows CCR by controlling the translation of operon specific regulators (Moreno et al., 2007). The translational regulators such as AlkS are responsible for activating the expression of secondary substrate operons in presence of appropriate inducers; however, under CCR conditions, the RNA-binding protein, Crc, binds to the mRNAs of the regulators, thereby repressing the activation of the secondary substrate genes in presence of glucose. This mechanism is different from the previously discussed mechanisms of CCR involving DNA-binding proteins.

CCR is a ubiquitous and one of the most important regulatory phenomena that gives microorganisms competitive advantage under varying environmental conditions. Some pathogenic bacteria such as *Chlamydia trachomatis* and *Mycoplasma pneumoniae* are known to entirely lack CCR due to their adaptation to nutrient rich growth conditions (Nicholson et al., 2004; Halbedel et al., 2007). Additionally, some bacteria such as *Pseudomonas aeruginosa* consume glucose as a secondary substrate and show reverse CCR (Collier et al., 1996). Though there are some common regulatory motifs of CCR present in closely related groups of microorganisms, it appears that each group has evolved its own specific nuances of the mechanism. Thus, although the basic principles of CCR are well understood in some model microorganisms, the detailed
molecular mechanisms of CCR in all classes of microorganisms and the common evolutionary themes, if any, remain to be well understood.

2.1.4 Glucose-xylose co-utilization in *E. coli*

The ability to utilize different sugars as carbon source is a very useful trait in industrial microorganisms. *E. coli* is a superior host in this respect. However, sequential utilization of sugars is non-ideal for industrial fermentations for the reasons mentioned in the previous section. Mutants devoid of CCR can address the problems associated with sequential sugar utilization. Co-utilization of sugars can potentially increase the productivity of fermentation processes (Nichols et al., 2001) and simplify the processes in large-scale operations (Kim et al., 2010). Several engineering strategies have been previously implemented to obtain *E. coli* strains that can simultaneously metabolize glucose and xylose. Most of these strategies involved deletion, overexpression, or alteration of regulatory genes involved in CCR.

PTS plays a major role in CCR in *E. coli*. Earliest strategies to abolish CCR involved deletion of the genes involved in PTS, such as main glucose transporter EHBC$_{Glc}$ (encoded by *ptsG*) (Nichols et al., 2001). The mutant *E. coli* Δ*ptsG* was used to convert glucose and xylose into ethanol, and could utilize the two sugars simultaneously, but could not consume all the xylose in the medium (Nichols et al., 2001). Another *E. coli* mutant lacking PTS system, obtained by deletion of *ptsHIcrr* operon, showed glucose-xylose co-utilization under anaerobic conditions (Balderas-Hernández et al., 2011). Though such mutants constructed by deleting PTS co-utilize glucose and xylose, they show considerably reduced glucose uptake rates.

Other strategies for forcing co-utilization of glucose and xylose in *E. coli* involve altering the global regulatory mechanisms of CCR (involving cAMP-CRP based activation of catabolic genes). Allosteric alterations of the global regulatory protein CRP via point mutation(s) (designated as crp$^*$ or crp$^+$) can confer relaxed CCR phenotype. These mutations make CRP bind to the promoter regions of catabolic genes independent of cAMP levels (Khankal et al., 2009). A crp$^*$ mutant with an amino acid substitution from Gly$^{122}$ to Ser$^{122}$ was shown to co-utilize glucose and xylose (Yao et al., 2011).

More recent strategies to obtain co-utilization phenotype in *E. coli* have involved manipulations to the genes not conventionally associated with CCR. For example, enhanced co-utilization
of glucose, xylose, arabinose, mannose, and galactose was found on deletion of methylglyoxal synthase gene \textit{mgsA} (Yomano et al., 2009). Extensive metabolic manipulations to \textit{E. coli} are also known to cause elimination of CCR leading to co-utilization of glucose and xylose. For example, an \textit{E. coli} mutant constructed with eight gene deletions for production of ethanol did not regulate its sugar uptake and co-utilized glucose and xylose (Trinh et al., 2008). The co-utilization of sugars in this mutant was most likely due to a generalized response to severe metabolic manipulation.

In a completely novel approach to ensure co-utilization of glucose and xylose, Eiteman et al. (2008) used co-fermentation strategy with a consortium of \textit{E. coli} mutants. One of the mutants in the consortium was incapable of consuming glucose (ZSC113) and the other was incapable of consuming xylose (ALS1008). This method circumvented the complications arising due to changing ratios of glucose and xylose in the feed, as the utilization of sugars was independent to each other. Similar consortia were further used for production of lactate (Eiteman et al., 2008), and co-utilization of glucose, xylose, arabinose, and acetate (Xia et al., 2012).

Vinuselvi and Lee (2012) engineered \textit{E. coli} for co-utilization of cellobiose and xylose by activating the cryptic operons \textit{chb} and \textit{asc}. Cellobiose is a dimer of glucose and was hydrolyzed to glucose only inside the cell. As the PTS system was not used for transporting glucose into the cell, xylose uptake was not repressed and the mutant consumed cellobiose and xylose simultaneously.

A list of \textit{E. coli} mutants engineered for sugar co-utilization are listed in Table 2.2.
Table 2.2: Glucose-xylose co-utilizing *E. coli* mutants.

<table>
<thead>
<tr>
<th>Strain name</th>
<th>Genotype</th>
<th>Medium</th>
<th>Sugars g/L</th>
<th>Time h</th>
<th>Conditions</th>
<th>Co-utilization characteristics</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> K12</td>
<td><em>pts</em>G21, Δ<em>frd</em>ABCD, pLOI297</td>
<td>LB</td>
<td>G = 37, X = 36, A = 12</td>
<td>70</td>
<td>An</td>
<td>co-utilization of all G, X, and A; no residual sugars</td>
<td>(Nichols et al., 2001)</td>
</tr>
<tr>
<td><em>E. coli</em> PB103</td>
<td>Δ(<em>ptsH ptsI err</em>):kanR</td>
<td>M9</td>
<td>G = 1, X = 1</td>
<td>11</td>
<td>Ar</td>
<td>partial co-utilization of G and X; no residual sugars</td>
<td>(Hernández-Montalvo et al., 2001)</td>
</tr>
<tr>
<td><em>E. coli</em> K12</td>
<td><em>pts</em>G21, Δ<em>frd</em>ABCD, pUCLDH1</td>
<td>LB</td>
<td>G = 50, X = 50</td>
<td>150</td>
<td>An</td>
<td>co-utilization of G and X; incomplete consumption of X</td>
<td>(Dien et al., 2002)</td>
</tr>
<tr>
<td><em>E. coli</em> ATCC23724</td>
<td>mutations in <em>pts</em>G, <em>p</em>RB, and <em>ldh</em>A</td>
<td>MM</td>
<td>G = 50, X = 50</td>
<td>30</td>
<td>Ar+An</td>
<td>partial co-utilization of G and X; incomplete sugar consumption</td>
<td>(Andersson et al., 2007)</td>
</tr>
<tr>
<td><em>E. coli</em> MG1655</td>
<td>Δ<em>zuf</em> Δ<em>ndh</em> Δ<em>ldh</em>A Δ<em>sfc</em>:A Δ<em>mace</em>:B Δ<em>frd</em>:B Δ<em>poz</em>:B Δ<em>pta</em>:kanR</td>
<td>LB</td>
<td>G = 40, X = 40</td>
<td>48</td>
<td>An</td>
<td>co-utilization of G and X; no residual sugars</td>
<td>(Trinh et al., 2008)</td>
</tr>
<tr>
<td><em>E. coli</em> MG1655</td>
<td>Δ<em>zyl</em>A</td>
<td>MM</td>
<td>G = 30, X = 20</td>
<td>16.5</td>
<td>Ar+An</td>
<td>different inoculation times of two cultures were required to ensure consumption of sugar in same time.</td>
<td>(Eiteman et al., 2009)</td>
</tr>
<tr>
<td><em>E. coli</em> MG1655</td>
<td>Δ<em>pts</em>G Δ<em>glk</em> Δ<em>man</em>:Z</td>
<td>MM</td>
<td>G = 50, X = 50</td>
<td>30</td>
<td>Ar</td>
<td>partial co-utilization of G and X; no residual sugars</td>
<td>(Balderas-Hernández et al., 2011)</td>
</tr>
<tr>
<td><em>E. coli</em> MG1655</td>
<td>Δ<em>pts</em>G, pLOI1594</td>
<td>MM</td>
<td>G = 5, X = 5</td>
<td>25</td>
<td>An</td>
<td>partial co-utilization of G and X; no residual sugars</td>
<td>(Yao et al., 2011)</td>
</tr>
<tr>
<td><em>E. coli</em> W3110</td>
<td>mutated <em>crp</em> Gly^{122} to Ser^{122}</td>
<td>M9</td>
<td>G = 6, X = 5</td>
<td>14</td>
<td>Ar</td>
<td>co-utilization of G and X; residual X</td>
<td>(Yao et al., 2011)</td>
</tr>
<tr>
<td><em>E. coli</em> BW25113</td>
<td><em>chb</em>, asc activation</td>
<td>MM</td>
<td>X = 2, C = 6</td>
<td>35</td>
<td>Ar</td>
<td>co-utilization of X and C; no residual sugars</td>
<td>(Vinuselvi and Lee, 2012)</td>
</tr>
<tr>
<td><em>E. coli</em> BL21</td>
<td>Δ<em>pts</em>G Δ<em>poz</em>:B Δ<em>ldh</em>:A Δ<em>pta</em>:Δ<em>frd</em>:kanR</td>
<td>MM</td>
<td>G = 30, X = 30</td>
<td>25</td>
<td>An</td>
<td>co-utilization of G and X; no residual sugars</td>
<td>(Chiang et al., 2013)</td>
</tr>
</tbody>
</table>

Abbreviations used: G - glucose, X - xylose, A - arabinose, C - cellobiose, LB - Luria-Bertani broth, MM - minimal medium, Ar - aerobic, An - anaerobic
The multitude of studies on glucose-xylose co-utilizing *E. coli* underscores the industrial importance of such strains. As the studies report diverse media and environmental conditions, comparison of the performance of these mutants is difficult. Different strategies used to obtain the co-utilization strains, and their characterization under different media and environmental conditions suggest that no single strategy is superior and universally applicable to different fermentation processes.

Similar to *E. coli*, metabolic engineering strategies involving deletion of regulatory genes have been attempted to enable efficient utilization of xylose in *S. cerevisiae*. These metabolic engineering strategies have met with limited success in *S. cerevisiae*. For example, deletion of the gene *MIG1* did not force co-utilization of glucose and xylose (Roca et al., 2004). However, using alternative approaches such as adaptive evolution and evolutionary engineering, strains with significantly improved xylose utilization characteristics have been developed (Wisselink et al., 2009; Bera et al., 2011). These strains do not strictly co-utilize glucose and xylose, regardless.

### 2.1.5 Opportunities for advancement

Previously described glucose-xylose co-utilizing *E. coli* mutants were constructed by altering the regulatory genes involved in CCR. Though some of these mutants were able to co-utilize glucose and xylose, engineering of regulatory pathways is not the best approach to eliminate sequential sugar utilization. Several drawbacks are associated with engineering of regulatory network for forcing co-utilization of sugars.

1. Engineering of regulatory network can be carried out only for highly characterized organisms. Many industrial applications use non-conventional organisms. Such organisms may not be extensively characterized for their regulatory networks. Identifying engineering targets for such organisms may not be possible. In comparison, metabolic networks are conserved and relatively less complex. Metabolic networks are well characterized in many organisms, and can be easily modeled. Engineering approaches using metabolic networks are hence more convenient than the regulatory networks.

2. Engineering the regulatory network may sometimes not be successful. Engineering of reg-
ulatory networks has proven unsuccessful even for highly characterized microorganisms. For example, *S. cerevisiae* is a model eukaryotic organism and has been extensively characterized. However, deletion of the regulatory gene MIG1 (the primary protein involved in main glucose repression pathway of *S. cerevisiae*) was unsuccessful to force co-utilization of glucose and xylose (Roca et al., 2004).

3. Engineering of regulatory network may only temporarily eliminate CCR, and the mutant may revert back to the original sequential sugar utilization phenotype by using an alternative regulatory mechanism. Regulation of sugar utilization generally involves multiple regulatory mechanisms. For example, *E. coli* CCR involves global regulatory mechanism as well as inducer exclusion mechanism. Inactivation of one of the mechanisms may not be sufficient to completely eliminate CCR, and the organism may evolve to use the alternative mechanism to control sugar utilization.

4. Finally, engineering of regulatory network may cause undesired physiological changes. As some of the genes involved in CCR may express global regulatory proteins (example CRP in *E. coli*), elimination or alteration of these genes may have unnecessary global effects.

Given the drawbacks in engineering the regulatory networks, there is a need for a completely novel approach to address the problem of sequential sugar utilization in industrial microorganisms. The solution methodology should ideally be independent of the regulatory network to enable its application to non-conventional organisms, and it should be stringent to force the organism to permanently co-utilize sugars, thereby eliminating the chances of reversion. The solution conceived in this thesis achieves these criteria. The algorithm SimUp, which identifies the reaction deletion strategies, is based only on metabolic network, and requires no knowledge of regulatory network. Additionally, the strategies identified by SimUp inherently eliminate the possibility of reversion to the original phenotype by coupling sugar co-utilization phenotype to growth. When sugar co-utilization is coupled to growth, the organism cannot grow without consuming the sugars simultaneously, and thus, can never revert back to utilizing the sugars sequentially.
2.2 Mathematical modeling in metabolic engineering

Metabolic engineering concerns with altering the genetic network of microorganisms in order to achieve specific metabolic objectives. Metabolic engineering is based on systems-level understanding of the cellular metabolism. This approach is distinct from conventional genetic engineering which targets isolated pathways or sub-networks. Due to the enormity and complexity of the metabolic networks, metabolic engineering relies on computational methods for analysis and prediction. Copeland et al. (2012) list seven different computational areas that are commonly used in studies involving metabolic engineering. Among these seven areas, metabolic network reconstruction and metabolic flux analysis using constraint-based modeling (CBM) are the most extensively used techniques for industrial strain design.

CBM is a mathematical framework that enables modeling of microbial physiology using the reconstructed metabolic networks (metabolic models). Over a hundred different CBM methods have been developed (Lewis et al., 2012). These methods use metabolic models to predict microorganisms’ phenotypes (growth rates and fluxes through metabolic reactions) under different environmental and genetic perturbations (Edwards et al., 2001; Ibarra et al., 2002; Famili et al., 2003). The predictive capability of CBM has been extended for engineering of industrial microorganisms (Alper et al., 2005; Kennedy et al., 2009; Chemler et al., 2010), simulation of interactions in microbial communities (Stolyar et al., 2007; Taffs et al., 2009), and finding novel drug targets (Kim et al., 2011a; Shen et al., 2010). This section discusses the fundamentals of CBM, examples of a few popular CBM methods, and development and applications of CBM methods for strain design.

2.2.1 Constraint-based modeling basics

CBM techniques are based on a few fundamental concepts. Metabolic network reconstruction is the essential first step. Metabolic network reconstruction entails mathematical representation of the constituents of a metabolic network (metabolites) and the interactions within (reactions connecting the metabolites). The most popular approach to represent a metabolic network is a stoichiometric matrix, denoted by $S$. The $S$ matrix has the dimensions $M \times N$, where $M$ is the number of metabolites and $N$ is the number of reactions. Using the annotated
genome of a microorganism, its entire metabolic network can be assembled and represented as a stoichiometric matrix. The process of model reconstruction from the genome sequence data has been automated to a large extent (Devoid et al., 2013). Using frameworks such as RAST (Aziz et al., 2008) and Model SEED (Devoid et al., 2013), the genome sequence of a microorganism can be first annotated and converted into a genome-scale metabolic model with minimal manual effort. The reconstructed networks can then be subjected to various analyses by the use of CBM techniques (Lewis et al., 2012).

The earliest CBM methods were developed to predict the reaction rates (metabolic fluxes) in microorganisms under steady-state growth conditions. Equation (2.1) represents the rate of change in concentration of an intracellular metabolite,

\[
\frac{dc_A(t)}{dt} = \sum_{i \in \text{prod}} a_i v(t)_i - \sum_{j \in \text{cons}} b_j v(t)_j,
\]

where, \(c_A\) is the intracellular metabolite concentration (mM), \(a_i\) are the stoichiometric coefficients associated with the metabolite in the production reactions, \(b_j\) are the stoichiometric coefficients associated with the metabolite in the consumption reactions, and \(v_i\) and \(v_j\) represent the rates (mmol/gDW/h) of the production reactions and the consumption reactions, respectively. Equation (2.1) can be extended over the entire metabolic network as shown in equation (2.2):

\[
\frac{dc(t)}{dt} = Sv(t),
\]

where, \(S_{M \times N}\) is the stoichiometric matrix of a metabolic network with \(M\) metabolites and \(N\) reactions, \(v \in \mathbb{R}^N\) is a vector of all reaction fluxes, and \(c \in \mathbb{R}^M\) is a vector of all intracellular metabolite concentrations. Under steady-state growth conditions, when there is no change in the intracellular metabolite concentration, equation (2.2) is reduced to,

\[
Sv = 0.
\]

A solution to the equation (2.3) represents the steady-state distribution of metabolic fluxes (flux distribution vector). The system of equations represented by equation (2.3) is typically an under-determined system (\(M < N\)) and has multiple solutions. Among the possible solutions,
physiologically meaningful flux distributions can be identified using the most fundamental CBM technique - **Flux Balance Analysis (FBA)** (Varma and Palsson, 1994; Orth et al., 2011). FBA is formulated as a linear optimization problem. It identifies meaningful flux distributions by optimizing physiologically relevant objective functions such as maximization of growth rate or ATP production (Schuetz et al., 2007). The basic premise of FBA is that an organism distributes its fluxes through metabolic reactions to achieve specific physiological objectives (for example, maximization of growth rate). Thus, a flux distribution identified by FBA using an appropriate objective should be very close to the actual flux distribution of the organism.

A typical formulation for FBA is shown in equation (2.4):

\[
\max_v \quad f^T v \\
\text{s.t.} \quad Sv = 0 \\
\quad v^L \leq v \leq v^U,
\]  

(2.4)

where, \( S \) is the stoichiometric matrix, \( v \) is the flux variable, \( v^L \) and \( v^U \) are the lower and upper bounds of the reactions fluxes \( v \), and \( f \in \mathbb{R}^N \) is a vector of objective coefficients. For the growth-maximization objective, all the elements of the vector \( f \) are 0 except for the index corresponding to the biomass formation reaction, which has the value 1. FBA has been used to calculate maximum growth rate of organisms under various environmental and genetic perturbations (Edwards et al., 2001; Ibarra et al., 2002; Famili et al., 2003). Different environmental conditions in FBA are simulated by altering the bounds \( (v^L \text{ and } v^U) \) of the appropriate reactions. For example, aerobic and anaerobic growths can be simulated using FBA by changing the upper and lower bounds of the flux variable corresponding to the oxygen uptake reaction. If the index \( j \) corresponds to oxygen uptake reaction, then for the anaerobic case, \( v^L_j = v^U_j = 0 \), and for aerobic case, \( v^L_j = -20 \text{ mmol/gDW/h} \) and \( v^U_j = 0 \). By convention, the uptake reactions have negative fluxes \( (v_j < 0) \). Similarly, genetic perturbations such as gene deletions can be simulated by setting the values of both upper and lower bounds for the reaction(s) corresponding to the gene to 0. The flux value through that particular reaction is thus set to 0. FBA was shown to quantitatively predict growth rates of *E. coli* and *S. cerevisiae* mutants, especially after adaptive evolution (Förster et al., 2003; Fong and Palsson, 2004).
2.2.2 Alternatives to FBA

FBA predicts flux distribution with the implicit assumption that organisms naturally maximize their growth rates. However, this assumption may not be universally valid. For the organisms with genetic perturbations, the physiological objective may not be growth maximization. Alternative objectives such as minimization of metabolic adjustment (Segrè et al., 2002) and minimization of regulatory changes (Shlomi et al., 2005) have been suggested for mutants with metabolic gene deletions. CBM methods based on such alternate physiological objectives have been shown to predict phenotypes of gene deletion mutants with better accuracy.

MOMA (Minimization Of Metabolic Adjustment), developed by Segrè et al. (2002), uses an alternative objective function to identify flux distribution of gene deletion mutants. The objective function is based on the premise that the mutants tend to maintain their flux distribution close to the wild-type flux distribution immediately after gene deletion (sub-optimal growth phase). Mathematically, the objective function of MOMA is formulated as minimization of the Euclidean norm between the flux distributions of the mutants and the wild-type. MOMA has been shown to predict the sub-optimal growth rates of gene knockout mutants with increased accuracy than FBA (Segrè et al., 2002). Formulation of MOMA is very similar to FBA and is shown in equation (2.5) below:

$$\min_v (v - w)^T (v - w)$$

s.t. $Sv = 0$

$$v^L \leq v \leq v^U,$$

where, the vector $v \in \mathbb{R}^N$ is the mutant flux distribution and the vector $w \in \mathbb{R}^N$ is the wild type flux distribution.

The objective function of MOMA tends to prohibit large flux changes after gene knockouts while allowing small changes in multiple fluxes. However, it has been shown that under certain conditions mutants tend to reroute their fluxes by making few large changes instead of multiple small changes to the pathways. Regulation of On-Off Minimization (ROOM), developed by Shlomi et al. (2005) takes this fact into account. The objective function of ROOM minimizes
the number of reactions requiring flux changes (Hamming distance\textsuperscript{1} from the reference flux vector) instead of the flux values. ROOM was shown to quantitatively better predict the flux distributions in single gene knockout mutants of \textit{E. coli} (Shlomi et al., 2005).

Both ROOM and MOMA require a reference wild-type flux distribution. However, the flux distribution of the wild-type calculated by FBA may not be unique owing to the existence of multiple optima (Mahadevan and Schilling, 2003). The latest CBM developed by Kim and Reed (2012), RELATCH (for RELATive CHange), uses C\textsuperscript{13} metabolic flux analysis data, and gene expression data to identify the reference flux distribution and gene expression of the wild-type. With the correct reference flux distribution, the perturbed metabolic flux distribution of the mutants is more accurately predicted by RELATCH compared to both MOMA and ROOM (Kim and Reed, 2012).

Using two additional parameters, RELATCH formulation also enables simulation of both sub-optimal growth phase (phase immediately after gene deletion) and growth after adaptive evolution. The sub-optimal growth phase of mutants is characterized by activation of latent reactions, which are not active under the wild-type growth. However, the mutants tend not to activate multiple latent pathways and keep the flux distribution close to the wild-type. In addition, the fluxes through the already active reactions do not change significantly during the sub-optimal growth phase. After adaptive evolution, however, the mutants reach (close to) the maximum possible growth rate, mainly by improving the capacity of the already active reactions. These scenarios are simulated by RELATCH using two parameters, \(\alpha\) and \(\gamma\). Whereas \(\alpha\) penalizes the activation of inactive pathways in the mutants, \(\gamma\) limits the contribution of the already active enzymes. By varying the values of \(\alpha\) and \(\gamma\) the effects of genetic perturbations can be simulated under either unevolved (high \(\alpha\) and low \(\gamma\)) or evolved (low \(\alpha\) and high \(\gamma\)) conditions. RELATCH provides the best quantitative accuracy with flux distribution prediction than the previously formulated CBM methods.

\textsuperscript{1}Hamming distance between two strings or vectors of equal length is the number of positions at which the corresponding elements are different.
2.2.3 Constraints for improved CBM predictions

The prediction accuracy of CBM methods can be improved by implementing additional theoretical and experimentally-derived constraints. Thermodynamics-based Metabolic Flux Analysis (TMFA), developed by Henry et al. (2007), improves predictions of FBA by imposing thermodynamic constraints. The thermodynamic constraints ensure that the directions of fluxes within the metabolic network are consistent with the thermodynamic feasibilities of the corresponding reactions. The feasibility of a reaction is calculated from the Gibb’s free energy change (ΔG) for the reaction. For the reactions with unknown free energy change, standard Gibb’s free energy change (ΔG°) is first calculated using the group contribution method, which is then used to calculate ΔG (Henry et al., 2007).

All the reactions in a metabolic network are not active at all times, as gene expression in microorganisms is regulated. Metabolic models which do not use any regulatory constraints tend to over-estimate the metabolic capacity of microorganisms. To improve the predictive accuracy of metabolic networks, regulatory constraints can be added to metabolic models. Regulatory constraints were first introduced in FBA as Boolean rules by Covert et al. (2001). A more refined version of regulatory constraints was introduced by Chandrasekaran and Price (2010) in a method called Probabilistic Regulation of Metabolism (PROM). PROM implements regulatory constraints as quantitative bounds instead of Boolean rules, and predicts flux distribution that is close to the actual values by penalizing the regulatory rules that disagree with the experimentally observed gene expression levels.

Beg et al. (2007) improved the predictions of FBA by introducing molecular crowding constraints, which account for the physical limitations of the cell. The method called FBA with Molecular Crowding (FBAwMC) takes into account the crowding of enzymes in the cell by adding crowding constraints. FBAwMC is able to correctly predict the sequence of substrate utilization from a mixture of carbon sources (Beg et al., 2007), which regular FBA cannot predict. A similar concept has been extended to crowding of cell membrane by accounting for the fact that the cell membrane has a limited surface area and can accommodate a limited number of membrane proteins. By introducing the membrane crowding constraints, FBA predictions were improved for glucose uptake rates and the switch between respiratory and fermentative
Additional methods such as E-flux (Colijn et al., 2009), GIMME (Becker and Palsson, 2008), and IOMA (Yizhak et al., 2010) have been developed to improve the predictions of FBA by adding constraints based on experimental data obtained from transcriptomics, proteomics, and metabolomics experiments.

2.3 Model-based strain design

The CBM methods have found extensive applications in metabolic engineering design of industrial microorganisms. Typically, the strain design paradigms are formulated as bilevel optimization problems (one optimization problem nested inside another). The objective of the outer optimization problem maximizes the yield of the desired product and the inner optimization problem simulates the growth conditions. The problem is formulated such that the variables used for optimization of the outer problem change the connectivity of the metabolic network. The inner optimization problem simulates the growth of the altered (because of the outer problem) metabolic network using a common CBM method such as FBA or MOMA. A solution to the bilevel optimization problem is such that both the inner and the outer problems are optimized. In short, an optimum solution gives the best possible yield for the desired metabolite by altering the metabolic network, while ensuring maximum possible growth with the altered metabolic network.

OptKnock (Burgard et al., 2003) was the first strain design algorithm formulated. It has been used for designing strategies for overproduction of compounds such as lactic acid (Fong et al., 2005), 1,4-butanediol (Yim et al., 2011), and amino acids (Pharkya et al., 2003). OptKnock is a bilevel optimization problem with the outer problem maximizing the product formation and the inner problem maximizing the growth rate using FBA. The outer problem is optimized using binary variables that can either delete or retain a reaction in the metabolic network. Thus, OptKnock finds the reaction deletions that lead to maximization of product formation while forming the biomass at a specified rate. As there exists a trade off between the production of biomass and of the desired product, minimum biomass formation rate is chosen to ensure maximum flux towards the desired product. In its simplest form, the formulation for
OptKnock is shown below in equation (2.6):

\[
\begin{align*}
\max_{v, y} & \quad f_p^T v \\
\text{s.t.} & \quad \max_v f^T v \\
\text{s.t.} & \quad S v = 0 \\
& \quad v_i^L (1 - y_i) \leq v_i \leq v_i^U (1 - y_i), \quad i = 1, \ldots, N \\
& \quad \sum_{i=1}^{N} y_i \leq K \\
& \quad v_{\text{bio}} \geq v_{\text{bio}}^{min} \\
& \quad y \in \{0, 1\},
\end{align*}
\]

(2.6)

where, \( f_p \in \mathbb{R}^N \) is the vector of objective coefficients corresponding to maximization of the desired product, \( f \in \mathbb{R}^N \) is the vector of objective coefficients corresponding to maximization of biomass, \( y \) is a vector of binary variables (one corresponding to each reaction), \( K \) is the maximum limit on the reactions that can be deleted, and \( v_{\text{bio}}^{min} \) is the limit on minimum biomass production. The bounds for the flux variable \( v \in \mathbb{R}^N \) depend on the value the binary variable \( y \) assumes. When \( y_i = 0 \), the corresponding reaction is retained in the network and flux value of that reaction \( (v_i) \) lies between \( v_i^L \) and \( v_i^U \). When \( y_i = 1 \), the corresponding reaction is deleted from the network as the flux value of the reaction is forced to be 0.

An important feature of the OptKnock-like approaches for strain design is the inherently imposed coupling of growth to the product formation. In absence of such growth coupling, the organism can evolve to grow without production of the desired chemical within a few generations. The solutions found by OptKnock-like algorithms force the organism to produce the desired metabolite in order to grow.

OptKnock is the simplest strain design algorithm. However, the combinatorial nature of the problem makes it computationally intractable with genome-scale models. Faster algorithms that use alternative approaches to obtain solutions have been formulated. GDLS (Genetic Design through Local Search) developed by Lun et al. (2009) solves the bilevel problem using local search by sequentially predicting gene deletions. EMILiO (Enhancing Metabolism with Iterative Linear Optimization) developed by Yang et al. (2011) does not use binary variables. Using continuous variables, EMILiO solves the optimization problem using sequential linear
programming. Due to the use of continuous variables, EMILiO can identify up and down regulation of fluxes in addition to gene deletions to improve the yields of the desired products.

Other algorithms that can be considered as variations to OptKnock have been developed. OptORF finds gene deletions for both metabolic and regulatory network (Kim and Reed, 2010), whereas BiMOMA uses MOMA as the inner optimization problem instead of FBA (Kim et al., 2011b). More than 20 different strain design algorithms have been published and have been documented in (Lewis et al., 2012). A library of all the published constraint-based modeling algorithms is maintained at http://cobramethods.wikidot.com/seas. Some of the important strain design algorithms are described in Table 2.3.

### 2.3.1 Opportunities for advancement

Numerous bilevel optimization methods for strain design have been formulated and have been used to solve metabolic engineering problems. Increase in production yield is the most widely addressed metabolic engineering problem, as it is the most easily engineered trait. However, there exist potential opportunities for extending the bilevel optimization framework to solve alternative metabolic engineering challenges. Bilevel algorithms have been previously used for other applications (apart from metabolic engineering) such as identification of the correct objective function for metabolic models (Gianchandani et al., 2008), identification of enzyme capacity constraints (Yang et al., 2008), and as a tool to improve existing genome-scale models based on experimental data (Tervo and Reed, 2012). The bilevel algorithms can be extended to solve regulatory problems such as sugar co-utilization.

Algorithms such as OptORF that engineer the regulatory network in addition to the metabolic network have been previously formulated. These algorithms still are limited by the extent of knowledge of the regulatory networks. Methods such as OptORF can either not be implemented, or are inaccurate for organisms with uncharacterized regulatory networks.

Using clever formulations based on novel premises, the scope of bilevel optimization algorithms in metabolic engineering can be extended beyond the engineering for increasing the

---

2The algorithm FOCAL (FOrced Coupling ALgorithm) can be used to couple a non-measurable flux to a measurable flux. FOCAL was also used to couple utilization of glucose and xylose to growth, and predicted one strategy to force co-utilization of glucose and xylose. A similar strategy was also independently identified in this thesis using the SimUp algorithm
Table 2.3: Selected strain design algorithms published.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>OptStrain</td>
<td>Foreign metabolic pathways can be added in addition to gene knockouts.</td>
<td>(Pharkya et al., 2004)</td>
</tr>
<tr>
<td>OptGene</td>
<td>Genetic algorithm is used to improve strain design. May not converge to the global optimum.</td>
<td>(Patil et al., 2005)</td>
</tr>
<tr>
<td>OptReg</td>
<td>Allows up and down regulation of reactions in addition to gene deletions.</td>
<td>(Pharkya and Maranas, 2006)</td>
</tr>
<tr>
<td>GDLS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Simplifies computation using a local search gene deletion mutants. The genotype is iteratively improved by subsequent gene deletions.</td>
<td>(Lun et al., 2009)</td>
</tr>
<tr>
<td>RobustKnock</td>
<td>Predicts gene deletion strategies that lead to the overproduction of chemicals of interest by accounting for the presence of competing pathways in the network.</td>
<td>(Tepper and Shlomi, 2009)</td>
</tr>
<tr>
<td>OptForce</td>
<td>Identifies possible genetic manipulations by classifying whether a reaction flux MUST change to meet the desired target production.</td>
<td>(Ranganathan et al., 2010)</td>
</tr>
<tr>
<td>EMILiO&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Identifies gene knockout and flux ranges of reactions to give maximum yield. Uses successive linear programming.</td>
<td>(Yang et al., 2011)</td>
</tr>
<tr>
<td>BiMOMA</td>
<td>Identifies gene knockout strategies using MOMA as the inner problem.</td>
<td>(Kim et al., 2011b)</td>
</tr>
<tr>
<td>SimOptStrain</td>
<td>Finds gene addition and deletion targets in a single step to increase production yields.</td>
<td>(Kim et al., 2011b)</td>
</tr>
<tr>
<td>OptORF</td>
<td>Identifies gene knockout strategies while incorporating the known regulatory rules and including regulatory genes as potential targets.</td>
<td>(Kim and Reed, 2010)</td>
</tr>
<tr>
<td>GDBB&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Uses truncated branch and bound method to identify gene knockout strategies. May yield sub-optimal solutions.</td>
<td>(Egen and Lun, 2012)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Genetic Design through Local Search

<sup>b</sup> Enhancing Metabolism with Iterative Linear Optimization

<sup>c</sup> Genetic Design through Branch and Bound
product yields. Two such premises that haven’t been previously explored are used in this thesis.

- Regulatory networks in microorganisms generally control metabolism. Direct changes to the metabolic network may force the microorganism to alter its regulatory network as an adaptive response to these changes. Using this principle, new algorithms can be formulated that can address regulatory problems using the knowledge of metabolism alone.

- Mutants can show different growth/lethality characteristics on different growth media, a trait known as conditional lethality. Conditional lethality can be potentially used for engineering novel strategies where multiple media components are involved such as a mixture of sugars.

Based on these premises, this thesis introduces a completely novel algorithm to solve the problem of sequential sugar utilization. The problem incorporates the logic of altering the regulatory network using the knowledge of metabolism alone and utilizing the trait of conditional lethality of organisms by simulating multiple FBA problems nested under a single outer problem. The algorithm uses metabolic models of microorganisms and finds reaction deletion strategies to alter the metabolic network in such a manner that the organism is forced to co-utilize sugars and eliminate CCR.

2.4 Summary of the literature review

2.4.1 Sugar utilization

Utilization of sugars is a highly regulated phenomenon in microorganisms. Understanding the principles of sugar utilization in microorganisms is important to engineer their utilization characteristics. Efficient utilization of multiple sugars is important for production of biofuels and bio-based chemicals using second generation feedstocks. Accordingly, attempts have been made to engineer sugar utilization in important industrial organisms such as *E. coli* and *S. cerevisiae.* Following are some of the important considerations in sugar utilization in microorganisms.

- Most microorganisms utilize multiple sugar in a sequential manner. The phenomenon that controls the sugar uptake pattern is known as carbon catabolite repression (CCR),
• CCR is a complex regulatory phenomenon and involves multiple mechanisms. In *E. coli*, the glucose uptake system known as phosphoenolpyruvate-carbohydrate phosphotransferase system (PTS) plays a crucial role in CCR.

• Multiple *E. coli* mutants capable of co-utilizing glucose and xylose have been constructed by deleting regulatory genes involved in CCR. Some mutants show strict co-utilization, whereas others show partial elimination of CCR.

• Though construction of mutants by altering CCR genes may result in co-utilizing mutants, it is not the best approach to construct co-utilizing mutants. This approach may be difficult to use in less characterized microorganisms and may fail as the organisms may evolve to sequentially utilize sugars.

2.4.2 Metabolic modeling for strain engineering

Constraint-based modeling (CBM) methods have been extensively used to simulate microbial behaviour. Multiple CBM methods have been formulated with varying objectives such as prediction of mutant phenotypes, prediction of interactions between microbial communities, and for metabolic engineering. Following are some important characteristics of CBM and its applications in strain engineering.

• CBM can be used to accurately predict growth rates and gene lethality in microorganisms. Flux Balance Analysis (FBA) is the most common CBM which typically uses growth maximization objective function and stoichiometric constraints to predict flux distribution.

• Alternatives to FBA such as MOMA, ROOM, and RELATCH have been developed that predict the phenotypes of the gene deletion mutants with higher accuracy.

• Strain design algorithms based on CBM have been developed, which predict gene knockout strategies to achieve metabolic objectives such as improved product yield. Such algorithms are formulated as bilevel optimization problems. There is a potential for extending such optimization algorithms to solve regulatory problems.
2.5 Hypotheses, objectives, and organization of the thesis

Based on the opportunities for advancement discussed above, this thesis was focused on following objectives:

1. Construction of glucose-xylose co-utilizing mutants

Two problems were identified with construction of glucose-xylose co-utilizing strains: 1) engineering for glucose-xylose co-utilization requires detailed knowledge of the regulatory mechanism controlling sugar utilization in the organism, 2) deletion of regulatory genes to force co-utilization of sugars can have undesired physiological changes. To address these two problems, a completely novel approach based on an original hypothesis was adopted in this thesis.

**Hypothesis 1:** Sugar co-utilization can be forced in an organism by deleting appropriate metabolic genes without altering (deleting or replacing) any genes from the regulatory network.

**Objectives:** Based on the first hypothesis, following are the two main objectives of this thesis:

1. To develop a novel strain design algorithm based on metabolic models to solve the sequential sugar utilization problem in industrial microorganisms.

2. To construct and characterize the algorithm predicted microorganisms for glucose-xylose co-utilization.

The results based on these objectives are distributed in two chapters. Chapter 3 describes the development of the novel algorithm SimUp, and Chapter 4 discusses the results from construction and characterization of the *E. coli* mutants based on SimUp predictions.

2. Identification of mutations in sugar co-utilizing mutants

If an organism is forced to co-utilize sugars by metabolic gene deletions, the regulatory network responsible for CCR should to be altered to eliminate CCR. This could involve accumulation of mutations in one or more of the regulatory genes. Recent studies have shown that even metabolic genes might have roles to play in CCR (for example, *mgsA*). Identification of such mutated genes could yield novel engineering targets for forcing co-utilization, and could also provide deeper insights into the sugar utilization mechanisms of the microorganism. The second hypothesis of this thesis is based on this premise.
Hypothesis 2: *If sugar co-utilization is forced on a microorganism by deleting its metabolic genes, the organism will accumulate mutations in the genes that control sugar utilization pattern.*

Objectives: Based on the second hypothesis, one of the major objectives of this thesis is to identify novel mutations in glucose-xylose co-utilizing mutant(s).

Targeted Sanger sequencing and whole genome sequencing was used to identify novel mutations, if any, in one of the mutants described in Chapter 4. The results based on the sequencing studies are summarized in Chapter 5.

3. Production of valuable chemical from sugar co-utilizing mutant

Sugar co-utilization is considered useful in LCB-based industrial processes. After metabolic engineering of organisms for sugar co-utilization, further metabolic engineering is required to produce useful chemicals from sugar co-utilizing mutants. Effect of sugar co-utilization is anticipated to positively affect the production of the desired chemicals by simplifying the fermentation process and potentially increasing the productivity. Accordingly, the effects of sugar co-utilization on product formation are hypothesized.

Hypothesis 3: *Sugar co-utilization can improve the fermentation processes used for producing valuable chemicals.*

Objectives: Based on the above hypothesis, one of the objectives of the thesis is to engineer the glucose-xylose co-utilizing mutants of *E. coli* to produce D-ribose.

The results from engineering and characterization of D-ribose production using a glucose-xylose co-utilizing mutant and the wild-type are summarized in Chapter 6. The effect of sugar co-utilization is tested on common fermentation parameters such as yield, productivity, and titre.

An additional objective has been explored in this thesis. Systematic investigation on the effect of order of gene deletions at sub-optimal growth phase has never been reported. Using the mutants constructed in this study, characterization of differences in the growth physiology of multiple gene knockout mutants (with different order of gene deletions) was carried out. As the investigation was not directly related to the sugar co-utilization problem, the results of this study have been discussed in Appendix A of this thesis.
Finally, Chapter 7 summarizes the conclusions obtained in this thesis and discusses the contribution of this thesis to the current knowledge in the area of metabolic engineering and sugar utilization physiology. Chapter 8 suggests the future directions to the research presented in this thesis.

2.6 Unifying theme of the thesis

The central theme of this thesis is model-guided metabolic engineering of industrial microorganisms for sugar co-utilization. The thesis addresses the problem of sequential sugar utilization in industrial microorganisms. Emphasis has been placed on using modeling-based techniques to identify metabolic engineering strategies and experimental validation of the predicted strategies. This thesis also explores multiple related frontiers. The major thematic undertones of this thesis aim to address the inaccuracies in metabolic models, enhance the understanding of metabolism, demonstrate the utility of glucose-xylose co-utilizing mutants, and gain novel insights into CCR mechanism of \textit{E. coli}.

As summarized in the literature review, numerous algorithms have been developed for predicting metabolic engineering strategies. However, experimental validation of model-predictions still remains relatively uncommon in practice. Experimental validation of model predictions is important, as it can potentially reveal incorrect predictions and/or incorrect assumptions in the models. By addressing the inconsistencies between experimental results and model predictions, the accuracy of model predictions can be improved. This thesis provides one such example of model correction based on experimental results.

Characterization of metabolically engineered mutants also has potential to yield novel insights into metabolism. Due to the complexity of metabolic networks, model-based predictions may sometimes be non-intuitive. Investigation of such non-intuitive predictions can enhance our understanding of metabolism. A non-intuitive gene knockout strategy is validated in this thesis, which underscores the effect of stoichiometric constraints of metabolism.

A key objective of metabolic engineering is to obtain useful strains for industrial applications. Metabolic engineering strategies work by channeling carbon-flux through the desired pathways. A component of this thesis also focuses on demonstrating the utility of the glucose-xylose co-
utilizing mutants constructed in this study. One of the mutants is reengineered to produce D-ribose from glucose-xylose mixture. The effect of diverting carbon flux differently from the wild-type can improve the yields of the target products. This has been demonstrated for the product D-ribose in this thesis.

Mutants with multiple gene knockouts (higher-order mutants) are important in the study of microbial physiology. Such mutants show properties that are not displayed by the wild-type or single gene knockout mutants. The mutants constructed in this thesis are unique, and provide an opportunity to understand cellular responses to drastic genetic alterations. This thesis focuses mainly on studying the regulatory phenomenon of CCR using the mutants constructed. Though CCR in *E. coli* has been studied in great detail, there remains a possibility of existence of unknown mechanisms. Discovery of a novel mechanism can yield novel engineering targets for glucose-xylose co-utilization and may shed light on the fundamental question of why organisms use CCR-controlled sequential sugar utilization. With focus on the mutational changes caused due to forced glucose-xylose co-utilization phenotype, whole genome of a mutant is sequenced. The aim of this sequencing study is to possibly uncover hitherto unknown participants of CCR.

Using the mutants constructed in the study, a slightly unrelated (to co-utilization) phenomenon in microbial physiology is also investigated. The effect of order of gene deletions in higher-order mutants is an unexplored phenomenon. Different order of gene knockouts is anticipated to result in different sub-optimal growth states. To investigate whether such different sub-optimal growth states exist, higher-order mutants with *pgi* deletion are characterized.

In summary, this thesis should be viewed as a targeted approach to solve a problem of industrial relevance, along with fundamental investigations on metabolic and genetic responses of the cell.
Chapter 3

SimUp: a Bilevel Optimization Algorithm

This chapter contains material from our publication (Gawand et al., 2013):


This chapter consists of both the main manuscript and Supporting Information from the citation above. Reproduction of the material from the publication in this thesis is a right that has been granted by Elsevier (the publisher) to the authors of the manuscript.

3.1 Abstract

Computational algorithms based on genome-scale metabolic models are being increasingly used for metabolic engineering of industrial microorganisms. Most algorithms developed previously have focused on increasing the product yield by engineering metabolic network; however, improvements in other fermentation characteristics might require engineering of regulatory processes. Engineering of regulatory processes can be non-trivial due to the complexity of regulatory networks. In addition, regulatory networks may not be completely elucidated for non-conventional microorganisms. In this chapter, we describe a novel bilevel optimization algo-
algorithm, SimUp\textsuperscript{1}, that predicts strategies to eliminate the regulatory process of carbon catabolite repression (CCR) responsible for sequential utilization of sugars in industrial microorganisms. SimUp uses preconstructed metabolic models to identify reaction deletions to force co-utilization of two sugars. It is the first algorithm that eliminates a regulatory process by manipulating the metabolic network, thereby obviating the need for detailed knowledge of the regulatory network. Using SimUp and central metabolism model of \textit{Escherichia coli}, we identified reaction knockout strategies to force \textit{E. coli} to co-utilize glucose and xylose. Three of the eleven solutions predicted by SimUp were non-intuitive, and could not be explained based on the connectivity of the metabolic network. Stoichiometric feasibility analysis was carried out to understand the logic behind these non-intuitive strategies.

\section*{3.2 Introduction}

Constraint-based modeling has been successfully used to predict the physiology of microorganisms under environmental and genetic perturbations using the genome-scale reconstructions of metabolic networks (Ibarra et al., 2002; Fong and Palsson, 2004). Multiple algorithms have since been developed that use the genome-scale metabolic models to identify genetic strategies to increase the production yield of chemicals (Kim et al., 2012b; Lewis et al., 2012). OptKnock was the first bilevel optimization algorithm that was used to predict metabolic engineering strategies to increase the production of chemicals such as lactic acid and succinic acid, while coupling the production of these chemicals to growth (Burgard et al., 2003). \textit{E. coli} mutants based on OptKnock-predicted strategies were constructed to produce lactic acid (Fong et al., 2005) and 1,4-butanediol (Yim et al., 2011). More efficient algorithms such as EMILiO (Yang et al., 2011) and GDLS (Lun et al., 2009) were subsequently developed to address the issue of scalability associated with OptKnock. These new algorithms had significantly improved computational times, and unlike OptKnock, could be efficiently used with genome-scale metabolic models. The algorithm OptREG, developed by Pharkya and Maranas (2006), expanded the solution space of OptKnock by allowing reaction additions and up/down-regulation in addition to reaction deletions. Similarly, the algorithm OptORF, developed by Kim and Reed (2010),

\textsuperscript{1}The name of the algorithm, SimUp, is a portmanteau of the term ‘Simultaneous Uptake’.
expanded the solution space by including regulatory targets in addition to metabolic targets. All these previously developed metabolic engineering algorithms focused on improving the yield of the desired products.

In addition to yield, other features such as productivity and titre of the desired product also play a major role in determining the success of a fermentation process (Van Dien, 2013; Stephanopoulos, 2007). Simplicity and ease of control of fermentation processes become critically important at industrial scales. For example, ethanol fermentation processes based on sugar mixtures, such as those obtained from lignocellulosic hydrolysate, are difficult to control and are inefficient under fed-batch mode, and hence have proven difficult to operate at scale (Kim et al., 2012a). The complexity of mixed-sugar fermentation processes arises from the microorganism’s differential preference towards different sugars, resulting in a sequential utilization pattern of sugars from a sugar mixture. The sequential utilization of sugars is controlled by the regulatory phenomenon known as carbon catabolite repression (CCR) (Escalante et al., 2012; Görke and Stülke, 2008a). Microorganisms which show simultaneous utilization of sugars are very useful for development of fermentation processes based on lignocellulosic biomass (Stephanopoulos, 2007; Kim et al., 2010; Huffer et al., 2012; Kim et al., 2012a).

One of the common challenges in engineering regulatory processes such as CCR is that the participating proteins play global roles in cellular physiology. Thus, the effect of a major manipulation to a regulatory gene can be difficult to predict, and can be deleterious to the cell. Additionally, the regulatory network of non-conventional microorganisms may be not be completely elucidated, making its engineering difficult. In this study, we develop a novel algorithm, SimUp, to predict the metabolic reaction deletion strategies to eliminate CCR phenotype in the industrial microorganisms. The central idea and the novelty of SimUp lies in the fact that it alters the metabolic network to overcome a regulatory process, without any direct alteration to regulatory genes.

The SimUp algorithm was formulated as a bilevel optimization problem, and was used to identify strategies to force co-utilization of glucose-xylose in E. coli. Though similar in framework to the previously developed algorithms, SimUp’s distinguishing feature is the use of multiple external media conditions in a single optimization problem. The section below describes the logic and formulation of the SimUp algorithm.
3.3 Logic of SimUp

The SimUp algorithm is based on a simple metabolic logic. The lethality of metabolic gene deletions depends on the external nutrient conditions, a phenomenon known as conditional lethality (Harrison et al., 2007). For instance, the gene deletions that are lethal on glucose as a carbon source may not be lethal on xylose as a carbon source. Using the phenomenon of conditional lethality, it is possible to find gene deletions that are supported only by a mixture of sugars, and not supported by the individual components of the mixture. Accordingly, the SimUp algorithm searches for the reaction deletion combinations that are lethal in presence of either glucose or xylose, but are not lethal in presence both glucose and xylose provided as a mixture. As the reaction deletions do not allow the organism to grow on either glucose or xylose alone, the organism cannot utilize these sugars one at a time. However, as the organism can survive on a mixture of glucose and xylose, the organism is forced to consume both the sugars simultaneously. A schematic of the bilevel optimization problem for the SimUp algorithm is shown in Figure 3.1.

3.4 Materials and Methods

3.4.1 SimUp algorithm formulation

The SimUp algorithm was formulated as a bilevel optimization problem as shown in equation (3.1).

$$\max_{y, v} \quad F(\mu)$$

s.t.

$$\max_v \quad \sum_{k \in \{g,x, gx\}} (e^k)^T v^k$$

s.t. $S^k v^k = 0 \quad \forall k \in \{g,x, gx\}$

$$y_j(v_j^L)^k \leq v_j^k \leq y_j(v_j^U)^k \quad \forall k \in \{g,x, gx\}, \forall j = 1, \ldots, N$$

$$v_j^k \in \mathbb{R} \quad \forall k \in \{g,x, gx\}, \forall j = 1, \ldots, N$$

$$\sum_{j=1}^{N} (1 - y_j) \leq KO$$

$$y_j \in \{0,1\} \quad \forall j = 1, \ldots, N,$$
Figure 3.1: Schematic of the bilevel optimization problem for SimUp. Three different optimization problems, each simulating growth in presence of different sugar conditions, are nested in an outer problem. The outer problem maximizes the growth on glucose and xylose mixture while penalizing the growth on individual sugars. The inner problems simulate the growth of the organism under three different media conditions: 1) growth on glucose and xylose, 2) growth on glucose alone, and 3) growth on xylose alone. To maximize the outer objective function, growth is desired only under condition 1 (represented in green), and not under conditions 2 and 3 (represented in red).

where, \( \mu \) represents growth rate (h\(^{-1}\)), \( c \) is the vector containing coefficients of the objective function, \( S \) is the stoichiometric matrix, \( v \) (mmol/gDW/h) is the flux variable, \( v^L \) and \( v^U \) are the lower and upper bounds for the flux variable, respectively, \( N \) is the number of reactions, \( KO \) is the limit on the number of knockouts, and \( y \) is the binary variable for the outer objective problem. The superscripts \( g \), \( x \), and \( gx \) represent the different environmental conditions as explained below in the definition of the outer objective function \( F(\mu) \). The detailed description of the problem and the step-by-step procedure to solve the problem are discussed in the following sections.
The outer objective function

The outer objective function for the SimUp algorithm is defined in equation (3.2), where, $\mu$ represents growth rate, the subscript $w$ denotes the wild-type, and the superscripts denote the substrate availability conditions: $gx$ denotes the presence of both glucose and xylose, $g$ denotes the presence of glucose alone, and $x$ denotes the presence of xylose alone. Thus, the term $\mu^g_w$ represents the growth rate of the wild-type in presence of glucose alone. In the objective function $F$, the growth rate of the mutant under each growth condition is normalized to the growth rate of the wild-type under the exact same condition. Maximization of the objective function results in maximization of the growth rate of the mutant in presence of glucose and xylose, while minimizing the growth rate in presence of glucose alone and xylose alone. As the maximum possible growth rate of the mutant can not exceed the corresponding growth rate of the wild-type, the maximum value $F$ can attain is 1. This is the case when the growth rate of the mutant on glucose and xylose is equal to growth rate of the wild-type ($\mu^{gx} = \mu^{gx}_w$), and growth rate of the same mutant on glucose alone and xylose alone is 0.

$$F(\mu) = \frac{\mu^{gx}}{\mu^{gx}_w} - \frac{\mu^g}{2\mu^{gw}_w} - \frac{\mu^x}{2\mu^{xw}_w}$$ (3.2)

Variables

The decision variables for the outer optimization problem were the binary variables stored as the vector $y$. For each reaction in the metabolic model, there was one corresponding binary variable which could either retain or delete the reaction in the metabolic network. The deletion or retention of a reaction in the network was brought about by multiplying the corresponding binary variable to the upper and lower bounds of that reaction as shown in equation (3.3). If the value of the binary variable was equal to 1, the corresponding reaction was active in the network, and if the value of the variable was equal to 0, the reaction was deleted from the network.

$$v^L_j \cdot y_j \leq v_j \leq v^U_j \cdot y_j \quad \forall j = 1 \ldots N$$

$$y_j = \begin{cases} 
1 & \text{if the reaction is active,} \\
0 & \text{if the reaction is inactive.} 
\end{cases}$$ (3.3)
The nested optimization problems in the SimUp algorithm were three FBA problems, each simulating a different growth environment represented by the superscripts $g$, $x$, and $gx$ of the decision variable $v$. Corresponding to each environmental condition, there were three flux variables for the inner FBA problems represented by, $v^g$, $v^x$, and $v^{gx}$.

**Constraints**

The SimUp formulation consists of two levels of optimization. The variable $y$ of the outer optimization problem (maximization of $F(\mu)$) was constrained by the limit on the number of reaction knockouts, denoted by $KO$. The variables of the inner FBA problems were constrained by the stoichiometric constraints ($Sv = 0$), and by the lower and upper bounds $v^L$ and $v^U$. The lower and upper bounds on the flux variables of the inner optimization problems were used to define the media conditions for each nested FBA problem. For example, to simulate the growth on glucose medium, the lower and upper bounds for the glucose uptake reaction ($v^L_j$ and $v^U_j$, where $j$ corresponds to glucose uptake reaction), were appropriately adjusted, while setting the lower and upper bounds of the xylose uptake reaction ($v^L_j$ and $v^U_j$, where $j$ corresponds to xylose uptake reaction) to 0.

### 3.4.2 Solution to the SimUp algorithm

The SimUp algorithm was solved in five major steps listed below:

1. segregation and stacking of the inner FBA problems,
2. formulation of the dual problem for the stacked FBA problem (primal problem),
3. transformation of the bilevel SimUp algorithm to a single-level mixed-integer linear program (MILP),
4. transformation of the bilinear constraints in the dual problem into linear constraints using Glover transform, and
5. solving the final MILP using the central metabolism model of *E. coli*.

The details of the above-listed steps are discussed in the following section.
Segregation and stacking of the FBA problems

First, the inner FBA problems of the SimUp algorithm were reformulated by segregating the reversible and irreversible reactions. Segregation of the problems was necessary for formulation of the dual problem in the subsequent step. Segregation of FBA problems involved defining new variables $v_R$ and $v_I$ corresponding to the reversible and irreversible flux variables, respectively. Accordingly, the stoichiometric matrix $S$ was segregated into $S_R$ and $S_I$, and the bound vectors were segregated as $v^L_R$ and $v^U_R$ for the lower bounds, and $v^L_I$ and $v^U_I$ for the upper bounds. Finally, the flux variables bound with equality constraints (where $v^L = v^U$; for example, ATP maintenance reaction) were implicitly added to the stoichiometric constraints resulting in the modified stoichiometric constraints $Sv = b$. Equation (3.4) shows the formulation of a segregated FBA problem.

$$\max_{v_R, v_I} \sum_{j=1}^{N_R} c_{Rj}^T v_{Rj} + \sum_{j=1}^{N_I} c_{Ij}^T v_{Ij}$$
subject to

$$S_{Ri} v_R + S_{Ii} v_I = b_i \quad \forall i = 1 \ldots M$$
$$v^L_{Rj} \leq v_{Rj} \leq v^U_{Rj} \quad \forall j = 1 \ldots N_R$$
$$v^L_{Ij} \leq v_{Ij} \leq v^U_{Ij} \quad \forall j = 1 \ldots N_I,$$

where, $M$ is the number of metabolites, $N_R$ is the number of reversible reactions, and $N_I$ is the number of irreversible reactions. As the constraints of the three FBA problems were linearly independent, the segregated FBA problems could be stacked into a single linear program (LP). After the segregation and stacking of FBAs, the bilevel optimization problem for SimUp could
be reformulated as shown in equation (3.5).

\[
\max_y \quad F(\mu) \\
\text{s.t.} \\
\max_{v_R,v_I} \sum_{k \in \{g,x,gx\}} \left( \sum_{j=1}^{N_R} (c_R)^k_j (v_R)_j^k + \sum_{j=1}^{N_I} (c_I)^k_j (v_I)_j^k \right)
\]

\[
\text{s.t.} \\
S_R i v_R^k + S_I i v_I^k = b_i^k \quad \forall i = 1 \ldots M, \quad \forall k \in \{g,x,gx\}
\]

\[
(y)_j^k (v_R)_j^k \leq (y)_j^k (v_I)_j^k \leq (y)_j^k (v_R)_j^k \quad \forall j = 1 \ldots N_R, \quad \forall k \in \{g,x,gx\}
\]

\[
(y)_j^k (v_R)_j^k \leq (v_I)_j^k \leq (y)_j^k (v_I)_j^k \quad \forall j = 1 \ldots N_I, \quad \forall k \in \{g,x,gx\}
\]

\[
(v_R)_j^k \in \mathbb{R} \quad \forall j = 1 \ldots N_R, \quad \forall k \in \{g,x,gx\}
\]

\[
(v_I)_j^k \in \mathbb{R} \quad \forall j = 1 \ldots N_I, \quad \forall k \in \{g,x,gx\}
\]

\[
\sum_{j=1}^{N_R+N_I} (1 - y_j) \leqKO
\]

\[
y_j \in \{0,1\} \quad \forall j = 1, \ldots, N_R + N_I
\]

The stacked LP in the SimUp problem was the primal problem of the LP theory. The primal objective denoted by \(P(v)\) was given as,

\[
P(v) = \sum_{k \in \{g,x,gx\}} \left( \sum_{j=1}^{N_R} (c_R)^k_j (v_R)_j^k + \sum_{j=1}^{N_I} (c_I)^k_j (v_I)_j^k \right) \quad (3.6)
\]

**Formulation of the dual problem**

To convert the bilevel problem into a single level problem, a dual LP was formulated corresponding to the primal LP of the stacked FBAs. For every constraint in the primal problem, there exist a dual-variable, and for every variable in the primal primal problem, there exists a dual-constraint. The constraints in the primal problem were the stoichiometric constraints, the constraints on the reversible fluxes, and the constraints on the irreversible reaction fluxes. Accordingly, the dual LP had corresponding dual variables, denoted by \(\lambda\), for each of the above constraints. The dual objective function with all the dual variables is shown in equation (3.7)
below:

\[ D(\lambda) = \sum_{k \in \{g,x,gx\}} b_k^k (\lambda_{\text{stoich}})_i^k + \sum_{j=1}^{N_R} (v_{Rj}^L)^k (y_j)_{Rj}^k + \sum_{j=1}^{N_I} (v_{Ij}^U)^k (y_j)_{Ij}^k \]  

(3.7)

The dual problem of minimization of \( D(\lambda) \) was subject to dual constraints corresponding to the primal variables \( v_R \) and \( v_I \). By adding the constraints corresponding to the variables of the primal problem, the complete dual LP was formulated as shown in equation (3.8) below:

\[
\begin{align*}
\min_{\lambda} & \quad D(\lambda) \\
\text{s.t.} & \quad \sum_{i=1}^{M+2N_R+N_I} \lambda_i^k \begin{bmatrix} S_i^k & I_R & I_R & 0 \end{bmatrix}^T = c_{Rj}^k \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall j = 1 \ldots N_R \\
& \quad \sum_{i=1}^{M+2N_R+N_I} \lambda_i^k \begin{bmatrix} S_i^k & 0_R & 0_R & I_I \end{bmatrix}^T \geq c_{Ij}^k \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall j = 1 \ldots N_I \\
& \quad -(1 - y_j) \lambda_{\text{min}}^k \leq (\lambda_{Ij}^U)^k \leq \lambda_{\text{max}}^k \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall j = 1 \ldots N_I \\
& \quad -(1 - y_j) \lambda_{\text{min}}^k \leq (\lambda_{Rj}^L)^k \leq \lambda_{\text{max}}^k \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall j = 1 \ldots N_R \\
& \quad -(1 - y_j) \lambda_{\text{min}}^k \leq (\lambda_{Ij}^U)^k \leq \lambda_{\text{max}}^k \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall j = 1 \ldots N_I \\
& \quad (\lambda_{\text{stoich}})_i^k \in \mathbb{R} \\
& \quad \forall k \in \{g,x,gx\}, \quad \forall i = 1 \ldots M,
\end{align*}
\]

(3.8)

where, \( I_R \) and \( I_I \) are the identity matrices of dimension \( N_R \times N_R \) and \( N_I \times N_I \), respectively, and \( 0_R \) and \( 0_I \) are the zero matrices of the dimension \( N_R \times N_R \) and \( N_I \times N_I \), respectively.

**Linearization of the objective function of the dual problem**

The objective function of the dual problem defined in equation (3.7) is non-linear due to the presence of the binary variables \((y_j)\). A bilinear product, i.e., a product of a continuous variable
and a binary variable can be linearized using Glover transform (Glover, 1975). To transform the objective function into a linear function, three new continuous variables were introduced, each corresponding to a bilinear product in the dual objective. Each of these variables was constrained by two linear constraints according to the method described by Glover (1975). For example, to convert the bilinear product, \( y_j \lambda_{LR}^L \) into a continuous variable, a new variable \( z_{LR}^L \) was defined such that,

\[
(z_{LR}^L)_j = y_j \lambda_{LR}^L
\]  

subject to,

\[
(\lambda_{LR}^L)^{\text{min}} y_j \leq (z_{LR}^L)_j \leq (\lambda_{LR}^L)^{\text{max}} y_j \quad (3.10a)
\]

\[
(\lambda_{LR}^L)_j - (\lambda_{LR}^L)^{\text{max}} (1 - y_j) \leq (z_{LR}^L)_j \leq (\lambda_{LR}^L)_j - (\lambda_{LR}^L)^{\text{min}} (1 - y_j) \quad (3.10b)
\]

The bilinear dual objective function was thus converted into an equivalent linear objective function by adding three new variables and six new constraints.

**Conversion of the bilevel problem to single-level problem**

The final step in solution of the SimUp problem was to convert the bilevel problem into a single-level LP using the strong duality theorem. According to the strong duality theorem of LP theory, the value of the objective function of the primal and the dual are equal at their corresponding optima. The inner LP problem could thus be reformulated as a set of constraints (as opposed to an optimization problem) by writing the problem as a primal-dual equality \( P(v) = D(\lambda) \), subject to both primal and dual constraints. Using this principle, the bilevel SimUp problem was converted into a single-level problem for maximization of the objective function \( F(\mu) \), subject to the strong duality constraint, primal constraints, dual constraints, and Glover transform constraints. Due to the presence of binary variable \( y \), the final single-level problem was a mixed integer linear program (MILP), which could be solved using an off-the-shelf LP solver. The final formulation of the SimUp problem is shown in the equation below:
\[
\begin{align*}
\max_{\nu, \lambda, z, y} & \quad F(\mu) \\
\text{s.t} & \quad P(\nu) = D(\lambda) \\
& \quad S_R v_R^k + S_I v_I^k = b_i^k \quad \forall k \in \{g, x, gx\}, \; \forall i = 1 \ldots M \\
& \quad (y)_j (v_R^k)_j \leq (v_I^k)_j \leq (y)_j (v_I^k)_j \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (y)_j (v_I^k)_j \leq (v_I^k)_j \leq (y)_j (v_I^k)_j \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_I \\
& \quad \sum_{i=1}^{M+2N_R+N_I} \lambda_i^k \begin{bmatrix} S_R^k & I_R & I_R & 0_I \end{bmatrix}^T = c_{R_j}^k \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad \sum_{i=1}^{M+2N_R+N_I} \lambda_i^k \begin{bmatrix} S_I^k & 0_R & 0_R & I_I \end{bmatrix}^T \geq c_{I_j}^k \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_I \\
& \quad -(1 - y_j) \lambda_{\min} \leq (\lambda_R^U)_j^k \leq \lambda_{\max} \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad -\lambda_{\min} \leq (\lambda_R^L)_j^k \leq (1 - y_j) \lambda_{\max} \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad -(1 - y_j) \lambda_{\min} \leq (\lambda_R^L)_j^k \leq \lambda_{\max} \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_R^L)_j^k \leq (z_R^L)_j^k \leq (\lambda_R^L)_j^k \lambda_{\min} (1 - y_j) \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_R^L)_j^k - (\lambda_R^L)_j^k \max (1 - y_j) \leq (z_R^L)_j^k \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_R^U)_j^k \leq (\lambda_R^U)_j^k \max y_j \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_R^U)_j^k - (\lambda_R^U)_j^k \max (1 - y_j) \leq (z_R^U)_j^k \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_I^U)_j^k \leq (\lambda_I^U)_j^k \max y_j \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_I \\
& \quad (\lambda_I^U)_j^k - (\lambda_I^U)_j^k \max (1 - y_j) \leq (z_I^U)_j^k \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_I \\
& \quad \sum_{j=1}^{N_R+N_I} (1 - y_j) \leq KO \\
& \quad (v_I)_j^k \geq 0 \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_I \\
& \quad (v_R)_j^k \in \mathbb{R} \quad \forall k \in \{g, x, gx\}, \; \forall j = 1 \ldots N_R \\
& \quad (\lambda_{\text{stoich}})_i^k \in \mathbb{R} \quad \forall k \in \{g, x, gx\}, \; \forall i = 1 \ldots M \\
& \quad y_j \in \{0, 1\} \quad \forall j = 1 \ldots N_R + N_I
\end{align*}
\]
The MILP was solved using CPLEX 11.2 (ILOG) in AMPL (ILOG) environment. A previously described model for central metabolism of *E. coli* was used as input to the problem. A step-by-step protocol to solve the SimUp algorithm for the *E. coli* model is described in Appendix B.

### 3.4.3 Stoichiometric feasibility analysis

To understand the logic behind the non-intuitive solutions predicted by SimUp, additional analysis of the metabolism of the mutants was carried out. This analysis involved assessing the feasibility of formation of each biomass precursor. Growth of an organism is a direct consequence of its ability to synthesize all the biomass precursors using the available metabolic pathways. As SimUp identified strategies for deletion of metabolic reactions, the mutants were restricted to use only certain pathways to make the biomass precursors. If a mutant could not synthesize all the biomass precursors from a given carbon source, the reaction deletion strategy of that mutant was considered lethal on that particular carbon source.

To identify exactly which biomass precursors could be synthesized by the mutants, under given media conditions, stoichiometric feasibility analysis was carried out. To assess the stoichiometric feasibility of the biomass precursors, an FBA-like problem was formulated using the objective function to maximize biomass precursor instead of growth. The stoichiometric matrix was modified by adding an exchange flux for the biomass precursor being maximized. The maximization problem shown in the equation (3.12) was iteratively solved to find the feasibility of formation of each biomass precursor in the background of reaction deletions of the predicted mutants. A similar method has been used in a previous study to find network gaps in metabolic network of *E. coli* (Feist et al., 2007).

\[
\begin{align*}
\max_v & \quad v_{\text{precursor}} \\
\text{s.t.} & \quad Sv = 0 \\
& \quad v^L \leq v \leq v^U \\
& \quad v \in \mathbb{R}^N,
\end{align*}
\tag{3.12}
\]

where, \(v_{\text{precursor}}\) is the flux through the exchange reaction for the biomass precursor, \(S\) is the stoichiometric matrix, \(v\) is the flux vector, \(v^L\) and \(v^U\) are the lower and upper bounds on the
Chapter 3. SimUp: a Bilevel Optimization Algorithm

flux variables, respectively, and $N$ is the number of reactions. All the constraints on the cofactors were removed by allowing free exchange and supply. The problem for stoichiometric feasibility was solved in the MATLAB (MathWorks Inc.) environment using COBRA Toolbox v2.0 (Schellenberger et al., 2011).

3.5 Results

The objective of the SimUp algorithm was to find reaction deletion combinations to produce a specific phenotype: growth on a mixture of glucose and xylose, but no growth on either glucose or xylose alone. The limit on the number of reaction knockouts ($KO$) for SimUp was a predetermined parameter. Solution to SimUp was started with the knockout limit set to 1 ($KO = 1$). None of the single reaction knockout mutants were found to have the desired phenotype (corresponding to the SimUp objective $F(\mu) = 1.0$). Similarly, no solutions were found for the value of $KO$ set to 2. The minimum number of reaction deletions required to achieve the desired phenotype was found to be 3. Multiple triple reaction knockout mutants showed the desired phenotype required to force co-utilization of glucose and xylose. A total of 10 solutions with three reaction deletions were identified by SimUp. Only one solution was identified with $KO$ set to 4.

The eleven solutions predicted by SimUp were classified into groups based on the logic behind the solutions. For the classification, the solutions were considered as auxotrophies that could be rescued only with a combination of two sugars: glucose and xylose. Growth on individual sugars (either glucose or xylose) was not possible because the individual sugars could not be used to synthesize the complete set of biomass precursors. The complete set of biomass precursors could be synthesized only if both the sugars were consumed simultaneously. Using this logic, the solutions were grouped together based on the unique set of biomass precursors that could be synthesized from each sugar. A group of equivalent solutions was called a strategy.

As SimUp was solved using the central metabolism model of $E. \ coli$, the solutions were further verified using the genome-scale metabolic models, iJR904 (Reed et al., 2003) and iAF1260 (Feist et al., 2007), by simulating the SimUp predicted gene deletions. As iAF1260 was a more recent model than iJR904, the simulations with iAF1260 were considered more accurate. It
was found that the strategies identified by SimUp using central metabolism model were not sufficient to obtain similar results with the genome-scale models. The genome-scale models required additional reaction knockouts to obtain the desired phenotype. All the reaction deletion combinations and their classification into strategies are summarized in Table 3.1.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Reaction deletions</th>
<th>Growth rate $(h^{-1})^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Central model</td>
<td>iAF1260$^b$</td>
</tr>
<tr>
<td>Strategy A</td>
<td>pgi, gnd, eda</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pgi, gnd, edd</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>pgi, gnd, fbp</td>
<td></td>
</tr>
<tr>
<td>Strategy B</td>
<td>pgi, rpe, eda</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pgi, rpe, edd</td>
<td>DRPA, XYLI2, PPS</td>
</tr>
<tr>
<td></td>
<td>pgi, rpe, tpi</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pgi, rpe, fba</td>
<td></td>
</tr>
<tr>
<td>Strategy C</td>
<td>pgi, gnd, fbp, pfk</td>
<td>XYLI2, F6PA, DRPA</td>
</tr>
<tr>
<td></td>
<td>pgi, gnd, tpi</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Growth rates were calculated using glucose and xylose exchange fluxes set to -5 mmol/gDW/h each.

$^b$ Reactions deleted from the genome-scale model in addition to the reactions deleted from the central metabolism model.

DRPA: deoxyribose-phosphate aldolase,
XYLI2: xylose isomerase reaction converting glucose into fructose (*in vitro* evidence),
PPS: phosphoenolpyruvate synthetase,
F6PA: fructose-6-phosphate aldolase.
3.5.1 Strategy A and the mutant LMSE$_1$

Strategy A consisted of three solutions which were grouped together based on the unique set of biomass precursors that could be synthesized from glucose and xylose. All the solutions allowed the synthesis of only glucose-6-phosphate (G6P) from glucose, and all the remaining precursors were synthesized from xylose. One of the solutions, with three reaction deletions, pgi, gnd, and eda, was used to construct the mutant LMSE$_1$, with three corresponding gene deletions, pgi, gnd, and eda. The experimental characterization of the mutant LMSE$_1$ is provided in Chapter 4. The solutions belonging to strategy A and the metabolic network for the mutant LMSE$_1$ are shown in Figure 3.2.

The deletion of the gene pgi, encoding phosphoglucose isomerase, restricted the entry of glucose into glycolysis and conversion of xylose to G6P using the reductive pentose phosphate pathway (PPP) and reverse activity of phosphoglucose isomerase (Fig. 3.2). The deletion of the gene gnd, encoding 6-phosphogluconate dehydrogenase, prevented glucose from entering the PPP. Finally, the deletion of the gene eda, encoding the second enzyme of the Entner-Doudoroff Pathway, ensured that glucose could not be used to synthesize glyceraldehyde-3-phosphate (G3P), which could otherwise be channeled into glycolysis. By deleting these three genes, glucose was confined to synthesis of G6P, and could not be used in the downstream metabolic reactions. All the remaining biomass precursors were provided by xylose alone, which entered the PPP as xylulose-5-phosphate (X5P). Thus, the mutant was predicted to grow only when both glucose and xylose were provided and not on glucose or xylose individually.

The gene deletion combination of the mutant LMSE$_1$ was also tested on the genome-scale model of E. coli, iAF1260 (Feist et al., 2007). The genome-scale model predicted no growth on glucose, but non-zero growth on xylose. This discrepancy was attributed to the fact that glycogen was not included as a biomass precursor in the model iAF1260, thereby rendering G6P synthesis non-essential for growth. The central metabolism model assumes G6P as a growth precursor due to its role in glycogen synthesis. As the mutant did not require G6P synthesis for growth, it was predicted to grow on xylose alone by the model iAF1260.
Figure 3.2: a) The solutions belonging to strategy A, and b) the metabolic map of the mutant LMSE\(_1\). The highlighted solution (in red) was used to construct the mutant LMSE\(_1\).

3.5.2 Strategy B and the mutant LMSE\(_2\)

The strategy B was a group of five solutions that allowed synthesis of only ribose-5-phosphate (R5P) from glucose, and only xylulose-5-phosphate (X5P) from xylose. The mutant LMSE\(_2\), with three gene deletions, pgi, rpe, and eda, each corresponding to a reaction, was constructed to verify the strategy B (experimental characterization discussed in Chapter 4). The solutions belonging to the strategy B and the metabolic network for the mutant LMSE\(_2\) are shown in Figure 3.3.

The genes pgi and eda were common to the mutant LMSE\(_1\), and served the same purpose of restricting the entry of glucose into glycolysis and the Entner-Doudoroff Pathway (Fig. 3.3). Due to the deletion of rpe, glucose could not be used to synthesize the downstream metabolites of the PPP beyond R5P. Deletion of rpe also prevented xylose from entering the PPP. The only metabolite that could be synthesized from xylose was X5P. In presence of both glucose and xylose, R5P (from glucose) and X5P (from xylose) could be used to synthesize all the
remaining PPP metabolites, which could then be used to synthesize the remaining biomass precursors. The strategy for the mutant LMSE_2 was verified with the model iAF1260. The equivalent strategy with the genome-scale model iAF1260 needed deletion of six reactions (Table 3.1).

### 3.5.3 Strategy C and the mutant LMSE_5

The strategy C consisted of three solutions, two with three reaction deletions and one with four reaction deletions. The solutions in the strategy C could not be explained from the network connectivity similar to the strategies A and B. The mutant LMSE_5 was constructed based on the solution with four reaction deletions, pgI, gnd, fbp, and pfk. Due to the presence of two isozymes, pfkA and pfkB, the mutant LMSE_5 had five gene deletions (pgI, gnd, fbp, pfkA, and pfkB). The experimental characterization of the mutant LMSE_5 is discussed in Chapter 4. The solutions in the strategy C and the related metabolic network for the mutant LMSE_5 are shown in Figure 3.4.
Figure 3.4: a) The solutions belonging to strategy C, and b) the metabolic map of the mutant LMSE$_5$. The highlighted solution (in red) was used to construct the mutant LMSE$_5$.

With the deletion of the four reactions, glucose could not be used to synthesize fructose-6-phosphate (F6P). Furthermore, the gluconeogenic flux and the entry of glucose into the PPP were stopped. Thus, the reaction deletions were lethal on glucose. Xylose, however, could enter through the PPP as X5P, followed by conversion into R5P (using the reactions rpi and rpe) (Fig. 3.4). With X5P and R5P synthesized from xylose, all the remaining metabolites of the PPP could possibly be synthesized and channeled into glycolysis. Thus, considering the connectivity of the metabolic network, it appears that the strain could grow on xylose alone. However, SimUp predicted that the deletion of the four reactions would be lethal on xylose. An equivalent strategy on the genome scale model iAF1260 needed deletion of six reactions (Table 3.1). As the metabolic network was fully connected with respect to xylose, lethality on xylose in the strategy C was most likely due to stoichiometric constraints. To investigate the reason for lethality of the mutant LMSE$_5$ on xylose, stoichiometric feasibility analysis was carried out.
3.5.4 Feasibility analysis of biomass precursors

Lethality caused due to the metabolic gene deletions can be explained by the mutants’ inability to synthesize the complete set of biomass precursors (Suthers et al., 2009). To identify the cause of lethality of the mutant LMSE\textsubscript{5} (on xylose), the feasibility of formation of each of its biomass precursor was analyzed under three different substrate conditions: glucose alone, xylose alone, and glucose and xylose. Similar analysis was also carried out for the mutant LMSE\textsubscript{2}.

Figure 3.5 shows the results for the feasibility analysis for the mutants LMSE\textsubscript{2} and LMSE\textsubscript{5}. As expected, the mutant LMSE\textsubscript{2} could not synthesize any precursor except R5P from glucose and X5P (not a biomass precursor) from xylose (Fig. 3.5a). However, in presence of both glucose and xylose, all the precursors could be synthesized, and growth of the mutant was possible. For the mutant LMSE\textsubscript{5}, growth on glucose was not possible because R5P and erythrose-4-phosphate (E4P) could not be synthesized from glucose (as deletion of gnd prevented glucose from entering the PPP) (Fig. 3.5b). The only metabolites that were feasible from xylose were R5P and F6P. No other biomass precursors, such as G3P or E4P, could be synthesized, despite the existing reactions that connected these metabolites to xylose. To identify whether stoichiometric imbalance was causing the lethality on xylose, secretion reaction for each biomass precursor was iteratively added to the metabolic network of the mutant LMSE\textsubscript{5}.

Figure 3.5c shows the feasibilities of biomass precursors in the mutant LMSE\textsubscript{5} after addition of a secretion reaction for the biomass precursor F6P. Allowing secretion of F6P enabled the mutant LMSE\textsubscript{5} to grow on xylose alone, suggesting that the cause of reduced growth of the mutant LMSE\textsubscript{5} on xylose was intracellular accumulation of F6P. The mutant was most likely forced to synthesize excess F6P to meet the imposed stoichiometric constraints due to reaction deletions. As sugar phosphates cannot be secreted out of the cell, the likely accumulation of F6P resulted in the limited growth phenotype of the mutant LMSE\textsubscript{5} on xylose. Thus, the co-utilization phenotype of the mutant LMSE\textsubscript{5} was attributed to stoichiometric imbalance, in addition to the topological constraints of the metabolic network.
Figure 3.5: Stoichiometric feasibilities for the mutants LMSE₂ and LMSE₅. G – glucose as carbon source; X – xylose as carbon source; GX – glucose and xylose as carbon source. a) Feasibilities of biomass precursors for the mutant LMSE₂, b) feasibilities of biomass precursors for the mutant LMSE₅, c) feasibilities for all the biomass precursors with additional secretion reaction for F6P in the metabolic network of the mutant LMSE₅. F6P – fructose-6-phosphate, G3P – glyceraldehyde-3-phosphate, 3PG – 3-phosphoglycerate, PEP – phosphoenolpyruvate, PYR - pyruvate, R5P – ribose-5-phosphate, E4P – erythrose-4-phosphate, OA – oxaloacetate, AKG – α-ketoglutarate.

3.6 Conclusions

This chapter describes a novel bilevel optimization algorithm, SimUp, that predicts metabolic reaction deletion strategies for co-utilization of glucose and xylose. Eleven different solutions were predicted using the central metabolism model of E. coli. The solutions were classified into three different strategies. Two of the three strategies, strategy A and strategy B, were found to be intuitive as they could be understood by following the network connectivity. However, strategy C could not be justified solely by the network connectivity and required additional analysis on the stoichiometric feasibility of each biomass precursor. Strategy C forced co-utilization of glucose and xylose most likely due to the imbalance in fructose-6-phosphate metabolism.
Chapter 4

Glucose and Xylose Co-utilizing Mutants of *Escherichia coli*

This chapter contains material from our publication (Gawand et al., 2013):


This chapter consists of the experimental section published in the manuscript from the citation above. The names of the (LMSE-series\(^1\)) mutants have been slightly changed from the original publication in accordance with the new naming convention adopted. For example, the mutant denoted as LMSE1 in the original publication is denoted as LMSE\(_1\), et cetera. Reproduction of the material from the publication in this thesis is a right that has been granted by Elsevier (the publisher) to the authors of the manuscript.

4.1 Abstract

One of the current challenges in using lignocellulosic biomass as industrial feedstock is the complex composition of the sugar streams that are obtained from saccharification of the biomass.

---

\(^{1}\)The name of the series is derived from the name of the research group: Laboratory of Metabolic Systems Engineering.
Most industrial microorganisms, including *Escherichia coli*, regulate their sugar utilization using carbon catabolite repression (CCR), causing the preferred sugar (glucose) to be utilized before the non-preferred sugars (xylose or arabinose). This sequential utilization of sugars causes diauxic growth which is not ideal for industrial fermentation processes. Several strains of *E. coli* have been previously constructed to eliminate CCR, thereby allowing co-utilization of glucose and xylose. However, most of the previously described mutants contain mutations in the regulatory genes involved in CCR. This chapter describes construction and characterization of *E. coli* mutants that co-utilize glucose and xylose without genetic alterations to CCR. These mutants were based on the designs predicted by the SimUp algorithm described in Chapter 3. Two of the five mutants constructed, mutant LMSE\textsubscript{2} and mutant LMSE\textsubscript{5}, showed glucose-xylose co-utilization on mixtures containing equal concentrations of glucose and xylose. The mutant LMSE\textsubscript{2} was further characterized on mixtures with unequal concentrations of glucose and xylose.

### 4.2 Introduction

*Escherichia coli* can naturally metabolize both glucose and xylose, the two major sugar components of the lignocellulosic hydrolysate. However, like most common microorganisms, *E. coli* consumes the two sugars sequentially, preferring glucose over xylose. Presence of glucose represses the expression of transporters and metabolic genes required for utilization of xylose due to the phenomenon known as carbon catabolite repression (CCR). Presence of CCR is an undesirable trait in industrial microorganisms as it causes delay in the utilization of the non-preferred sugars resulting in increased residence times in the fermenter. Additionally, complex feeding regimes are required for fed-batch fermentation processes using organisms with CCR, making the fermentation processes difficult to control.

Mutants of industrial microorganisms that can co-utilize glucose and xylose can potentially reduce fermentation batch times and can simplify fermentation processes (Kim et al., 2010). Due to the extensive information available on CCR in *E. coli* and *S. cerevisiae*, engineering of the regulatory process in these microorganisms to obtain co-utilizing mutants has been previously attempted. These engineering strategies involved deletion of important regulatory gene(s), or
replacing an important regulatory gene with its mutant allele. For example, a mutant devoid of PTS (constructed by deleting the operon containing *ptsH*, *ptsI*, and *crr*, encoding HPr, EI, and EIIGlc, respectively) was constructed to force non-PTS mediated glucose uptake. This mutant showed diminished glucose repression on xylose (Hernández-Montalvo et al., 2001). *E. coli* carrying a mutation in *ptsG* (encoding the glucose transporter EIIBC\textsubscript{Glc}) was found to co-utilize glucose and xylose at low sugar concentrations (Nichols et al., 2001). Allosteric alteration of the global regulatory protein crp (designated as crp\textsuperscript{*} or crp\textsuperscript{+}) is known to confer relaxed CCR in *E. coli* (Khankal et al., 2009). Recently, a crp\textsuperscript{*} mutant (with an amino acid substitution from Gly\textsubscript{122} to Ser\textsubscript{122}) was shown to co-utilize glucose and xylose (Yao et al., 2011).

Mutations in various metabolic genes are also known to confer co-utilization phenotype to *E. coli*. Deletion of mgsA, for example, has been implicated in improved co-utilization of glucose and xylose in an ethanologenic strain of *E. coli* (Yomano et al., 2009). A minimal *E. coli* mutant, also engineered to produce ethanol, co-utilized glucose and xylose using an unknown mechanism (Trinh et al., 2008). An alternative strategy to force co-utilization of glucose and xylose involves cultivation of a co-culture of substrate-selective *E. coli* mutants (Eiteman et al., 2009). Though these mutants of *E. coli* show improved co-utilization of glucose and xylose, none of the mutants have shown complete elimination of CCR.

Similar approaches for altering the regulatory genes in *S. cerevisiae* have not been particularly successful. For instance, deletion of the regulatory gene MIG1 did not alleviate glucose repression on xylose utilization (Roca et al., 2004). Evolutionary engineering of *S. cerevisiae* also did not yield a strain capable of co-utilizing glucose and xylose (Wisselink et al., 2009). These examples suggest that the glucose-xylose co-utilization problem has not been entirely solved even in the most intensively studied organisms. Due to the natural tendency of the organisms to use the most efficient substrate first, which is coordinated by multiple interacting regulatory mechanisms, the problem of engineering a co-utilization phenotype is especially challenging. This problem is compounded further by the lack of knowledge of CCR in many organisms of interest. Thus, there remains a need for novel approaches that can lead to co-utilization phenotypes in industrial organisms without extensive manipulations to their regulatory networks.

This chapter describes the construction and characterization of the mutants predicted by the SimUp algorithm described in Chapter 3. These mutants were characterized using phenotypic
tests and batch cultivation studies.

4.3 Materials and Methods

4.3.1 Strains, media, and plasmids

All the strains and plasmids were procured from Coli Genetic Stock Center strain (CGSC), Yale University. *E. coli* K-12 MG1655 (*F*, *LAM*, *rph-1*) was the wild-type control strain. This strain was used to make the multiple knockout mutants listed in Table 4.1. All the strains were maintained on Luria-Bertani (LB) plates (16 g/L agar). The kanamycin resistant or chloramphenicol resistant transformants were selected on LB plates containing 25 µg/mL kanamycin or 20 µg/mL chloramphenicol (stock made in ethanol), respectively. For isolation of the mutant with more than two gene deletions, LB plates were supplemented with 5 g/L glucose and 5 g/L xylose. The only plasmid used in this study was the helper plasmid pCP20 for elimination of *kanR* gene from the mutants. pCP20 is a temperature sensitive plasmid, has ampicillin resistance and chloramphenicol resistance genes, and shows temperature inducible FLP (flippase) synthesis which enables recombination of FRT (flippase recognition target) sites and elimination of any region present between them. The plasmid was maintained in *E. coli* BW25113 and was isolated using Miniprep plasmid extraction kit (Sigma Aldrich).

4.3.2 Phenotypic characterization

Phenotypic characterization of all the constructed mutants involved characterizing the growth of these mutants on both complex and minimal media. Each mutant was tested under seven different media conditions. The complex medium used was LB broth. The minimal medium used was described previously by Causey et al. (2003). The medium contained, per L: 3.5 g of KH$_2$PO$_4$; 5.0 g of K$_2$HPO$_4$; 3.5 g of (NH$_4$)$_2$HPO$_4$; 0.25 g of MgSO$_4$.7H$_2$O; 15 mg of CaCl$_2$.2H$_2$O; 0.5 mg of thiamine, and 1 mL of trace metal stock. The trace metal stock was prepared in 0.1 M HCl and consisted of per L: 1.6 g of FeCl$_3$; 0.2 g of CoCl$_2$.6H$_2$O; 0.1 g of CuCl$_2$; 0.2 g of ZnCl$_2$.4H$_2$O; 0.2 g of NaMoO$_4$; and 0.05 g of H$_3$BO$_3$. 4-Morpholinopropanesulfonic acid (MOPS) (0.1 M) was added to the medium to control pH. The trace metal solution, thiamine solution, and MOPS were filter-sterilized using 0.22 micron filters.
Table 4.1: List of the mutants constructed in this study.

<table>
<thead>
<tr>
<th>Name</th>
<th>Genotype</th>
<th>Strategya</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMSE\textsubscript{1}</td>
<td>\textit{E. coli} $\Delta$pgi, $\Delta$eda, $\Delta$gnd::kanR</td>
<td>A</td>
</tr>
<tr>
<td>LMSE\textsubscript{2}</td>
<td>\textit{E. coli} $\Delta$pgi, $\Delta$eda, $\Delta$rpe::kanR</td>
<td>B</td>
</tr>
<tr>
<td>LMSE\textsubscript{3}</td>
<td>\textit{E. coli} $\Delta$pgi, $\Delta$gnd, $\Delta$fbaB::kanR</td>
<td>C</td>
</tr>
<tr>
<td>LMSE\textsubscript{4}</td>
<td>\textit{E. coli} $\Delta$pgi, $\Delta$gnd, $\Delta$tpi::kanR</td>
<td>C</td>
</tr>
<tr>
<td>LMSE\textsubscript{5}</td>
<td>\textit{E. coli} $\Delta$pgi, $\Delta$gnd, $\Delta$pfkB, $\Delta$pfkA, $\Delta$fkp::kanR</td>
<td>C</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Refer to \textbf{Chapter 3} for definition and description of strategies.

The seven conditions used for phenotypic characterization were: LB, LB with 5 g/L glucose, LB with 5 g/L xylose, LB with 5 g/L each of glucose and xylose, MM with 10 g/L glucose, MM with 10 g/L xylose, and MM with 5 g/L each of glucose and xylose. The growth of the mutants was observed for 48 hours, and was qualitatively described as positive growth, negative growth, or limited growth based on the optical density measurements at 550 nm (OD\textsubscript{550}).

### 4.3.3 Batch cultivation

Three strains, LMSE\textsubscript{1}, LMSE\textsubscript{2}, and LMSE\textsubscript{5}, were cultivated in 5 L fermenters (Minifors, INFORS HT) with 3 L working volume to characterize their growth rates and sugar consumption profiles. The medium used for growth characterization was minimal medium with glucose and xylose, 5 g/L each, as carbon source. The pH was maintained at 7.0 using 6 M KOH, and dissolved oxygen (DO) levels were maintained above 50% by varying air-flow rate and impeller speed.

Seed cultures were prepared by inoculating a fresh (less than 48 h old) colony overnight in 3 mL minimal medium. The overnight grown culture was then transferred to 500 mL Erlenmeyer flasks with 150 mL minimal medium containing glucose and xylose (5 g/L each). The flasks were incubated till OD\textsubscript{550} of 1.5-2.2 was achieved (around 8 – 12 hours) . The fermenters were inoculated with appropriate amount of harvested biomass to get an initial OD\textsubscript{550} of 0.1.

Samples were drawn regularly from the fermenters until the end of the batch, which was indicated by the rise in pH and DO levels. The samples were analyzed for biomass (OD\textsubscript{550}; OD\textsubscript{550} of 0.1 corresponded to 35 mg of dry cell weight/L (Causey et al., 2003)), and sugar
concentrations. The study was carried out in triplicates for all the mutants under consideration.

In addition to equal concentrations of glucose and xylose, the mutant LMSE2 was further characterized on different concentrations of glucose and xylose. Two different sugar combinations were chosen: a) glucose concentration (6.7 g/L) twice that of xylose (3.3 g/L), and b) xylose concentration (6.7 g/L) twice that of glucose (3.3 g/L).

4.3.4 Metabolite analysis

Glucose and xylose were separated using a Bio-Rad HPX-87H cation-exchange column. 5 mM H$_2$SO$_4$ was used as the mobile phase (0.5 mL/min flow rate, 42°C column temperature, and 20 µL injection volume).

4.4 Results

4.4.1 Phenotypic characterization of the mutants

All the five mutants were characterized for growth on complex and minimal medium. Figure 4.1 shows the phenotypic characterization results of all the mutants.

During construction, the mutants were isolated on LB supplemented with glucose and xylose, as some of the mutants were expected not to grow on LB alone. The phenotypic studies showed that all the mutants were capable of growing on LB except the mutant LMSE2, which could not grow on LB or even LB supplemented with glucose. Growth of the mutant LMSE2 was observed only when LB was supplemented with xylose. The results indicate that construction and selection of higher-order mutants may require supplementation of complex medium, which is generally regarded as sufficiently rich to support growth of most mutants.

The phenotypic studies on minimal medium were carried out to verify the predictions of the SimUp algorithm. SimUp was formulated to predict mutants with a specific growth phenotype: mutants that could not grow on glucose or xylose, but could grow on a mixture of glucose and xylose. SimUp algorithm was not used to predict phenotypes on the complex medium due to its undefined composition. Only two mutants, the mutant LMSE2 and the mutant LMSE5, conformed to this phenotype when experimentally tested. The other three mutants, LMSE1, LMSE3, and LMSE4, showed growth on either glucose-based or xylose-based medium,
not conforming to the SimUp predicted phenotypes. The mutants LMSE$_1$ and LMSE$_3$ showed no growth on glucose, but could grow on xylose alone. The growth on xylose was possible most likely due to non-deleted alternative pathways that were not included in the central metabolism model. No equivalent strategy could be identified for these mutants using a genome-scale model. As the mutants grew on xylose alone, they were not expected to show co-utilization of glucose and xylose. The mutant LMSE$_1$ was further characterized in a batch culture with mixture of glucose and xylose to check its sugar utilization characteristics (results discussed in the next section). The mutant LMSE$_4$ showed growth on glucose as well as xylose. As the mutant LMSE$_4$ could utilize either glucose or xylose individually, it was anticipated that the mutant would not co-utilize the two sugars. This assumption was based on the fact that the mutant had no manipulations to the genes involved in CCR. If CCR in the mutant was active, the mutant would utilize glucose and xylose sequentially. Based on this assumption the mutant LMSE$_4$ was not further characterized. The mutant LMSE$_5$ showed limited growth on xylose-based medium (Fig. 4.1b), which was classified as non-lethal sick phenotype. As discussed in Chapter 3, the sick phenotype of the mutant LMSE$_5$ was most likely due to the imbalance in the metabolism resulting in accumulation of fructose-6-phosphate.

Based on the observed phenotypes, we characterized one mutant from each strategy to verify whether the mutants could co-utilize glucose and xylose. The mutants used for further studies were, LMSE$_1$, LMSE$_2$, and LMSE$_5$.

### 4.4.2 Growth characteristics of the LMSE mutants

The batch characteristics of the characterized mutants are shown in Figure 4.2. The experimentally calculated growth parameters are summarized in Table 4.2.

The wild-type *E. coli* was characterized for growth on glucose and xylose as a control. The wild-type showed the expected diauxic growth pattern by consuming glucose before xylose, and showing the characteristic lag phase. No consumption of xylose was observed in presence of glucose. The lag phase lasted for around 30 minutes (between 8 and 9 hours), which corresponded exactly to the exhaustion of glucose in the medium (Figure 4.2a). During the lag-phase the wild-type expressed the genes required for utilization and metabolism of xylose, and resumed the growth on xylose in the medium. The growth rate, glucose uptake rate, and xylose uptake
Figure 4.1: Phenotypic characterization of the LMSE mutants. a) Shows the growth characteristics on seven different media combinations including LB and minimal medium, b) shows the limited growth phenotype of the mutant LMSE<sub>5</sub>.

The mutant LMSE<sub>1</sub>

The sugar consumption and growth profiles of the mutant LMSE<sub>1</sub> are shown in Figure 4.2b. The mutant LMSE<sub>1</sub> had three genes deleted (Table 4.1), which confined glucose to synthesis of glucose-6-phosphate (G6P), and the remaining metabolism was supported by xylose. As the gene deletions did not allow synthesis of G6P from xylose, the mutant was expected to consume glucose and xylose simultaneously, provided G6P was a biomass precursor. Phenotypic
characterization of the mutant showed that it could grow on minimal medium with xylose as the sole carbon source, suggesting G6P was not required for growth (this phenotype was later predicted by the genome-scale metabolic model of *E. coli*, iAF1260). As the mutant could grow on xylose alone, the mutant did not co-utilize glucose and xylose, but consumed xylose alone. Glucose in the medium was not utilized as G6P could not be converted into any other metabolite.

In the model iAF1260, G6P is not used for synthesis of any metabolite except glycogen, and since glycogen is not a biomass precursor, it was likely that the mutant LMSE\textsubscript{1} did not synthesize G6P from xylose. Based on these results, the central metabolism model was amended by removing G6P as a biomass precursor. The updated central metabolism model was used for all further simulations and to predict the mutants LMSE\textsubscript{2} and LMSE\textsubscript{5}.

**The mutant LMSE\textsubscript{2}**

For the mutant LMSE\textsubscript{2}, deletion of only three genes was required to confer the expected phenotype. Additional deletions based on iAF1260 (Chapter 3, Table 3.1) were not needed, indicating that the additional reactions could not support growth, most likely due to the limited flux carrying capacity of the enzymes involved. The characterization results of the mutant LMSE\textsubscript{2} are shown in Figure 4.2. When cultivated in a batch, the mutant LMSE\textsubscript{2} showed simultaneous utilization of glucose and xylose, thereby confirming the SimUp prediction. The mutant co-utilized glucose and xylose in a strict 1:1 ratio. Additionally, the growth rate of the mutant was found to be lower than the wild-type growth rate (Table 4.2). The lower growth rate of the mutant was most likely due to the sub-optimal growth phase in which the mutant was characterized. The uptake rate of both glucose and xylose were lower compared to the wild-type.

The mutant LMSE\textsubscript{2} was also characterized on different ratios of glucose and xylose (Fig. 4.3). The mutant showed different growth rates and sugar uptake rates under different media conditions. Additionally, the mutant could not consume the residual sugar after one of the sugar was exhausted in the medium. The biomass yields for the mutant LMSE\textsubscript{2} grown on different concentrations of glucose and xylose were found to be higher than growth on medium with equal concentrations of glucose and xylose. The higher biomass yield was most likely a
Figure 4.2: Batch cultivation studies of the LMSE mutants on glucose and xylose. a) Growth characteristics of the wild-type \textit{E. coli}, b) growth characteristics of the mutant LMSE\textsubscript{1}, c) growth characteristics of the mutant LMSE\textsubscript{2}, d) growth characteristics of the mutant LMSE\textsubscript{5}. The symbol, \( \bullet \) represents biomass expressed as dry cell weight, \( \circ \) represents glucose concentration, and \( \blacktriangle \) represents xylose concentration. The error bars represent standard deviation of three replicates.

The xylose uptake rate in the mutant LMSE\textsubscript{2} was consistently higher than glucose uptake rate. Under conditions of low xylose concentration (3.3 g/L) and high glucose concentration (6.7 g/L), glucose uptake rate was found to drop, suggesting that xylose metabolism was required consequence of absence of overflow metabolism due to slower growth rate.
for glucose uptake. With higher concentration of xylose (6.7 g/L), and lower concentration of glucose (3.3 g/L), the growth rate improved, but was still lower than the conditions with equal concentrations of glucose and xylose. It appears that despite the fact that the mutant could consume sugars at unequal rates, the highest growth rate of the mutant was achieved when the two sugars were consumed at almost same rate.

The higher dependence of the mutant on xylose can be explained based on the metabolic network of the mutant. Xylose was consumed by the mutant using the pentose phosphate pathway (PPP). Xylose enters the PPP as xylulose-5-phosphate (X5P), which can be catabolized using two reactions of the PPP. In one reaction, catalyzed by transaldolase, X5P reacts with ribose-5-phosphate (R5P) to form sedoheptulose-7-phosphate (S7P) and glyceraldehyde-3-phosphate (G3P); in other the reaction, catalyzed by transketolase, X5P reacts with erythrose-4-phosphate (E4P) to be converted into fructose-6-phosphate (F6P) and G3P. As F6P and G3P also participate in glycolysis and biomass formation, the mutant LMSE\(_2\) most likely depended more on xylose by catabolizing it using both the reactions.
Table 4.2: Growth characteristics of the constructed mutants.

<table>
<thead>
<tr>
<th>Strains</th>
<th>glu (g/L)</th>
<th>xyl (g/L)</th>
<th>( \mu^a )</th>
<th>( q_{glc}^b )</th>
<th>( q_{xyl}^b )</th>
<th>( Y_{x/s}^c )</th>
<th>( \mu^{FBAa} )</th>
<th>( Y^{FBAc}_{x/s} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>wild-type</td>
<td>5.0</td>
<td>5.0</td>
<td>0.60</td>
<td>9.3</td>
<td>–</td>
<td>0.011</td>
<td>0.86</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.28(^d)</td>
<td>–</td>
<td>6.53</td>
<td>0.008</td>
<td>0.49(^d)</td>
<td>0.015(^d)</td>
</tr>
<tr>
<td>LMSE(^1)</td>
<td>5.0</td>
<td>5.0</td>
<td>0.28</td>
<td>0.00</td>
<td>6.86</td>
<td>0.008</td>
<td>0.49</td>
<td>0.015</td>
</tr>
<tr>
<td>LMSE(^2)</td>
<td>5.0</td>
<td>5.0</td>
<td>0.38</td>
<td>3.98</td>
<td>4.91</td>
<td>0.008</td>
<td>0.67</td>
<td>0.014</td>
</tr>
<tr>
<td>LMSE(^2)</td>
<td>3.3</td>
<td>6.7</td>
<td>0.34</td>
<td>2.76</td>
<td>3.78</td>
<td>0.010</td>
<td>0.51</td>
<td>0.014</td>
</tr>
<tr>
<td>LMSE(^2)</td>
<td>6.7</td>
<td>3.3</td>
<td>0.26</td>
<td>2.17</td>
<td>2.48</td>
<td>0.010</td>
<td>0.32</td>
<td>0.012</td>
</tr>
<tr>
<td>LMSE(^5)</td>
<td>5.0</td>
<td>5.0</td>
<td>0.20</td>
<td>3.86</td>
<td>1.12</td>
<td>0.007</td>
<td>0.36</td>
<td>0.012</td>
</tr>
</tbody>
</table>

\(^a\)expressed in h\(^-1\). \(^b\)expressed in mmol/gDW/h, \(^c\)expressed in gDW/mmol, \(^d\)calculated for xylose.

All the values were calculated based on the results presented in Figure 4.2 and Figure 4.3.

The mutant LMSE\(^5\)

In phenotypic characterization, the mutant LMSE\(^5\) showed no growth on glucose, limited growth on xylose, and significant growth only when both glucose and xylose were provided (Fig. 4.1). When cultivated in a mixture of glucose and xylose, the mutant co-utilized glucose and xylose in agreement with SimUp predictions. The growth rate of the mutant was severely reduced as compared to the wild-type, most likely due to the deletion of five major genes from the central metabolic pathways. Unlike the mutant LMSE\(^2\), the glucose uptake rate of the in the mutant LMSE\(^5\) was much higher as compared to the uptake rate of xylose. The dominant pathway for glucose utilization was the Entner-Doudoroff pathway in the mutant, and growth seemed to be glucose dependent. The mutant also did not show strict co-utilization of glucose and xylose, consuming glucose faster than xylose, thereby showing a small lag between glucose and xylose consumption. The most likely reason for lower rate of xylose consumption, as predicted by the stoichiometric analysis, was accumulation of excess F6P. F6P could not be metabolized in the mutant LMSE\(^5\) due to deletion of the genes \( pfkA \) and \( pfkB \). Thus, xylose consumption was most likely limited by rate of catabolism of F6P, which was very low in the mutant LMSE\(^5\).

Despite the growth limitations of the mutant LMSE\(^5\), this mutant was the first mutant
which was engineered based on stoichiometric imbalance, and could not be predicted based on metabolic network connectivity. The mutant LMSE\textsubscript{5} was an example of the superior predictive capabilities of stoichiometric models and constraint-based modeling techniques.

4.5 Conclusions

This chapter describes the construction and characterization of the LMSE-series mutants which were based on the predictions of the SimUp algorithm. Of the five multiple gene knockout mutants constructed, two mutants showed co-utilization of glucose and xylose. Whereas the mutant LMSE\textsubscript{2} showed strict co-utilization of glucose and xylose, the mutant LMSE\textsubscript{5} utilized the two sugars simultaneously but at different rates. The growth rate and the sugar utilization rate in both the mutants LMSE\textsubscript{2} and LMSE\textsubscript{5} was lower than the wild-type, indicating that the co-utilization phenotype in the mutants came at the cost of prolonged batch times. The mutant LMSE\textsubscript{5} validated the first metabolic engineering strategy based on stoichiometric imbalance. The mutants that did not co-utilize sugars (LMSE\textsubscript{1}, LMSE\textsubscript{3}, and LMSE\textsubscript{4}) highlight the redundancies in the metabolic network. Finally, the mutants LMSE\textsubscript{2} and LMSE\textsubscript{5} validated the original hypothesis of this thesis that co-utilization of sugars can be forced without making any direct manipulations to the regulatory network.
Chapter 5

Genome Sequencing of the

*Escherichia coli* Mutant LMSE$_2$

This chapter contains material from our publication (Gawand et al., 2013):


The studies on Sanger sequencing and characterization of the mutant gene *ptsG* have been reported in the above publication. Additional studies on the whole genome sequencing reported in this chapter are unpublished. Re-production of the material in this thesis is a right that has been granted by Elsevier (the publisher) to the authors of the manuscript.

5.1 Abstract

The *E. coli* mutant LMSE$_2$ co-utilizes glucose and xylose due to the deletion of the three metabolic genes, *pgi, rpe,* and *eda*. The regulatory network responsible for carbon catabolite repression (CCR) has not been externally altered in the mutant. Sequencing studies were carried out to investigate the exact mechanism for sugar co-utilization in this mutant. Sanger sequencing was first used to sequence all the genes involved in CCR, followed by whole genome sequencing using next-generation Illumina Sequencing. Two genes with significant mutations
were identified: *ptsG* (encoding the major glucose transporter EIIBC\textsubscript{Glc}), and *icdA* (encoding the TCA cycle enzyme isocitrate dehydrogenase). The mutated genes were further characterized by complementation studies. The mutations identified in the genes were not found responsible for the glucose-xylose co-utilization phenotype of the mutant. The results suggest that the mutant LMSE\textsubscript{2} most likely co-utilized the two sugars by altering its enzyme expression levels, without accumulating any significant mutations in the CCR network. These preliminary results indicate that the cells are able to transition between distinct metabolic states without altering their regulatory network.

### 5.2 Introduction

Wild-type E. coli consumes glucose and xylose sequentially, resulting in a diauxic growth pattern (Görke and Stülke, 2008a). The E. coli mutants capable of co-utilizing glucose and xylose are known to carry mutations in one or more genes involved in CCR (Hernández-Montalvo et al., 2001; Khankal et al., 2009; Nichols et al., 2001; Yao et al., 2011). The mutant LMSE\textsubscript{2}, described in Chapter 4, co-utilizes glucose and xylose due to the forced metabolic dependence on both these sugars. The mutant has three metabolic gene deletions: *pgi*, *rpe*, and *eda*. No alterations were made to the regulatory network involved in CCR of the mutant. From the growth characteristics and sugar utilization characteristics of the mutant LMSE\textsubscript{2}, the exact mechanism responsible for co-utilization of sugars could not be interpreted. One hypothesis for the observed co-utilization phenotype was the possible accumulation of point mutations in the genes involved in CCR. Additionally, random mutations that could have been accumulated by the mutant before the deletion of the three genes could also have had a role to play in adaptation to sugar co-utilization. We analyzed the genotype of the mutant to identify any such mutations.

Whole genome sequencing studies have been previously used to characterize genetic responses to metabolic alterations in microorganisms (Oud et al., 2012; Warner et al., 2009). Mutational profiling is commonly used for reverse engineering of microorganisms to achieve complex traits such as improved stress tolerance (Atsumi et al., 2010). Herring et al. (2006) used comparative genome sequencing, a microarray-based approach, to detect single nucleotide
polymorphisms (SNPs) responsible for improved growth of *E. coli* on glycerol. New sequencing-based technologies, which do not require prior knowledge of the genome sequence, are now commercially available and can be used for SNP profiling of mutants. Due to the large number of low-cost reads produced by the new sequencing-based techniques, these methods find diverse applications such as discovery of variants, *de novo* assemblies of bacterial and eukaryotic genomes, and obtaining the transcriptome of cells (RNA-Seq) (Metzker, 2010). The most popular sequencing techniques currently used are Roche/454, Illumina/Solexa, and SOLiD by Life technologies, which differ from each other in various aspects such as template preparation, sequencing chemistry, and read lengths.

As a preliminary investigation, we first sequenced only the genes involved in CCR of the mutant LMSE₂ using Sanger sequencing. Mutations in the upstream regions of the genes can affect their expression levels by altering the promoter sequences. Accordingly, Sanger sequencing primers for the genes were designed to cover upstream as well as downstream regions of the genes. Table 5.1 shows the list of all the genes sequenced.

Subsequently, to investigate the possible role of unrelated genes in the co-utilization phenotype of the mutant LMSE₂, we used next generation Illumina Sequencing to sequence the entire genome of the mutant.

The mutated targets identified by Sanger sequencing and whole genome sequencing were further investigated using complementation studies. Briefly, the effect of the mutated genes was tested by first deleting the target gene from the wild-type *E. coli*, and expressing the mutated gene, under its own promoter, on a suitable vector. Any changes in the physiology due to the mutated gene were then checked by characterizing the constructed mutants on minimal medium with glucose and xylose. Differences in the sugar utilization patterns between the two complementary strains were compared. Following sections discuss the results of the sequencing studies on the mutant LMSE₂.
Table 5.1: Genes of the mutant LMSE₂ sequenced using Sanger sequencing.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Region of the gene sequenced</th>
<th>Protein</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ptsG</td>
<td>-300 bps +100 bps</td>
<td>ECIIBC</td>
<td>main glucose transporter</td>
</tr>
<tr>
<td>ptsH</td>
<td>-300 bps +100 bps</td>
<td>HPr</td>
<td>histidine protein</td>
</tr>
<tr>
<td>ptsI</td>
<td>-100 bps +100 bps</td>
<td>EI</td>
<td>enzyme I</td>
</tr>
<tr>
<td>crr</td>
<td>-100 bps +100 bps</td>
<td>EIIA</td>
<td>PTS enzyme IIA domain</td>
</tr>
<tr>
<td>crp</td>
<td>-200 bps +100 bps</td>
<td>CRP</td>
<td>cAMP receptor protein</td>
</tr>
<tr>
<td>mgsA</td>
<td>-200 bps +100 bps</td>
<td>MGS</td>
<td>methylglyoxal synthase</td>
</tr>
<tr>
<td>xylF</td>
<td>-100 bps +100 bps</td>
<td>XylF</td>
<td>periplasmic binding domain (transporter)</td>
</tr>
<tr>
<td>xylG</td>
<td>-100 bps +100 bps</td>
<td>XyIG</td>
<td>ATP binding component (transporter)</td>
</tr>
<tr>
<td>xylH</td>
<td>-100 bps +100 bps</td>
<td>XylH</td>
<td>integral membrane component (transporter)</td>
</tr>
<tr>
<td>xylR</td>
<td>-100 bps +100 bps</td>
<td>XylR</td>
<td>xylose regulator (transcription factor)</td>
</tr>
<tr>
<td>xylE</td>
<td>-250 bps +100 bps</td>
<td>XylE</td>
<td>xylose proton symporter</td>
</tr>
</tbody>
</table>

5.3 Materials and Methods

5.3.1 Sanger sequencing

The genes involved in CCR of *E. coli* were sequenced using Sanger sequencing. Upstream regions of all the genes were included in the sequencing to identify mutations in the promoter regions. The list of the genes sequenced, their description, and the regions sequenced are shown in Table 5.1. The list of primers used for sequencing is provided in Appendix C. The target genes were first PCR amplified from the wild-type *E. coli* and the mutant LMSE₂. The PCR products were purified using the GeneJET PCR purification kit, Fermentas and were sent for sequencing at the The Sanger Sequencing Facility at The Center for Applied Genomics (TCAG), Hospital for Sick Children, Toronto. Samples were prepared as per the guidelines provided by TCAG. Alignment of the sequences was carried out using Geneious 5.5.6 (Biomatters Ltd.).

5.3.2 Whole genome sequencing

Whole genome sequencing of the wild-type *E. coli* and two isolates of the mutant LMSE₂ (LMSE₂ (sequenced) and LMSE₂ (unsequenced)) were carried out at the McGill University
and Génome Québec Innovation Center. The mutant LMSE₂ (sequenced) refers to the isolate in which the CCR genes were sequenced using Sanger sequencing. The mutant LMSE₂ (un-sequenced) was an unsequenced variant, isolated as a separate colony during construction of the mutant. Both the mutants were PCR confirmed and showed co-utilization of glucose and xylose. The genomic DNA (gDNA) for the samples was prepared from the overnight grown cultures using GeneJET genomic DNA purification kit, Fermentas. The gDNA was sequenced using Illumina HiSeq 2000 sequencer in 100 bp, paired-end format. The assembly, annotation, and analysis of the SNPs was carried out by the Bioinformatics Facility of the McGill University and Génome Québec Innovation Center. A description of the complete genome pipeline used for the bioinformatics analysis is provided in Appendix F.

### 5.3.3 Complementation studies

The mutated genes identified using Sanger sequencing and the whole genome sequencing were tested using complementation studies. Objective of the complementation studies was to investigate whether the identified mutations were responsible for the observed co-utilization phenotype in the mutant LMSE₂. The mutated \textit{ptsG} (denoted as \textit{ptsG}*) was amplified from the mutant LMSE₂ and ligated into the vector pBT4 (Lynch and Gill, 2006) under its native promoter. Similarly, the mutated \textit{icdA} gene (denoted as \textit{icdA}*) was amplified and ligated into the vector pBT4 under its native promoter. The constructed vectors with \textit{ptsG}* and \textit{icdA}* were transformed into \textit{E. coli} ∆\textit{ptsG} and \textit{E. coli} ∆\textit{icdA}, respectively. The mutants thus obtained were denoted as SQ-002 (\textit{E. coli} ∆\textit{ptsG} pBT4.ptsG*) and SQ-003 (\textit{E. coli} ∆\textit{icdA} pBT4.icdA*). Both these mutants were characterized on the minimal medium (Appendix E) with 5 g/L each of glucose and xylose as carbon source. The mutants were characterized for growth and sugar utilization in 250 mL Erlenmeyer flasks with 50 mL culture volume.

The inoculum was prepared by growing the cultures overnight in LB broth with tetracycline antibiotic (10 µg/mL) for selection pressure. The overnight grown culture was washed with minimal medium and the harvested cells were used for inoculation of the fresh minimal medium to get a starting OD<sub>550</sub> of 0.1. The growth of the cultures and sugar consumption was monitored using OD<sub>550</sub> measurements and HPLC.
5.4 Results

5.4.1 Sanger sequencing results

To identify whether the molecular mechanism of CCR-elimination in the mutant LMSE$^2$ was similar to the previously reported mutants, all the genes involved in CCR of the mutant were sequenced. Among the genes sequenced, only one gene, *ptsG*, was found to carry a mutation. The gene *ptsG* encodes for the main glucose transporter EIIBC$^{\text{Glc}}$. The SNP identified was a base-pair substitution from T to G at position 89 within the gene, which resulted in the amino acid substitution from Lys$^{30}$ to Arg$^{30}$.

The protein EIIBC$^{\text{Glc}}$ consists of two domains: a hydrophobic transmembrane EIIC domain and a hydrophilic EIIB domain, which are connected by a conserved linker (Jeckelmann et al., 2011). The identified mutation was present in the transmembrane EIIC domain of the protein (Fig. 5.1a). Mutations in this region have been associated with non-specific utilization of ribose, mannitol, and fructose (Siebold et al., 2001). Mutations in *ptsG* are also known to induce co-utilization of glucose and xylose (Nichols et al., 2001).

To characterize the identified mutation in *ptsG*, the mutated gene *ptsG$^*$ was expressed in the wild-type *E. coli* devoid of *ptsG*. The strain so obtained was designated as SQ-002. Results from the characterization of the two mutants (*E. coli ∆ptsG* and SQ-002) are shown in Figure 5.1b and 5.1c. Deletion of *ptsG* has been known to drastically reduce glucose uptake in *E. coli*, forcing it to utilize glucose via alternative transporters such as GalP (Flores et al., 2005). In the presence of a mixture of glucose and xylose, the *E. coli ∆ptsG* mutant utilized xylose in preference to glucose, indicating the absence of glucose uptake due to deletion of *ptsG*. Utilization of xylose despite the presence of glucose in the medium suggested absence of CCR. Introduction of the mutated *ptsG$^*$ gene in *E. coli ∆ptsG* restored the order of sugar utilization to the wild-type. The mutant SQ-002 consumed glucose before xylose. However, a strict repression of xylose utilization, as found in the wild-type, was not observed in the mutant SQ-002. This difference could be due to the fact that the EIIBC$^{\text{Glc}}$ was expressed on a plasmid, and the expression levels may not have matched the wild-type levels. As the expression of *ptsG$^*$ improved glucose uptake in the mutant *E. coli ∆ptsG*, and also did not allow utilization of xylose simultaneously with glucose, it could be concluded that the mutation identified in
ptsG* had not severely affected the wild-type gene. These results indicate that the mutation in ptsG* was not responsible for the complete elimination of CCR observed in the mutant LMSE2.

Figure 5.1: a) Schematic showing the position of the SNP in the gene ptsG of the mutant LMSE2, b) growth characteristics of the mutant E. coli ∆ptsG, c) growth characteristics of the mutant SQ-002. The symbol, ● represents biomass, ○ represents glucose concentration, and ▲ represents xylose concentration.
5.4.2 Whole genome sequencing

Whole genome sequencing was carried out to discover the non-CCR genes, if any, that were involved in the regulation of sugar utilization and had a role to play in glucose-xylose co-utilization in the mutant LMSE₂. Three strains were selected for the sequencing study: E. coli wild-type, LMSE₂ (partially sequenced using Sanger sequencing), and an unsequenced variant of LMSE₂. Table 5.2 summarizes the results from the whole genome sequencing studies. In total, 10 genes were found to carry SNPs with varying effects. These variants were identified using non-stringent filtering criteria with minimum read depth set to 2 and the minimum RMS (root mean square) mapping quality for SNPs set to 10 (see Appendix F).

The only gene with a mutation that had a role to play in CCR was ptsG, which was also identified using Sanger sequencing. Four genes, namely, purE, gatC, glpR, and nrfF, showed mutations in all the three samples sequenced (Table 5.2), indicating that these mutations were present in the original wild-type strain used for the study. The genotype of the wild-type used for the study was not exactly same as the E. coli K-12 MG1655 reference genome. As these genes were mutated in the wild-type, which clearly showed active CCR, these genes were not considered responsible for co-utilization of sugars in the mutant LMSE₂.

The gene rhsA showed mutation only in one of the LMSE₂ strains indicating that this gene could not have been responsible for the co-utilization phenotype of the mutant. Mutation of the gene rhsA was classified as random mutation. The genes pgI and rpe were found to have mutations in the mutant LMSE₂ (5.2). As both these genes were deleted from the mutant LMSE₂, no coverage of these genes was anticipated from the genome sequencing. The deletion of both pgI and rpe was confirmed using PCR and phenotypic test prior to the genomic DNA extraction for sequencing. It is clear from Table 5.2 that the coverage for the mutations in genes pgI and rpe was very low (between 18 - 44) compared to the mean coverage of the genome which was around 2000. Furthermore, the genes pgI and rpe were non-continuously and/or incompletely covered, unlike the wild-type which showed complete coverage of the genes. A comparison of the coverage of the regions corresponding to the genes pgI and rpe in the mutants and the wild-type is shown in Appendix G. Based on these observations, the reason for the slight coverage observed for the genes pgI and rpe in the mutant LMSE₂ was attributed
to contamination.

The genes *edd* and *icdA* were the two genes that showed mutations in both the LMSE$_2$ mutants. These mutations were most likely caused in response to the gene deletions in the mutant. The gene *edd* encodes the first enzyme of the Entner-Doudoroff pathway. As the gene *eda*, the second gene of the Entner-Doudoroff pathway, was already deleted in the mutant LMSE$_2$, the mutation in *edd* was not expected to cause any important phenotypic changes.

The only metabolite that is produced by the gene product of *edd* is 2-dehydro-3-deoxy-D-gluconate-6-phosphate, which can be metabolized only through the Entner-Doudoroff pathway. It was hence unlikely that the mutations in *edd* caused glucose-xylose co-utilization in the mutant LMSE$_2$.

The only gene with significant mutations that could have played a role in co-utilization phenotype of the mutant LMSE$_2$ was *icdA*. However, all the mutations in the gene were found to be silent mutations. Though these mutations did not cause any structural changes to the protein encoded by the gene *icdA*, they could have altered the protein concentration levels of the enzyme isocitrate dehydrogenase. Changes in the gene sequence can affect the translational efficiency by altering the codon bias. A direct correlation between translational efficiency based on the codon usage and the expression levels of proteins has been previously proposed, and is based on the hypothesis that highly expressed genes need to be more efficiently translated to gain competitive advantage for the limited resources in the cell (ribosomes and amino acids) (Sharp and Li, 1987). To characterize the mutant gene *icdA* $^\star$, and to test whether the changes in the codon bias of the mutated gene had any role to play in co-utilization of glucose and xylose, complementation studies were carried out.
Table 5.2: The SNPs identified using whole genome sequencing.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Position</th>
<th>Reference allele</th>
<th>Altered allele</th>
<th>Coverage depth</th>
<th>Effect</th>
<th>Codon change</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>seq)</td>
<td>unseq)</td>
<td>wild-type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>purE</td>
<td>547831</td>
<td>AGGGG</td>
<td>AGGGGG</td>
<td>2175</td>
<td>3028</td>
<td>1795</td>
<td>downstream</td>
</tr>
<tr>
<td>ptsG</td>
<td>1157180</td>
<td>T</td>
<td>G</td>
<td>1949</td>
<td>2586</td>
<td>–</td>
<td>missense cTg/cGg</td>
</tr>
<tr>
<td>icdA</td>
<td>1195443</td>
<td>C</td>
<td>T</td>
<td>872</td>
<td>1181</td>
<td>–</td>
<td>silent caC/caT</td>
</tr>
<tr>
<td>icdA</td>
<td>1195455</td>
<td>C</td>
<td>T</td>
<td>753</td>
<td>1035</td>
<td>–</td>
<td>silent acC/acT</td>
</tr>
<tr>
<td>icdA</td>
<td>1195468</td>
<td>T</td>
<td>C</td>
<td>606</td>
<td>835</td>
<td>–</td>
<td>silent Tta/Cta</td>
</tr>
<tr>
<td>icdA</td>
<td>1195470</td>
<td>A</td>
<td>G</td>
<td>607</td>
<td>833</td>
<td>–</td>
<td>silent ttA/ ttG</td>
</tr>
<tr>
<td>icdA</td>
<td>1195500</td>
<td>C</td>
<td>T</td>
<td>257</td>
<td>371</td>
<td>–</td>
<td>silent aaC/aaT</td>
</tr>
<tr>
<td>icdA</td>
<td>1195503</td>
<td>G</td>
<td>C</td>
<td>221</td>
<td>324</td>
<td>–</td>
<td>silent geC/geC</td>
</tr>
<tr>
<td>icdA</td>
<td>1195506</td>
<td>A</td>
<td>G</td>
<td>210</td>
<td>306</td>
<td>–</td>
<td>silent aaA/aaG</td>
</tr>
<tr>
<td>edd</td>
<td>1931545</td>
<td>T</td>
<td>A</td>
<td>2048</td>
<td>2676</td>
<td>–</td>
<td>nonsense Aaa/Taa</td>
</tr>
<tr>
<td>gatC</td>
<td>2171384</td>
<td>ACC</td>
<td>ACCCC</td>
<td>1748</td>
<td>2337</td>
<td>1630</td>
<td>frameshift -/GG</td>
</tr>
<tr>
<td>rpe</td>
<td>3512423</td>
<td>GCC</td>
<td>GC</td>
<td>33</td>
<td>–</td>
<td>–</td>
<td>frameshift -</td>
</tr>
<tr>
<td>rpe</td>
<td>3512462</td>
<td>T</td>
<td>G</td>
<td>18</td>
<td>44</td>
<td>–</td>
<td>missense gAc/gCc</td>
</tr>
<tr>
<td>gfpR</td>
<td>3558477</td>
<td>CG</td>
<td>C</td>
<td>2361</td>
<td>3309</td>
<td>1748</td>
<td>frameshift -</td>
</tr>
<tr>
<td>rhsA</td>
<td>3764327</td>
<td>TAAAAAAA</td>
<td>TAAAAAA</td>
<td>3474</td>
<td>–</td>
<td>–</td>
<td>frameshift -</td>
</tr>
<tr>
<td>pgi</td>
<td>4233409</td>
<td>C</td>
<td>CA</td>
<td>–</td>
<td>29</td>
<td>–</td>
<td>frameshift -/A</td>
</tr>
<tr>
<td>nrfF</td>
<td>4294403</td>
<td>AC</td>
<td>ACGC</td>
<td>1654</td>
<td>2333</td>
<td>1184</td>
<td>downstream</td>
</tr>
</tbody>
</table>
5.4.3 Characterization of \( \text{icdA}^* \)

The mutant gene \( \text{icdA}^* \) had eight silent mutations. These changes caused no change in the protein sequence, but altered the mRNA sequence. Similar to the studies on \( \text{ptsG}^* \), the mutant \( \text{icdA}^* \) was expressed under the control of its native promoter in the deletion mutant of \( E. \text{coli} \ \Delta \text{icdA} \). This final mutant was designated as SQ-003 (\( E. \text{coli} \ \Delta \text{icdA} \ \text{pBT4.}\text{icdA}^* \)). The two mutants \( E. \text{coli} \ \Delta \text{icdA} \) and SQ-003 were characterized on minimal medium with glucose and xylose. The results of the characterization studies are shown in Figure 5.2. The deletion mutant \( \text{icdA} \) did not show any growth on the minimal medium indicating that the mutation of \( \text{icdA} \) is lethal for \( E. \text{coli} \) on minimal medium. These results were in agreement to the results reported by Baba et al. (2006). On addition of the \( \text{icdA}^* \), the growth of the mutant \( E. \text{coli} \ \Delta \text{icdA} \) was restored, indicating the fully functional expression of the mutant \( \text{icdA}^* \) gene. However, despite the slow growth rate, the mutant consumed glucose and xylose sequentially similar to the wild-type strain. The growth characteristics of the mutant showed clear repression of xylose in presence of glucose. These results indicate that the mutant \( \text{icdA}^* \) too had no role to play in the glucose-xylose co-utilization phenotype of the mutant LMSE\(_2\).

The overall results from the characterization studies of the mutated genes in the mutant LMSE\(_2\) suggest that co-utilization of glucose and xylose in the mutant was accomplished without any significant genetic changes. The most likely mechanism by which the mutant LMSE\(_2\) could accomplish co-utilization was either by changing the gene expression levels of the proteins involved or by altering the phosphorylation states of the proteins participating in CCR mechanism. These results are important as they indicate that \( E. \text{coli} \) can alternate between different metabolic states due to constraints on the metabolism alone, without any significant genetic changes to the regulatory network.

These studies, however, are preliminary in nature. Significant detailed studies need to be carried out to ascertain that the expression levels of \( \text{ptsG}^* \) and \( \text{icdA}^* \) had no role to play in glucose-xylose co-utilization. The best approach to accomplish such a study would be using chromosomal replacement of the native genes with their corresponding mutant variants. The combined effect of expressing both the mutant genes simultaneously also needs further investigation.
Figure 5.2: Complementation studies on the mutant SQ-003 for characterization of the mutant gene $icdA^*$. The symbol, $\bullet$ represents biomass, $\diamond$ represents glucose concentration, and $\triangle$ represents xylose concentration.

### 5.5 Conclusions

In this study, the genetic changes in the mutant LMSE$_2$ were investigated to identify the molecular mechanism of the glucose-xylose co-utilization. Two genes, $ptsG$ and $icdA$, were identified to carry mutations. The mutated variants of these genes, however, did not show any role in glucose-xylose co-utilization phenotype of the mutant. The results from the study suggest that *E. coli* can change its metabolic state and sugar utilization profile without any significant mutations in its regulatory network. The most likely mechanisms to achieve such a change in metabolic state is by changing the expression levels of different genes, or by changing the phosphorylation state of the regulatory proteins. Additional studies are required to ascertain the precise mechanism by which simultaneous glucose and xylose uptake is possible in the mutant LMSE$_2$. 
Chapter 6

D-Ribose Production Using the
Escherichia coli Mutant LMSE$_2$

This chapter contains material from our publication (Gawand and Mahadevan, 2014):


Reproduction of the material from the publication is a right that has been granted to the authors by the publisher, Mary Ann Liebert, Inc.

6.1 Abstract

D-Ribose is a commercially important functional sugar used as a nutritional supplement and as a starting compound in synthesis of antiviral drugs. In this study, we report engineered E. coli mutants that can produce D-ribose from glucose and xylose. Two endogenous haloacid dehalogenase-like (HAD) phosphatases from E. coli, HAD12 and HAD13, encoded by the genes $ybiV$ and $yidA$, respectively, were expressed in E. coli wild-type and the glucose-xylose co-utilizing mutant LMSE$_2$. All the mutants constructed in this study produced D-ribose. The mutant RB-006 (LMSE$_2$ expressing $ybiV$) showed the highest D-ribose titre of 1.16 g/L from 5 g/L each of glucose and xylose. Additionally, using xylose feeding, D-ribose titre was improved
to 3.36 g/L. Xylulose and acetate were formed as the major by-products in the fed-batch study. This study is the first example of engineered *E. coli* for production of D-ribose. This study also demonstrates reengineering of a glucose-xylose co-utilizing mutant of *E. coli* for production of a valuable chemical.

### 6.2 Introduction

D-Ribose is a naturally produced pentose sugar with a commercial demand for several applications. D-Ribose is used as a sweetener as well as a starting compound for synthesis of riboflavin, inosine monophosphate, and several antiviral drugs (De Wulf and Vandamme, 1997; Toivari et al., 2010). Additionally, D-ribose is used as a nutritional supplement for muscle recovery among athletes (Zarzeczny et al., 2001), and also aids in functional recovery of patients with ischemic cardiovascular disease (Shecterle et al., 2011). D-ribose can also be potentially used as a starting compound for production of its stereoisomer L-ribose, which is has a market price of around $1000/kg as compared to $20-30/kg for D-ribose (Woodyer et al., 2008).

Earliest methods of synthesizing D-ribose involved enzymatic hydrolysis of yeast RNA, or chemical synthesis from glucose, arabinose, gluconic acid, and xylose (Wu et al., 2009). However, these methods suffer poor conversion yields. Fermentation-based processes were then developed for replacing the chemical and biochemical methods for commercial production of D-ribose (De Wulf and Vandamme, 1997). Several organisms such as *Penicillium brevicompactum*, *Pseudomonas reptiliwora*, and *Candida pelliculose* are known to produce D-ribose naturally. However, fermentation processes using these organisms have found limited applications due to low yields of D-ribose and difficulties in culturing them.

Current industrial fermentation processes for D-ribose production employ transketolase deficient *Bacillus* spp. that were isolated using random mutagenesis (De Wulf and Vandamme, 1997). Absence of transketolase (or another pentose phosphate pathway (PPP) enzyme D-ribulose-5-phosphate-3-epimerase) results in accumulation of ribose-5-phosphate (R5P), which is converted into D-ribose by endogenous phosphatase(s) in *Bacillus* spp. Similar transketolase mutants of other industrially relevant organisms such as *E. coli* (Zhao and Winkler, 1994) and *Corynebacterium glutamicum* (Ikeda et al., 1998) do not show accumulation of D-ribose despite supplementation of aromatic amino acids and vitamins (pyridoxine). *Saccharomyces cerevisiae* strain with deletion of transketolases (TKL1 and TKL2) and phosphoglucone isomerase (PGI1), and overexpression of two genes, GHD2 (encoding NAD$^+$– dependent glutamate dehydrogenase) and DOG1 (encoding 2-deoxyglucose-6-phosphate phosphatase), was recently shown to produce D-ribose and ribitol (Toivari et al., 2010). The most preferred *Bacillus* strain for com-
Chapter 6. D-Ribose Production Using the *Escherichia coli* Mutant LMSE₂

Table 6.1: D-Ribose producing strains.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Growth conditions and supplements (if any)</th>
<th>Yield</th>
<th>Titre</th>
<th>Productivity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. subtilis</em> ATCC21951</td>
<td>100 g/L glu, 100 g/L xyl, batch mode</td>
<td>0.30</td>
<td>60</td>
<td>0.55</td>
<td>(De Wulf and Vandamme, 1997)</td>
</tr>
<tr>
<td><em>B. subtilis</em> SPK1</td>
<td>20 g/L glu, 20 g/L xyl, batch mode</td>
<td>0.58</td>
<td>23</td>
<td>0.72</td>
<td>(Park et al., 2004)</td>
</tr>
<tr>
<td><em>B. subtilis</em> SPK1</td>
<td>20 g/L glu, 20 g/L xyl, fed-batch mode a</td>
<td>0.43</td>
<td>46.6</td>
<td>0.88</td>
<td>(Park et al., 2004)</td>
</tr>
<tr>
<td><em>B. subtilis</em> JY200</td>
<td>5 g/L glu, 5 g/L xyl, yeast extract</td>
<td>0.05</td>
<td>0.25</td>
<td>0.03</td>
<td>(Park et al., 2006)</td>
</tr>
<tr>
<td><em>B. subtilis</em> JY200</td>
<td>20 g/L glu, 20 g/L xyl, fed-batch mode b</td>
<td>0.24</td>
<td>10.1</td>
<td>0.29</td>
<td>(Park et al., 2006)</td>
</tr>
<tr>
<td><em>B. subtilis</em> EC2</td>
<td>180 g/L glu, 0.3 g/L citrate, batch mode</td>
<td>0.59</td>
<td>84.3</td>
<td>1.96</td>
<td>(Wu et al., 2009)</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>20 g/L fru, 2 g/L glu, YNB and amino acids d</td>
<td>–</td>
<td>1.0 c</td>
<td>0.01</td>
<td>(Toivari et al., 2010)</td>
</tr>
</tbody>
</table>

a Feed: 100 g/L xyl;
b Feed: 200 g/L xyl + 50 g/L glu;
c Includes both D-ribose and ribitol concentration;
d The amino acids were used in triple concentration.

Commercial production of D-ribose is *B. subtilis* (De Wulf and Vandamme, 1997; Park et al., 2006). Many different variants of *B. subtilis*, with transketolase deletion, have been reported to produce D-ribose from glucose and other carbon sources at high yields and titres. A summary of D-ribose producing strains with their culture conditions is presented in Table 6.1.

Apart from *Bacillus* spp., *E. coli* is a strong candidate for D-ribose production, especially from glucose-xylose mixtures obtained from lignocellulosic biomass. *E. coli* has been extensively studied for growth on lignocellulosic hydrolysates as feedstocks (Mills et al., 2009; Saha and Cotta, 2012). Additionally, there exists a possibility for further engineering of D-ribose producing *E. coli* for production of L-ribose. The *E. coli* mutant LMSE₂, which contains three metabolic gene deletions, *pgi*, *rpe*, and *eda*,
Figure 6.1: a) Metabolic network showing modifications in the mutant LMSE$_2$ for production of D-ribose. Additional reactions, ybiV or yidA, have been added to the network showing phosphatase activity required for D-ribose production. The values beside the reaction names are the flux values expressed in mmol/gDW/h. The top values on the arrow (in black) are the wild-type fluxes and the bottom values (in grey) are the mutant LMSE$_2$ fluxes. Note that the directions of the arrows in the metabolic map represent the directions of the fluxes and not irreversibility of the reactions. The names of all the reactions and the metabolites in the figure can be found in the list of abbreviations provided in the beginning of this thesis. b) Robustness analysis of D-ribose production flux with respect to glucose exchange flux in the mutant LMSE$_2$, and c) robustness analysis of D-ribose production flux with respect to xylose exchange flux in the mutant LMSE$_2$. 
is capable of strictly co-utilizing glucose and xylose (Gawand et al., 2013). Due to the gene deletions, the glucose flux in the mutant is channeled into the PPP, converting glucose exclusively into R5P. Whereas, xylose is exclusively converted into xylulose-5-phosphate (X5P). Using R5P and X5P, the mutant LMSE\textsubscript{2} can synthesize all the biomass precursors using the PPP. A metabolic map of the mutant LMSE\textsubscript{2} is shown in Figure 6.1a.

As glucose is exclusively converted into R5P in the mutant LMSE\textsubscript{2}, overexpression of a phosphatase capable of cleaving the phosphate bond in R5P was expected to result in D-ribose production. Additionally, as the mutant LMSE\textsubscript{2} consumes xylose simultaneously with glucose, xylose was expected to be utilized for synthesis of biomass precursors, thereby increasing the yield of D-ribose per mole of glucose consumed. The mutant is incapable of converting to xylose to D-ribose due to the deletion of the gene \textit{rpe} (Figure 6.1a).

With the aim of producing D-ribose using the mutant LMSE\textsubscript{2}, we selected two endogenous phosphatases that could cleave the phosphate bond in R5P. A family of twenty-three haloacid dehalogenase-like (HAD) phosphatases in \textit{E. coli} was previously screened by Kuznetsova et al. (2006) against 80 different phosphorylated substrates. Among these twenty-three enzymes, four HAD phosphatases were active towards dephosphorylation of R5P. Two phosphatases, HAD12 (encoded by \textit{ybiV}) and HAD13 (encoded by \textit{yidA}), were chosen to be expressed in the mutant LMSE\textsubscript{2}, based on their kinetic properties (Kuznetsova et al., 2006). Multiple mutants were constructed by transforming \textit{E. coli} wild-type and the mutant LMSE\textsubscript{2} with the genes \textit{ybiV} and \textit{yidA} on an appropriate vector. The mutants thus constructed were characterized for production of D-ribose from a mixture of glucose and xylose.

6.3 Materials and Methods

6.3.1 FBA and robustness analysis

Flux distributions for the wild-type and the mutant LMSE\textsubscript{2} for D-ribose production were obtained using flux balance analysis (FBA). All the simulations were carried out using the \textit{E. coli} genome-scale model iJO1366 (Orth et al., 2011). The model was modified by addition of the phosphatase reaction and the flux values were obtained by simulating maximization of D-ribose production flux. The glucose and xylose exchange fluxes were fixed to −5 mmol/gDW/h each and the minimum biomass production flux (specific growth rate) was fixed to 0.1 h\textsuperscript{−1}. To obtain a realistic wild-type flux distribution, additional constraints were added to the model to divert 30% of total glucose uptake flux into the PPP. To simulate the LMSE\textsubscript{2} mutant phenotype, six reactions (PGI, RPE, EDA, XYLI2, PPS, and DRPA) were deleted from the model (see Chapter 3 for details). All the simulations were performed in MATLAB environment.
using COBRA Toolbox v.2 (Schellenberger et al., 2011). Robustness analysis was performed for D-ribose production flux against glucose uptake flux, and D-ribose production flux against xylose uptake flux in the mutant LMSE\textsubscript{2} background. The minimum biomass production rate was constrained to 0.1 h\textsuperscript{-1}. The analysis was carried out using the in-built function in COBRA Toolbox v.2.

6.3.2 Strains and plasmids

All the strains used in this study were obtained from Coli Genetic Stock Center (CGSC), Yale University. \textit{E. coli} K-12 MG1655 was used as the wild type strain. The mutant LMSE\textsubscript{2} (\textit{E. coli} Δ\textit{pgi} Δ\textit{eda} Δ\textit{rpe}::\textit{kanR}), described in Chapter 4, was engineered by adding genes, either \textit{yidA} or \textit{ybiV}. Both the genes were PCR amplified using forward and reverse primers with restriction enzyme sites for KpnI and HindIII, respectively (the primer sequences can be found in Appendix C). The amplified genes were ligated into the plasmid pTrc99A, and the plasmids were transformed into the mutant LMSE\textsubscript{2}. Appropriate controls were constructed by transforming \textit{E. coli} wild-type with the same plasmid constructs, and also by transforming the mutant LMSE\textsubscript{2} and the wild-type with the empty plasmid. All the constructed mutants with their genotypes and designations are listed in Table 6.2.

6.3.3 Phenotypic tests

All the six constructed mutants were first tested for growth and D-ribose production using phenotypic characterization. All the mutants were grown on Luria-Bertani (LB) plates supplemented with 5 g/L glucose, 5 g/L xylose, and 100 µg/mL ampicillin. Individual colonies were grown overnight in 3 mL LB broth supplemented with glucose (5 g/L), xylose (5 g/L), and ampicillin (100 µg/mL). The overnight grown cells were washed in minimal medium without sugars to remove all the LB, and were used to inoculate three different media: 1) minimal medium, 2) minimal medium supplemented with 0.5 g/L yeast extract, and 3) minimal medium supplemented with 0.5 g/L LB. Each tube was supplemented with 1 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG) to induce the plasmid. Mutants constructed from the wild-type (RB-001, RB-002, and RB-003) were checked for growth and production of D-ribose after 12 hours of incubation, and the mutants constructed from the mutant LMSE\textsubscript{2} (RB-004, RB-005, and RB-006) were checked for growth and D-ribose production after 24 hours of incubation. Growth in the tubes was measured as OD\textsubscript{550}, and D-ribose was detected using HPLC (method described below).

6.3.4 Media and batch cultivation

For batch fermentation, the constructed mutants were cultivated in 500 mL fermenters (Applikon MiniBio) with 300 mL working volume. Minimal medium with composition described in Appendix
**Chapter 6. D-Ribose Production Using the Escherichia coli Mutant LMSE₂**

E. coli was used for cultivation of all the mutants. Glucose and xylose concentration in the minimal medium was 5 g/L each. Additional supplementation of yeast extract (0.5 g/L) was added to the medium to enable growth of the LMSE₂-based mutants. Ampicillin (100 µg/mL) was added to the medium to maintain the selection pressure on the plasmid. The plasmid was induced using 1 mM IPTG when the culture reached an OD_{550} of 1.0. Fermenters were sampled every hour or every three hours depending upon the growth rates of the mutants. The batch was kept aerobic (>30% dissolved oxygen) by altering the impeller speed. The pH of the medium was maintained at 7.0 using 5 M KOH.

The inoculum was prepared by first inoculating a fresh colony in LB with 5 g/L glucose and 5 g/L xylose. The overnight grown cells were first washed and then transferred to minimal medium with 5 g/L glucose and 5 g/L xylose, and were grown until OD_{550} was between 1.5 – 2.0. The cells were then harvested and an appropriate amount was used to inoculate the fermenters to get an initial OD_{550} of 0.1.

Every sample was analyzed for OD_{550}, residual sugars, D-ribose, and co-products (acetate, xylulose, fructose, and erythrose). The sugars were separated using a Bio-Rad HPX-87H cation-exchange column (5 mM H₂SO₄ mobile phase, 0.4 mL/min flow rate, 42 °C column temperature, 20 µL injection volume).

**6.3.5 Fed-batch cultivation**

Fed batch cultivation of D-ribose production was carried out by feeding an appropriate amount of 300 g/L xylose feed. Xylose was added as an impulse-feed to match the residual concentration of glucose in the fermenter after all the initial xylose was consumed. Exhaustion of xylose in the medium was indicated by rise in pH and DO levels, as the mutant could not grow on residual glucose. Due to the very high concentration of xylose feed, the dilution effect in the fermenters was negligible. The impulse feed volume added was approximately 3 mL. The samples were withdrawn every 3 hours until the end of the batch.

**6.4 Results**

**6.4.1 FBA and robustness analysis**

Predicted flux distributions for production of D-ribose in the wild-type and the mutant LMSE₂ are shown in Figure 6.1a. The flux value for D-ribose production in the wild-type represents the maximum possible theoretical yield of D-ribose from glucose and xylose. However, such high yields are achievable only if most of the carbon flux in the wild-type is diverted towards D-ribose production and growth rate is as low as 0.1 h⁻¹. This flux distribution is very unlikely in the wild-type as majority of the carbon is diverted...
towards biomass formation. To achieve the FBA-predicted yields in the wild-type, major rerouting of the central metabolism fluxes (as shown in Fig. 6.1a) is required. The flux distribution of the mutant LMSE$_2$ shows that majority of D-ribose in the mutant is produced from glucose being diverted into the PPP. The mutant LMSE$_2$ has three gene deletions that cause the desired rerouting of the fluxes, and all of the glucose consumed is diverted into the PPP. The FBA-predicted fluxes of the mutant LMSE$_2$ are plausible and the mutant is expected to show high D-ribose yield. Robustness analyses of D-ribose production flux with respect to glucose exchange flux (Fig. 6.1b) and xylose exchange flux (Fig. 6.1c) in the mutant LMSE$_2$ show that D-ribose is almost exclusively produced from glucose. Decreasing the glucose uptake flux proportionally decreases the D-ribose production flux, whereas xylose exchange flux has no effect on the D-ribose production flux. These results suggest that xylose is used by the mutant only for making biomass precursors.

### 6.4.2 Phenotypic characterization

Phenotypic characterization of the mutants was carried out as a preliminary investigation to detect D-ribose production in *Escherichia coli*. The results for the phenotypic characterization are summarized in Table 6.2. The control mutants (RB-002 and RB-003) showed production of D-ribose under all the three media conditions tested. No D-ribose was detected in the control expressing the empty plasmid (RB-001). The mutants based on LMSE$_2$ (RB-005 and RB-006) did not show growth on minimal medium, but showed growth and D-ribose production on minimal medium supplemented with yeast extract and LB. The mutant LMSE$_2$ transformed with the empty plasmid did not show any D-ribose production.

The most likely reason for no observed growth of the mutants RB-005 and RB-006 on minimal medium was the high activity of phosphatase on R5P, resulting in low intracellular concentrations of R5P. This high phosphate activity could have precluded the production of biomass precursors such as erythrose-4-phosphate (E4P) through the PPP. Though overexpression of phosphatase could have been depleting the pools of R5P in the wild-type based control strains RB-002 and RB-003 as well, the phosphatase expression was expected to be more deleterious for the mutants RB-005 and RB-006. This is mainly because the wild-type based strains could use alternative pathways to make the phosphorylated metabolites such as fructose-6-phosphate (F6P) and glyceraldehyde-3-phosphate (G3P) (which could be used to synthesize other biomass precursors such as E4P). For example, RB-002 and RB-003 could synthesize G3P using either Embden-Meyerhof-Parnas pathway, or the Entner-Doudoroff pathway, or the PPP. In contrast, the mutants RB-005 and RB-006 had to synthesize all the biomass precursors using the PPP. Such limitations on metabolism could have caused the expression of phosphatase to be lethal for the mutants RB-005 and RB-006 than the control strains RB-002 and RB-003.
Table 6.2: Phenotypes of the mutants constructed in this study.

<table>
<thead>
<tr>
<th>Mutant name</th>
<th>Background</th>
<th>Gene added</th>
<th>Growth medium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MM&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>RB-001</td>
<td>wild-type</td>
<td>empty plasmid</td>
<td>G+ R-</td>
</tr>
<tr>
<td>RB-002</td>
<td>wild-type</td>
<td>yidA</td>
<td>G+ R+</td>
</tr>
<tr>
<td>RB-003</td>
<td>wild-type</td>
<td>ybiV</td>
<td>G+ R+</td>
</tr>
<tr>
<td>RB-004</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt;</td>
<td>empty plasmid</td>
<td>G+ R-</td>
</tr>
<tr>
<td>RB-005</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt;</td>
<td>yidA</td>
<td>G- R-</td>
</tr>
<tr>
<td>RB-006</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt;</td>
<td>ybiV</td>
<td>G- R-</td>
</tr>
</tbody>
</table>

G - growth; R - D-ribose; <sup>a</sup>MM - minimal medium; <sup>b</sup>YE - yeast extract (added 0.5 g/L); <sup>c</sup>LB - Luria-Bertani broth (added 0.5 g/L)

Addition of yeast extract at a low concentration of 0.5 g/L most likely facilitated the growth of the mutants RB-005 and RB-006 by supplying the amino acids that otherwise these mutants could not synthesize in the presence of phosphatase. Based on the results from the phenotypic studies, yeast extract was added to all the subsequent fermenter cultivations. In addition, to enable initial growth in the fermenter, IPTG was added only after OD<sub>550</sub> had reached a value greater than 1.0.

### 6.4.3 Batch studies

Four D-ribose producing mutants RB-002, RB-003, RB-005, and RB-006 were characterized in detail using batch cultivation studies. The cells were cultivated in minimal medium with yeast extract as a supplement. Figure 6.2 shows the growth and D-ribose production characteristics of these mutants. All four mutants characterized produced D-ribose from the mixture of glucose and xylose. The mutant RB-006 accumulated highest concentration (1.16 g/L) of D-ribose. The most likely source of D-ribose in the mutant RB-006 (and RB-005) was glucose, as predicted by the simulations (Fig. 6.1b). The experimental results showed that D-ribose production stopped when the mutant RB-006 stopped consuming glucose (Fig. 6.2d). The stoichiometry of D-ribose production from glucose is shown in the equation below:

$$\text{glucose} + \text{ATP} + 2\text{NADP}^+ + \text{H}_2\text{O} \rightarrow \text{D-ribose} + 2\text{NADPH} + 2\text{H}^+ + \text{CO}_2 + \text{ADP} + \text{P}_i$$  \hspace{1cm} (6.1)

As is clear from the equation above that conversion of glucose to D-ribose requires a supply of both
ATP and NADP$^\text{+}$. The most likely reason for cessation of glucose uptake and its subsequent conversion of D-ribose could have been because of limited supply of ATP or an inability to maintain the balance of the cofactor NADPH.

In contrast to RB-006, the control strains RB-002 and RB-003 produced D-ribose from xylose. These strains did not produce much D-ribose during the consumption of glucose, but showed increased D-ribose production during the xylose consumption phase (Fig. 6.2a and 6.2b). The overall stoichiometry of D-ribose production from xylose is shown below:

\[
\text{xylose} + \text{ATP} \rightarrow \text{D-ribose} + \text{ADP} + \text{P}_i
\]

(6.2)

It is clear from the reaction equations 6.1 and 6.2 that both glucose and xylose can be stoichiometrically converted in D-ribose. However, the main difference between the two substrates is that glucose requires a supply of two additional NADP$^\text{+}$ molecules. As the mutant LMSE$_2$ cannot produce D-ribose from xylose, NADPH balance is an important consideration in optimizing D-ribose production in the mutant.

Another important difference between the wild-type and the mutant LMSE$_2$ was the sugar consumption pattern. Whereas the wild-type shows sequential utilization of sugars, the mutant LMSE$_2$ has been engineered to co-utilize glucose and xylose. Accordingly, the mutants based on the wild-type, RB-002 and RB-003, consumed glucose and xylose sequentially (Fig. 6.2a and 6.2b); however, the mutant RB-006, which was constructed in the LMSE$_2$ background, consumed some glucose simultaneously with xylose. Both mutants RB-005 and RB-006 consumed xylose preferentially over glucose. Even though the mutant LMSE$_2$ is designed to show strict co-utilization of glucose and xylose, the most likely reason for the lower consumption of glucose and higher dependence on xylose was the additional supplementation of yeast extract. As the mutants most likely synthesized some of the biomass precursors from yeast extract, the mutants did not utilize high amounts of glucose. Based on these results, the concentration of yeast extract was reduced to 0.2 g/L in the subsequent fed-batch experiments to force higher uptake of glucose. The mutants RB-003 and RB-005, both expressing the gene ybiV encoding for HAD12, produced more D-ribose compared to the corresponding mutants RB-002 and RB-004 expressing the gene yidA encoding for HAD13. This difference in D-ribose production was most likely due to the differences in the substrate specificity of the enzymes HAD12 and HAD13. HAD13 has higher promiscuity towards phosphorylated substrates and acts on such diverse substrates as E4P, glucose-1-phosphate, mannose-1-phosphate, fructose-1-phosphate, and glucose-6-phosphate (in addition to R5P). Whereas, the alternative substrate range for HAD12 is limited to fructose-1-phosphate, glucose-6-phosphate, and acetyl phosphate (Kuznetsova et al., 2006). Overexpression of such promiscuous enzyme as HAD13, could have caused
Figure 6.2: Growth plots of RB-series mutants, a) RB-002, b) RB-003, c) RB-005, and d) RB-006. The symbol, ● represents biomass, ○ represents glucose concentration, ▲ represents xylose concentration, ■ represents D-ribose concentration, and □ represents acetate concentration. Error bars represents standard deviation of three replicates.

general dephosphorylation of multiple intermediates inside the cell, resulting in reduced growth of the mutant RB-005.

Table 6.3 shows the carbon balance for the mutant RB-003 and the mutant RB-006. The mutant RB-003 consumed both glucose and xylose and produced mainly biomass and CO₂. Some amount of carbon was also lost as acetate due to the overflow metabolism. On the other hand, the mutant RB-006 did not consume all the glucose, and produced less amount of biomass as well as less amount of acetate. The overflow metabolism in the mutant was slower due to lower glycolysis rate (Fig. 6.1a). The
Table 6.3: Carbon balance for the mutants RB-003 and RB-006.

<table>
<thead>
<tr>
<th>Components</th>
<th>RB-003</th>
<th></th>
<th>RB-006</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>g/L</td>
<td>mmol C</td>
<td>g/L</td>
<td>mmol C</td>
</tr>
<tr>
<td>beginning of the batch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glucose</td>
<td>4.43</td>
<td>147.67</td>
<td>4.66</td>
<td>155.33</td>
</tr>
<tr>
<td>xylose</td>
<td>4.25</td>
<td>141.67</td>
<td>4.52</td>
<td>150.67</td>
</tr>
<tr>
<td>yeast extract</td>
<td>0.50</td>
<td>15.46(^a)</td>
<td>0.50</td>
<td>15.46</td>
</tr>
<tr>
<td>biomass</td>
<td>0.05</td>
<td>1.96(^b)</td>
<td>0.03</td>
<td>1.18</td>
</tr>
<tr>
<td>end of the batch</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>glucose</td>
<td>0.00</td>
<td>0.00</td>
<td>3.03</td>
<td>101.00</td>
</tr>
<tr>
<td>xylose</td>
<td>0.14</td>
<td>4.73</td>
<td>0.04</td>
<td>1.33</td>
</tr>
<tr>
<td>acetate</td>
<td>1.80</td>
<td>60.00</td>
<td>0.77</td>
<td>25.67</td>
</tr>
<tr>
<td>D-ribose</td>
<td>0.84</td>
<td>28.00</td>
<td>1.16</td>
<td>38.67</td>
</tr>
<tr>
<td>xylulose</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>biomass</td>
<td>1.81</td>
<td>70.89</td>
<td>0.75</td>
<td>29.38</td>
</tr>
<tr>
<td>CO(_2)</td>
<td>(143)(^c)</td>
<td></td>
<td>(126)(^c)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>306.75</td>
<td>163.63(^d)</td>
<td>322.63</td>
<td>196.04(^d)</td>
</tr>
</tbody>
</table>

\(^a\) The yeast extract was assumed to be composed of 70% protein. The carbon content of protein was assumed to be 0.53 g/g (Rouwenhorst et al., 1991).

\(^b\) The carbon content of *E. coli* biomass was assumed to be 47% on w/w (dry cell) (Heldal et al., 1985).

\(^c\) Assuming that the unaccounted carbon was lost as CO\(_2\).

\(^d\) Carbon lost as CO\(_2\) was not included to calculate the total carbon.

Results indicate that though the wild-type based mutant RB-003 is capable of producing D-ribose using expression of phosphatase, the mutant tends to waste a large fraction of carbon consumed as biomass, CO\(_2\), and the byproduct acetate. Although the LMSE\(_2\)-based mutant RB-006 grows more slowly and consumes sugars at a lower rate, it is capable of diverting more carbon consumed towards D-ribose production. As the enzymes HAD12 and HAD13 are promiscuous, multiple co-products could have been produced by the RB-series mutants. The samples were hence tested for fructose and erythrose. Neither fructose nor erythrose was detected in the fermentation samples. The only co-product detected in the batch cultures was acetic acid.

The results of the batch cultivation of the RB-series mutants are summarized in Table 6.4. The best performing mutants were RB-003 and RB-006. As anticipated, the yield of the LMSE\(_2\)-based mutant RB-006 was higher than the mutant RB-003 based on the wild-type, which was most likely due to the
Table 6.4: D-Ribose production characteristics by the RB-series mutants.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Mode</th>
<th>D-Ribose titre g/L</th>
<th>Yield mol/mol</th>
<th>Productivity g/L/h</th>
<th>Acetate titre g/L</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB-002</td>
<td>batch</td>
<td>0.38 ± 0.17</td>
<td>0.067 ± 0.012</td>
<td>0.048 ± 0.007</td>
<td>1.30 ± 0.28</td>
</tr>
<tr>
<td>RB-003</td>
<td>batch</td>
<td>0.84 ± 0.35</td>
<td>0.135 ± 0.007</td>
<td>0.094 ± 0.002</td>
<td>1.80 ± 0.72</td>
</tr>
<tr>
<td>RB-005</td>
<td>batch</td>
<td>0.42 ± 0.00</td>
<td>0.341 ± 0.039</td>
<td>0.012 ± 0.006</td>
<td>0.22 ± 0.05</td>
</tr>
<tr>
<td>RB-006</td>
<td>batch</td>
<td>1.16 ± 0.44</td>
<td>0.249 ± 0.085</td>
<td>0.050 ± 0.017</td>
<td>0.97 ± 0.43</td>
</tr>
<tr>
<td>RB-006</td>
<td>fed-batch</td>
<td>3.36 ± 0.12</td>
<td>0.328 ± 0.047</td>
<td>0.162 ± 0.004</td>
<td>1.96 ± 0.11</td>
</tr>
</tbody>
</table>

The errors represent standard deviation of three different cultivations.

direct channeling of glucose into D-ribose production. The concentration of D-ribose produced by the mutant RB-006 was also higher than the mutant RB-003. One aspect in which the mutant RB-003 was better than the mutant RB-006 was its productivity, which was due to higher growth rate and a higher sugar consumption rate of the mutant RB-003 compared to the mutant RB-006.

6.4.4 Fed-batch studies

The results obtained from the batch cultivation studies on the mutant RB-006 are shown in Figure 6.3. Due to the preferential consumption of xylose by the strain RB-006, xylose was exhausted in the fermenter faster than glucose. For around 5 g/L of xylose consumed, only 2.5 g/L of glucose was consumed. During this stage of consumption of glucose and xylose, approximately 2.5 g/L of D-ribose was synthesized. However, as RB-006 could not consume glucose alone, additional impulse of xylose was added to the fermenter to facilitate consumption of remaining glucose, and possibly increase the D-ribose titres. After addition of the feed, the concentration of both glucose and xylose was approximately equal to 4 g/L. Using the xylose fed, more glucose was consumed, and D-ribose titre increased to 3.36 g/L. In addition to D-ribose, most of the xylose consumed was converted into the byproduct xylulose. These results show that the mutant RB-006 could not consume glucose even in presence of xylose, and stopped growing as well as producing D-ribose. The most likely reason behind the poor glucose uptake and poor conversion of xylose to D-ribose could be the fact that utilization of both glucose and xylose require ATP. Glucose is phosphorylated while it is transported inside the cell, and xylose enters the PPP as X5P after being phosphorylated at the cost of an ATP. As overexpression of HAD12 dephosphorylated R5P, the flux through the PPP was very low leading to no or little production of intermediates required
to generate ATP. Due to low concentration of metabolites required to drive ATP production, complete conversion of glucose to D-ribose was not observed.

The final concentrations and yields of xylose and xylulose in fed-batch culture are listed in Table 6.4. The fed-batch studies on RB-006 suggest that additional engineering of the mutant is required for sufficient energy production to enable consumption of glucose and xylose. It appears that though D-ribose production was achieved in *E. coli* while enabling co-utilization of glucose and xylose, fine-tuning of phosphatase expression is required to achieve the desired results with fed-batch fermentations.

![Growth profile for fed-batch culture of the mutant RB-006](image)

**Figure 6.3:** Growth profile for fed-batch culture of the mutant RB-006. The symbol, ● represents biomass, ▲ represents glucose concentration, ▲ represents xylose concentration, ■ represents D-ribose concentration, ▲ represents acetate concentration, and ▲ represents xylulose concentration. Error bars represent standard deviation of three replicates.

### 6.5 Discussion

Unlike transketolase deficient *Bacillus* mutants, transketolase deficient *E. coli* mutants do not show accumulation or D-ribose (Zhao and Winkler, 1994). Causes for the absence of D-ribose production in the transketolase deficient *E. coli* could be either insufficient activity of the endogenous phosphatases or low intracellular concentration of R5P. We achieved production of D-ribose in *E. coli* by addressing both these concerns. Production of D-ribose was conceived in the mutant LMSE₂ based on the fact that glucose consumed by the mutant is channeled directly into the PPP, resulting in its exclusive conversion to R5P (Gawand et al., 2013). Additionally, it was presumed that R5P produced by the mutant could be dephosphorylated by overexpression of an endogenous phosphatase.
Overexpression of phosphatases, HAD12 or HAD13, showed D-ribose production in both wild-type and LMSE$_2$-based strains, RB-003 and RB-006, respectively. However, in RB-003, most of the carbon obtained from the sugars was converted into CO$_2$ and biomass, resulting in a decreased overall yield of D-ribose. In RB-006, glucose was converted almost exclusively into D-ribose resulting in higher yields. Another advantage of using RB-006 was that the mutant could co-utilize glucose and xylose, a trait very important in industrial fermentations (Kim et al., 2010).

Despite the advantages, the mutant RB-006 had certain limitations, such as incomplete conversion of glucose-xylose mixture into D-ribose, lower sugar-uptake and growth rate, and dependence on yeast extract for growth. Results obtained from the fed-batch characterization show that the xylose feed was almost exclusively converted into xylulose, and glucose was not completely consumed. Limited ATP production was most likely the reason for limited conversion of glucose and xylose into D-ribose and lower growth rate of the mutant. ATP is required for phosphorylation of both glucose and xylose as the first step of their catabolic pathways. It is also required for biomass production. Low ATP production in RB-006 was caused most likely due to the metabolic make-up of the mutant, which could support only a low glycolytic flux (Fig. 6.1a). Finally, the dependence on yeast extract for growth was due to high phosphatase activity causing depletion of R5P resulting in no flux through the PPP for production of biomass precursors.

All these issues associated with RB-006 are addressable and can be resolved by additional engineering of the mutant and/or a follow-up treatment such as adaptive evolution. A tight control over the expression of the phosphatase gene would solve multiple problems associated with the mutant. Low ATP production as well as limited production of biomass precursors in the mutant are mainly caused due to the limited flux through the PPP. High phosphatase activity leading to dephosphorylation of R5P depletes the R5P pools stalling the flux through the PPP. Controlled activity of phosphatase gene, allowing sufficient intracellular pools of R5P to drive the PPP, can ensure sufficient ATP production to allow continuous glucose and xylose uptake, and production of biomass precursors to reduce the dependence on yeast extract. Controlled expression of the phosphatase gene can be easily achieved using tightly regulated systems used to express toxic proteins in *E. coli* (Saïda et al., 2006). The mutant LMSE$_2$ is capable of growing without yeast extract (Gawand et al., 2013). If the phosphatase gene is completely repressed during the growth phase of the mutant RB-006, it would not require yeast extract supplementation, as all the biomass precursors would be provided from co-utilization of glucose and xylose. To eliminate the dependence on yeast extract and to improve the sugar uptake rates, adaptive evolution of the mutant can also be carried out.

For the *E. coli* mutants described in this study to compete with the industrial *Bacillus* strains,
multiple bioprocess considerations must also be investigated. One of the important characteristics of the industrial strains is high sugar tolerance. In *E. coli*, sugar tolerance is strain dependent (Son et al., 2011; Marisch et al., 2013; Negrete et al., 2010). The mutants constructed in this study were constructed in K-12 MG1655 background, which is a common industrial strain of *E. coli* and is known to grow at high glucose concentrations of 40 g/L (Son et al., 2011). Another consideration in industrial fermentations is the dissolved oxygen concentrations in the fermenters. Wild-type *E. coli* can convert up to 50% carbon into CO$_2$, thereby losing a significant amount of carbon under aerobic conditions (Colletti et al., 2011). Anaerobic and micro-aerobic conditions can potentially increase the product yields by conserving the carbon lost as CO$_2$ (Colletti et al., 2011). The current study characterized the mutants under high dissolved oxygen conditions in the fermenters (>30%). Cultivation of the mutants under reduced oxygen (microaerobic) conditions is expected to increase the yields of D-ribose by minimizing the loss as CO$_2$.

Optimum time for induction can also critically affect the productivities of the target compounds. The biomass produced by the mutant RB-006 was lower compared to RB-003, which increased the batch time and decreased the productivity of D-ribose. Optimization of the induction time of the plasmid can further improve the productivity of D-ribose by allowing sufficient biomass formation required to achieve high sugar consumption rates and high D-ribose production rates.

### 6.6 Conclusions

In this chapter, we demonstrated construction of *E. coli* mutants capable of producing D-ribose from a mixture of glucose and xylose. Using overexpression of endogenous phosphatase HAD12, and feeding xylose, we were able to achieve significant titres of D-ribose. Though good titres of D-ribose were achieved using the LMSE$_2$-based mutant RB-006, the titres were lower compared to the industrial strains. Significant amount of xylose was also converted into xylulose as a byproduct. These results indicate the need for additional metabolic engineering of the mutants to achieve industrially relevant titres of D-ribose.
Chapter 7

Conclusions

This chapter discusses the conclusions and contributions of this thesis in light of its objectives and hypotheses.

7.1 Conclusions

This thesis explored numerous problems with the main focus on development of a model-based metabolic engineering approach for engineering sugar co-utilization in industrial microorganisms. Following are the five studies reported in this thesis:

1. a novel bilevel optimization algorithm, SimUp, that identifies reaction knockout strategies to force co-utilization of two sugars was developed;

2. glucose-xylose co-utilization strategies for *Escherichia coli* using SimUp were identified;

3. five mutants (LMSE$_{1-5}$) based on SimUp predictions were constructed, among which the mutants LMSE$_2$ and LMSE$_5$ showed co-utilization of glucose and xylose;

4. the molecular mechanism of glucose-xylose co-utilization in the mutant LMSE$_2$ was investigated using Sanger sequencing, whole genome sequencing, and complementation studies; and

5. the mutant LMSE$_2$ was engineered for production of D-ribose.

This thesis was based on three main hypotheses listed in Chapter 2. All the studies were aimed at verifying the three hypotheses.

The first hypothesis stated that the regulatory phenomenon of carbon catabolite repression (CCR) could be circumvented by engineering metabolism of microorganisms, without direct manipulation to their regulatory genes. The hypothesis was tested by developing the algorithm SimUp to identify
metabolic reaction deletion strategies for elimination of CCR, and experimentally testing the SimUp-predicted strategies. The mutants constructed in this study were novel and could strictly co-utilize glucose and xylose. In agreement with the first hypothesis, these mutants could co-utilize the two sugars without any direct manipulations to their regulatory network. The mutants, however, suffered lower substrate uptake rates and lower specific growth rates than the wild-type. A unique feature of the mutants was that the growth of the mutants was strictly coupled to the sugar co-utilization phenotype. In other words, the mutants could grow only if they co-utilized the two sugars and could not grow on individual sugars. This trait gave two significant advantages: 1) the mutants could not revert back to their original wild-type phenotype of sequential sugar utilization, and 2) the sugar consumption rates could potentially be improved by adaptively evolving the mutants. Based on the results described in Chapter 3 and 4, it can be concluded that we successfully validated the first hypothesis and established a completely novel approach for engineering sugar co-utilization.

Studies on investigation of the CCR mechanism in *E. coli* using the mutant LMSE$_2$ were carried out to verify the second hypothesis of this thesis. The hypothesis stated that the gene-deletion mutants would accumulate additional mutations in the genes that control sugar utilization in response to the forced co-utilization of glucose and xylose. In addition, the mutants could have accumulated random mutations, in absence of the selection pressure, that could have played a role in co-utilization of glucose and xylose. Identification of both random and adaptive mutations was of interest to identify the regulatory mechanisms responsible for sugar co-utilization in the mutant LMSE$_2$. The glucose-xylose co-utilizing mutant LMSE$_2$ was unique in co-utilizing glucose and xylose without any direct manipulations to the regulatory genes involved in CCR. This characteristic of the mutant had a potential to unravel previously unknown regulatory mechanisms. Our investigation on the genome sequence of the unevolved mutant LMSE$_2$ found that the mutant did not have any mutations in its regulatory genes that were responsible for co-utilization of glucose and xylose. This observation suggested that the mutant could sense the requirement of xylose to survive and could overcome its wild-type phenotype to enable xylose utilization in presence of glucose. This shift in phenotype was accomplished without any additional mutations. The precise mechanism of how exactly the mutant began to utilize xylose could not be conclusively elucidated. The mechanism most likely involved changes in the expression levels of the *xyl* operon genes. The *xyl* operon in *E. coli* is positively regulated by intracellular xylose. It is hence likely that xylose was first transported through a non-specific transporter, which then activated the expression of *xyl* operon. The decreased glucose uptake rate responsible for altering the phosphorylation state of EIIAGlc could also have influenced the initial xylose uptake. Further investigations are required to understand these mechanisms and the mutant LMSE$_2$ can be used to unravel such cellular decision-making at molecular
level. Based on the results described in Chapter 5 of this thesis, no significant mutations were found to verify the second hypothesis. However, as the mutants were not adaptively evolved, the studies carried out were suitable only for detecting random mutations. Experiments with longer time-scale are required for identifying adaptive mutations and verification of the second hypothesis.

The third and final hypothesis of this thesis was that the mutants with co-utilization phenotype could be used to produce useful chemicals. Such mutants could improve the fermentation processes either by improving fermentation characteristics (yield, productivity, and titre), or by simplifying the fermentation process. We investigated the third hypothesis by reengineering the mutant LMSE$_2$ to produce D-ribose (mutant RB-006 in Chapter 6). Using the unique metabolism of the mutant LMSE$_2$, we were able to demonstrate higher yield and titre of D-ribose compared to the wild-type control. The main distinction between the metabolism of the wild-type and the mutant LMSE$_2$ was their sugar metabolism patterns. Whereas the wild-type produced D-ribose mainly from xylose, the mutant LMSE$_2$ produced D-ribose mainly from glucose. This difference in the metabolism allowed higher D-ribose production yields using the mutant LMSE$_2$. A few limitations of the mutant LMSE$_2$ (for D-ribose production) were reduced productivity and incomplete consumption of glucose. However, we clearly demonstrated that the mutants constructed in this thesis were of value as they offered improved fermentation characteristics than the wild-type, thereby confirming the third hypothesis.

In conclusion, we demonstrated a complete metabolic engineering study in this thesis, starting with model-based strain design and concluding with production of a high-value chemical using the mutants constructed. Throughout the thesis, importance of metabolic modeling was emphasized. Metabolic modeling was used not only for devising new engineering strategies, as demonstrated by SimUp, but also for understanding the metabolism of engineered organisms, as demonstrated by stoichiometric feasibility analysis of the mutant LMSE$_5$ (Chapter 3), and robustness analysis of the mutant LMSE$_2$ for D-ribose production (Chapter 6). In addition, the thesis also demonstrated the applications of metabolically engineered strains in study of fundamental phenomena (cellular response to forced phenotypes, Chapter 5) and improvement of metabolic models (updating the central metabolism model based on characterization of the mutant LMSE$_1$, Chapter 4). As a whole, this research demonstrates the important role metabolic engineering plays in the development and understanding of the industrial strains.

### 7.2 Contributions

Following are the main contributions of this thesis to the field of metabolic engineering and bio-based chemicals production:
1. **Novel strain design algorithm**: Strain design algorithms are generally formulated with an objective of improving the yield of the desired product. The SimUp algorithm reported in this thesis (Chapter 3) is the first strain design algorithm to circumvent the regulatory problem of CCR to enable sugar co-utilization. The novelty of SimUp lies in the formulation of its unique objective function and simulation of different media conditions in a single optimization problem. By solving a regulatory problem using metabolic models, this thesis contributes to the field of strain design by expanding the scope of model-based strain design algorithms. As SimUp does not require any knowledge of the regulatory network, it is anticipated to be especially useful for engineering of non-conventional industrial microorganisms.

2. **Novel glucose-xylose co-utilizing mutants of *E. coli***: Elimination of CCR has been an important metabolic engineering problem. Construction of sugar co-utilizing mutants by altering the regulatory genes of CCR may not result in strict co-utilization of sugars or may have undesired physiological effects. Designing engineering strategies may be difficult for organisms in which CCR is not fully understood. In this thesis, we constructed previously unreported mutants of *E. coli* capable of co-utilizing glucose and xylose (Chapter 4). None of the genes involved in CCR were altered in these mutants. In addition, the glucose-xylose co-utilization phenotype of the SimUp-based mutants was coupled to growth, thereby making the mutants very amenable to adaptive evolution. The mutants reported in this thesis are unique and demonstrate the utility of model-based strain design algorithms.

3. **Non-intuitive mutant LMSE_5**: Model-based strain design algorithms can predict non-intuitive metabolic engineering strategies that rational design approaches may not identify. Non-intuitive metabolic engineering strategies may arise due to the interactions between reactions coupled by co-factors or due to the stoichiometric imbalance of a central metabolite(s) participating in multiple reactions. The SimUp algorithm predicted a set of non-intuitive metabolic engineering strategies to force co-utilization of glucose and xylose. The mutant LMSE_5, based on one such non-intuitive strategy, successfully co-utilized glucose and xylose in agreement with the model-predictions. To the best of our knowledge, the mutant LMSE_5 is the first successful mutant that uses a metabolic engineering strategy based on stoichiometric imbalance, and is a working example of a model-predicted non-intuitive strategy. Successful prediction, construction, and characterization of the mutant LMSE_5 is a significant contribution towards model-driven metabolic engineering as a validation of superior predictive capabilities of metabolic models.

4. **Insights into CCR**: CCR is a very well studied phenomenon in *E. coli*. Multiple strategies for sugar co-utilization have been developed based on the understanding of CCR regulatory network of
E. coli. To force co-utilization of two sugars, it is generally assumed that working of CCR should be somehow disrupted. To understand the mechanism by which the mutant LMSE\textsubscript{2} overcame CCR, we carried out sequencing analysis on the mutant LMSE\textsubscript{2}. The sequencing studies and the subsequent complementation studies suggested that the mutant LMSE\textsubscript{2} had no alterations in its CCR network (Chapter 5). This study was the first to show that genetic alterations to the regulatory network may not be essential to cause a shift in metabolism and sugar uptake of an organism.

5. D-Ribose producing mutants of E. coli: D-Ribose is commercially produced using mutants of Bacillus spp. Similar mutants of other organisms, including E. coli, did not show production of D-ribose, most likely due to low intracellular pools of ribose-5-phosphate and/or low phosphatase activity to convert ribose-5-phosphate into D-ribose. This thesis reports construction of the first engineered mutants of E. coli capable of producing D-ribose (Chapter 6). D-ribose production was achieved by expression of endogenous phosphatases in the glucose-xylose co-utilizing mutant LMSE\textsubscript{2}. Using xylose feeding, significant titres of D-ribose were achieved. Though not fully optimized for industrial production, the mutants of E. coli constructed in this study were the first reported mutants capable of converting lignocellulosic sugars into D-ribose.
Chapter 8

Recommendations for Future Work

- **Improvements to the SimUp algorithm:** The SimUp algorithm is formulated as a mixed integer linear program (MILP) (Chapter 3). For such combinatorial formulations the computational effort increases exponentially with the number of the variables. The solutions for SimUp presented in this thesis were obtained using a central metabolism model for *Escherichia coli* (103 reactions and 73 metabolites). The time required to solve SimUp using a genome-scale model (iAF1260, 2077 reactions and 1039 metabolites) was found to be prohibitively long. Scalability of strain design algorithms to genome-scale models has been a subject of multiple studies. Various approaches have been adopted to reduce the computational effort of strain design algorithms. Preprocessing of metabolic models to reduce size, use of local search constraints (Lun et al., 2009), use of heuristic algorithms (Patil et al., 2005), and duality-based novel mixed integer programming techniques (Kim et al., 2011b) have been previously used to reduce the computational time of strain design algorithms.

Ability to use genome-scale models instead of central metabolism models will remarkably improve the utility of SimUp. The suggested directions to improve SimUp’s computational efficiency include preprocessing of metabolic models to reduce the number of variables and application of novel techniques developed by Kim et al. (2011b) for improving the solution time of MILP algorithms. Use of these techniques is expected to significantly slash the computational time for SimUp, making it accessible to be used with genome-scale reconstructed metabolic network of any organism.

In addition, the SimUp algorithm can be further improved by modifying the algorithm to enable the prediction of gene deletions instead of reaction deletions. Currently, SimUp predicts reaction deletions by setting the upper and lower bounds of the corresponding reactions equal to 0.
However, a reaction deletion may not necessarily correspond to one gene deletion. For example, deletion of a single reaction pfk (phosphofructokinase) in the strategy of the mutant LMSE$_5$ required deletion of two genes pfkA and pfkB. Gene-protein-reaction matrix stored in genome-scale models of organisms stores the relationships between genes, corresponding proteins, and reactions. This relationship matrix has been previously used to predict gene deletion strategies instead for strain design (Burgard et al., 2003; Kim et al., 2011b). Using the gene-protein-reaction matrix, SimUp can be modified to directly predict gene deletions instead of reaction deletions.

- **Application of SimUp to other industrially relevant microorganisms**: SimUp algorithm was successfully used to construct glucose-xylose co-utilizing mutants of *E. coli* (Chapter 4). Similarly, SimUp can be used to identify glucose-xylose co-utilization strategies for other industrially relevant organisms. As the algorithm is inherently agnostic to the metabolic model used, it can be used to identify strategies for any microorganism with a reconstructed metabolic network. *Saccharomyces cerevisiae* is an important industrial microorganism and cannot co-utilize glucose and xylose in its native form. Attempts to engineer carbon catabolite repression (CCR) network, as well as evolutionary engineering strategies have not been successful in developing strains of *S. cerevisiae* that can co-utilize glucose and xylose (Roca et al., 2004; Wisselink et al., 2009). It should, however, be noted that significant amount of research has been focused on improving xylose utilization by *S. cerevisiae*, without necessarily a focus on strict co-utilization of glucose and xylose. Many superior xylose utilizing strains of *S. cerevisiae* have been developed that show high xylose uptake rates. However, use of SimUp algorithm to identify strategies for *S. cerevisiae* may provide completely novel solutions to force co-utilization. Another organism that can be engineered for sugar co-utilization using SimUp is *Zymomonas mobilis*, which is an important industrial organism and has a potential to be used for production of bio-ethanol (Panesar et al., 2006).

The most important advantage of SimUp is that it does not require any knowledge of regulatory network involved in CCR. The algorithm can be especially useful in designing co-utilization mutants of organisms for which the regulatory network has not been completely elucidated. We suggest application of SimUp for engineering such non-conventional microorganisms.

- **Application of SimUp for co-utilization of other substrates**: Glucose-xylose co-utilization is a central problem in lignocellulosic biofuels industry. Co-utilization of other substrate combinations such as glucose-galactose, glucose-acetate, glucose-glycerol, and glucose-fructose are also important to enable the use of alternative feedstocks. For example, glucose and galactose are the two main sugars obtained from marine biomass, the third-generation renewable feedstock (Kim
et al., 2012a). SimUp algorithm can be easily extended to identify strategies to force co-utilization of any two carbon sources provided the metabolic model of the target organism includes the reactions to metabolize these compounds.

- **Improvement of growth rates of the co-utilizing mutants by adaptive evolution:** The mutants LMSE$_2$ and LMSE$_5$ constructed in this study showed glucose-xylose co-utilization, however at the cost of their growth rates and sugar uptake rates (Chapter 4). Decrease in the growth rates of mutants due to deletion of important gene(s) is caused due to the transition of the mutants into suboptimal growth phase (Segrè et al., 2002). The suboptimal growth rates of the mutants can be improved by subjecting the mutants to adaptive laboratory evolution. For adaptive laboratory evolution, microbial cells are cultivated under defined conditions for extended periods of time using serial transfers or in a chemostat. Such prolonged cultivation can allow for selection of superior phenotypes such as improved growth rate, improved tolerance to inhibitors, or improved substrate uptake rates (Dragosits and Mattanovich, 2013). The glucose-xylose co-utilization phenotype of the mutants LMSE$_2$ and LMSE$_5$ is strictly coupled to their growth, i.e., the mutants cannot grow without co-utilizing the two sugars (Chapter 3). The glucose-xylose uptake rates for these mutants can be easily improved by adaptively evolving the mutants on medium containing limiting concentrations of glucose and xylose. However, the possibility of the mutants evolving to grow on only glucose is also possible. The mutants may use alternative pathways that are otherwise repressed to allow growth on glucose alone. These alternative pathways were predicted by the genome-scale model of *E. coli* (Chapter 3). In case the mutants evolve to grow on glucose alone using alternative pathway(s), additional gene deletions might be required to force co-utilization in the evolved mutants.

The only limitation of the mutants LMSE$_2$ and LMSE$_5$ are their lower growth rates than the wild-type. Adaptive evolution to improve the growth rate of the mutants is an important follow up study. Mutants with growth rates comparable to the wild-type and simultaneous sugar consumption have potential to increase the productivities of fermentation processes for biofuels and bio-based chemicals.

- **Overexpression of the pentose phosphate pathway to increase growth rate:** The lower growth rate of the mutant LMSE$_2$ compared to the wild-type can be attributed to the lower uptake rates of glucose and xylose. Slow metabolism of glucose is also a limiting factor in the growth, as evident from the difference between the growth rates of the mutant grown on different ratios of glucose and xylose (Chapter 4). In the mutant LMSE$_2$, glucose is metabolized through the pentose phosphate pathway (PPP). One strategy to improve the growth rate of the mutant is to
overexpress the oxidative PPP, which will improve the rate of glucose metabolism. The oxidative PPP consists of three steps in which glucose-6-phosphate is converted into ribulose-5-phosphate. Overexpression of the oxidative PPP enzymes has been previously explored for improving xylose utilization in *S. cerevisiae* (Runquist et al., 2009). A similar approach can be adopted to increase glucose flux through the PPP in the mutant LMSE₂ to improve the growth rate of the mutant.

- **Characterization of the mutant genes *icdA* and *ptsG* from the mutant LMSE₂:** Chapter 5 discusses the sequencing studies carried out on the mutant LMSE₂ to identify mutations in CCR and non-CCR genes. The subsequent complementation studies suggested that the mutations identified in *icdA* and *ptsG* were not responsible for glucose-xylose co-utilization in the mutant. However, the complementation studies carried out in this thesis were only of preliminary nature. As both the mutated genes were expressed on a plasmid, the role of gene expression levels and the changes in the translation efficiency were not replicated exactly as present in the mutant LMSE₂. Exhaustive characterization of the mutated genes *icdA* and *ptsG* should involve chromosomal replacement of the genes using recombineering (Sharan et al., 2009). Additionally, the additive effect of the two mutations should also be tested by constructing a mutant carrying both the mutated genes. As the mechanism of co-utilization of the mutant LMSE₂ has not been clearly established, exhaustive characterization of the mutated genes is an important follow up study.

- **Gene expression studies on the mutant LMSE₂:** The exact mechanism of glucose-xylose co-utilization by the mutant LMSE₂ remains unexplained. The mutant was capable of expressing xylose transporters and xylose metabolizing enzymes in presence of glucose without accumulating any mutations in the regulatory genes of CCR. The most likely mechanisms by which the mutant co-utilized the two sugars was by varying its gene expression levels, or by changing the phosphorylation state of the regulatory proteins (especially EIIA\textsuperscript{Glc}) involved in the phosphorelay of glucose uptake (Chapter 2). The signal transduction molecule cAMP also plays a role, albeit questionable, in CCR of *E. coli* (Görke and Stülke, 2008a; Crasnier-Mednansky, 2008; Görke and Stülke, 2008b; Narang, 2009). Change in intracellular concentration of cAMP could also have played a role in xylose utilization in presence of glucose.

To identify the exact mechanism by which LMSE₂ co-utilized glucose and xylose, gene expression studies on the mutant LMSE₂ using RNAseq, analysis of phosphorylation state of the protein EIIA\textsuperscript{Glc}, and intracellular concentration levels of cAMP are suggested. The studies on phosphorylation state of EIIA\textsuperscript{Glc} and intracellular concentration of cAMP can be carried out by the methods described by Bettenbrock et al. (2007). These studies are expected to give detailed insights into the cellular response of the mutant LMSE₂ after being forced to co-utilize glucose and xylose.
• **Improvement of D-ribose producing mutant RB-006:** Chapter 6 describes D-ribose producing mutants constructed from the glucose-xylose co-utilizing mutant LMSE2. Though the mutant RB-006 produced significant amount of D-ribose from the mixture of glucose and xylose, the mutant had multiple limitations. Firstly, the mutant required supplementation of yeast extract for growth. Overexpression of phosphatase most likely caused a general dephosphorylation of the intracellular metabolites resulting in the poor growth on mineral medium, thereby requiring the supplementation. Secondly, the mutant consumed xylose preferentially over glucose, and could not consume glucose completely. The limited consumption of glucose resulted in lower D-ribose yields and inability to efficiently utilize the feed containing mixture of glucose and xylose. These problems were potentially caused due to the use of the expression system with a weakly regulated IPTG-inducible promoter.

We expect that a tight control of the phosphatase gene would solve multiple problems associated with the mutant RB-006. If the phosphatase gene is not expressed during the growth phase, the mutant is expected to grow without supplementation of yeast extract. Absence of yeast extract will force the mutant to co-utilize glucose and xylose, which will confer it the ability to completely consume glucose and efficiently utilize glucose and xylose feed. Controlled expression of the phosphatase gene can be easily achieved using tightly regulated systems used to express toxic proteins in *E. coli* (Saïda et al., 2006). Use of a non-promiscuous enzyme with activity specifically towards ribose-5-phosphate can also significantly improve D-ribose titres and improve growth rates. Between the enzymes encoded by the genes *yidA* and *ybiV*, the enzyme encoded by *ybiV* had less promiscuous activity than the enzyme encoded by *yidA* (Kuznetsova et al., 2006). Accordingly, strains expressing *ybiV* showed higher D-ribose production. However, the enzyme encoded by *ybiV* is known to have promiscuous activity towards glucose-6-phosphate and acetyl phosphate (Kuznetsova et al., 2006). By choosing an enzyme with minimal or no promiscuous activity, the titre of D-ribose and growth of the mutant is expected to improve.

Furthermore, *E. coli* can metabolize D-ribose by transporting it using an ABC transporter and subsequently phosphorylating it to ribose-5-phosphate using the enzyme ribokinase encoded by *rbsK*. D-ribose produced by the engineered mutant can thus be potentially re-utilized by *E. coli* as a carbon source. Deletion of *rbsK* from the D-ribose producing mutant is hence suggested for further improvement of D-ribose yield and titre.

Finally, adaptive evolution of the D-ribose producing mutant can be carried out to gradually eliminate its dependence on yeast extract. Studies on optimization of feed composition and feeding strategy are also suggested as future directions to improve D-ribose production.
Appendix A

Characterization of Higher-order Gene Knockout Mutants

A.1 Abstract

Metabolic networks are characterized by multiple redundant pathways that do not have a clear biological function. The redundancies in the metabolic networks are implicated in adaptation to random mutations and/or survival under different environmental conditions. Characterization of multiple gene knockout mutants (higher-order mutants) can uncover metabolic phenomena that are not observed in single gene knockout mutants, and can potentially aid in the identification of the roles of redundant metabolic pathways. In this study, we characterize the higher-order mutants constructed as intermediates to the LMSE$_{1-5}$ series mutants described in Chapter 4. Specifically, we investigate the effects of deletion of the glyoxylate-shunt gene, *aceA*, and the Entner-Doudoroff pathway gene, *eda*, on the growth-rate of the mutant *E. coli Δpgi*. Results indicate that deletion of *aceA* and *eda* reduces the growth rate of the mutant *E. coli Δpgi*, suggesting that activation of these genes plays an important role in adaptation of *E. coli Δpgi* mutant. We also investigate the effect of order of gene deletions on the growth of higher-order mutants. The results indicate that the order in which genes are deleted determine the phenotype of the mutants during the sub-optimal growth phase.

A.2 Introduction

Effects of deletion of non-lethal genes have been studied in great details in the model organism *E. coli* (Ibarra et al., 2002; Fong et al., 2006; Ishii et al., 2007; Baba et al., 2006). Fewer studies, however, have
Appendix A. Characterization of Higher-order Gene Knockout Mutants

focused on systematic characterization of multiple gene knockout mutants, also known as higher-order mutants, of *E. coli* (Nakahigashi et al., 2009; Butland et al., 2008). Characterization of higher-order mutants can uncover metabolic phenomena such as synthetic lethality (Suthers et al., 2009), synthetic rescues (Motter et al., 2008), and conditional lethality (Harrison et al., 2007), that cannot be observed in single-gene knockout mutants. Simple phenotypic characterization of higher-order mutants has lead to the discovery of unknown metabolic pathways in central carbon metabolism of *E. coli* (Nakahigashi et al., 2009). Systematic genetic perturbations and characterization of higher-order mutants can help in an in-depth understanding of metabolic responses, and possibly, roles of redundancies in the metabolic networks. The data from high-throughput characterization of such mutants can be integrated into genome-scale metabolic models, potentially leading to increased accuracy of their predictions.

Given the size of the metabolic network of *E. coli* (the latest genome-scale model accounts for 2251 metabolic reactions and 1136 metabolites (Orth et al., 2011)), it is extremely difficult to construct and characterize the complete set of double-gene knockout mutants. The best approach to study higher-order mutants is to systematically design and characterize mutants from important metabolic nodes. Important nodes in metabolism can be identified by analyzing the organization of the metabolic network. Topology of metabolic networks is well-characterized, and some metabolites are known to be highly connected compared to others (Jeong et al., 2000; Ma and Zeng, 2003). Such highly connected metabolites are known as hub-metabolites, and these metabolites generally have important physiological roles (Jeong et al., 2000; Ma and Zeng, 2003).

One of the hub-metabolites in metabolic network of *E. coli* is glucose-6-phosphate (G6P), which is the first branching point of the carbon flux between glycolysis, the pentose phosphate pathway (PPP), and the Entner-Doudoroff pathway. Additionally, the uptake of glucose in *E. coli*, and hence the cellular concentration of G6P, depends on the ratio of concentrations of pyruvate and phosphoenolpyruvate (PEP) (other two hub-metabolites). Gene deletion mutants around the G6P node show pronounced altered physiologies, and can provide insights into the glucose metabolism of *E. coli*. One of the most important genes in G6P metabolism is *pgi* which encodes for phosphoglucose isomerase (EC 5.3.1.9). The *pgi* knockout mutant of *E. coli* has been intensively studied (Charusanti et al., 2010; Fong et al., 2006). We use higher-order mutants of *E. coli* with *pgi* knockout, to investigate two fundamental questions: 1) Can growth rate of *E. coli* ∆*pgi* mutant be improved by deleting any additional genes? 2) Does the order in which genes are deleted in higher-order mutants affect the growth phenotype of the final mutants constructed?

The first question investigates the phenomenon known as synthetic recovery, in which the deletion of a gene improves the growth rate of a pre-constructed gene deletion mutant (Motter et al., 2008; Cornelius
Figure A.1: Metabolic pathways of *E. coli*, a) under normal growth conditions, and b) after deletion of the gene *pgi*. The major flux during normal growth is through glycolysis with no flux through either the Entner-Doudoroff pathway or the glyoxylate shunt. Both these pathways get activated immediately after deletion of *pgi*, and major flux diversion from glycolysis to the PPP takes place.

et al., 2011). It is well known that deletion of a major metabolic gene in microorganisms is immediately followed by a sub-optimal growth phase, where the growth rate of the mutant is lower than the wild-type (Fong et al., 2006; Segrè et al., 2002). The sub-optimal growth rate of the mutants can be gradually improved by adaptive evolution (Fong and Palsson, 2004). Additionally, it is also known that during sub-optimal growth phase, the mutant activates a number of pathways that are otherwise latent during the normal growth phase (Fong et al., 2006). Activation of pathways during the sub-optimal growth phase is considered to be a metabolic response that helps the mutants to cope with the sudden loss of an important gene. The Entner-Doudoroff pathway and the glyoxylate shunt are known to be activated in response to deletion of *pgi* in *E. coli* (Fig. A.1) (Fong et al., 2006).

Recently, modeling studies have shown that activation of latent pathways does not offer metabolic
advantage to the mutants growing in suboptimal growth phase (Cornelius et al., 2011). Rather, deletion of latent reactions was predicted to improve the suboptimal growth rates of the mutants (Cornelius et al., 2011). A logical explanation for this observation could be that the deletion of latent reactions eliminates non-optimal metabolic pathways, thereby forcing the mutants to grow at optimal/higher growth rates (Cornelius et al., 2011).

Due to the contradicting theories about the role of latent reactions in the suboptimal growth phase, their physiological significance remains unclear. We investigated the role of the glyoxylate shunt gene, aceA, and the Entner-Doudoroff pathway gene, eda, by characterizing various double-gene knockout mutants of E. coli (Table A.1). The double gene knockout mutants were also constructed by deleting genes in different orders to understand the effect of presence or absence of latent reactions (Entner-Doudoroff pathway and glyoxylate shunt) in an event of major gene (pgi) loss.

### A.3 Materials and Methods

#### A.3.1 Strains, media, and plasmids

All the strains and plasmids used in this study were obtained from Coli Genetic Stock Center (CGSC), Yale University. E. coli K-12 MG1655 was used as the wild-type control strain and to make the gene knockout mutants. A list of all the knockout mutants constructed in this study can be found in Table A.1. All the mutants were constructed using sequential P1 transduction method, by transferring the target gene deletion from the respective E. coli BW25113 mutants from the KIEO collection (Baba et al., 2006). The double mutants were also distinguished by the order in which the two genes were deleted. For example, E. coli \( \Delta \text{pgi}^1 \Delta \text{aceA}^2 \), and E. coli \( \Delta \text{aceA}^1 \Delta \text{pgi}^2 \) were two distinct mutants, where the superscripts denote the order in which the genes were deleted. The mutants were isolated on selection medium containing 25 \( \mu \)g/mL of kanamycin. The kanamycin cassettes were removed from the mutants using the rescue plasmid pCP20. All the gene deletions were confirmed using PCR (sequences of the primers used for confirming gene deletions are provided in Appendix C). To avoid any additional random mutations, the mutants were stored at \(-80^\circ\text{C} \) immediately after construction.

#### A.3.2 Mutant characterization

All the mutants were characterized for growth, and selected mutants were characterized for sugar utilization and product formation in aerobic shake-flask cultivations at 37\(^\circ\text{C}\). The minimal medium described previously (Causey et al., 2003) was used as the growth medium (Appendix E). Glucose at a concentration of 5 g/L was used as the only carbon source. Seed cultures were prepared by inoculating
Appendix A. Characterization of Higher-order Gene Knockout Mutants

a fresh colony in 10 mL Luria-Bertani broth. The overnight grown culture was used to inoculate 150 mL minimal medium in 500 mL baffled Erlenmeyer flasks to get an initial OD$_{550}$ of 0.1. Samples were withdrawn every hour or every two hours, and were analyzed for OD$_{550}$, glucose, and acetate concentrations. Glucose and acetate concentrations were measured using a Bio-Rad HPX-87H cation-exchange column (5 mM H$_2$SO$_4$ mobile phase, 0.4 mL/min flow rate, 42°C column temperature, 20 µL injection volume). The batch was assumed to be complete when the OD$_{550}$ reached a constant value. All studies were carried out in triplicates.

A.3.3 FBA and MOMA

Two commonly used constraint-based methods, flux balance analysis (FBA) (Orth et al., 2010) and minimization of metabolic adjustment (MOMA) (Segrè et al., 2002), were used to predict the growth rates of the mutants. Experimentally calculated glucose uptake rates (where available) were used to simulate the glucose exchange flux. All the simulations were performed using the latest genome-scale model of *E. coli*, iJO1366 (Orth et al., 2011). As FBA is not capable of estimating the growth rate changes with sequential gene deletions, MOMA was used to calculate the growth rate changes with different order of gene deletions. All the codes were implemented in MATLAB (The Mathworks Inc., Natick, MA), and CPLEX 11.2 (IBM ILOG) was used as the LP solver.

A.4 Results

A.4.1 Growth characterization of the mutants

Table A.1 lists all the mutants characterized for growth in this study, and Figure A.2 shows the growth profiles of the mutants. As can be noted from the figure, all the gene deletion mutants showed significantly lower growth than the wild-type, except *E. coli ΔaceA*, which had a growth rate slightly higher than the wild-type. Increase in growth rate of *E. coli ΔaceA* has been previously observed (Maharjan et al., 2005). Deletion of the Entner-Doudoroff gene *eda* caused the growth rate of the wild-type *E. coli* to drop significantly. It is known that the Entner-Doudoroff pathway is not active in the wild-type, or has a limited flux (Fong et al., 2006). Deletion of *eda*, resulting in disruption of the Entner-Doudoroff pathway, was hence not expected to cause significant change in the growth rate compared to the wild-type. However, the reason for the significant drop of growth rate in *E. coli Δeda* mutant could be accumulation of the intermediate compound, 2-dehydro-3-deoxy-gluconate-6-phosphate, which is known to be toxic to *E. coli* (Motter et al., 2008).

The growth rates of all the double mutants were found to be lower than *E. coli Δpgi*, suggesting
that both the Entner-Doudoroff pathway and the glyoxylate shunt play an important role during the
growth of *E. coli* Δpgi immediately after the gene deletion. If activation of either of the pathways
was responsible for sub-optimal growth of *E. coli* Δpgi, the growth rate of the mutant, in absence of the
second gene (eda or aceA), should have increased.

Slight but consistent differences in the growth rates of the mutants with different orders of gene
deletions were observed. To further investigate the effect of different orders of gene deletions, three
mutants, *E. coli* Δpgi, *E. coli* ΔpgiΔaceA, and *E. coli* ΔaceAΔpgi, were characterized in details
for sugar consumption and acetate formation profiles. The characteristics of these three mutants are
described further.

![Growth profiles of different single and double gene knockout mutants constructed in this study. a) Glyoxylate shunt mutants, b) Entner-Doudoroff pathway mutants. The error bars represent standard deviation of three replicates.](image)

**Figure A.2:** Growth profiles of different single and double gene knockout mutants constructed in this study. a) Glyoxylate shunt mutants, b) Entner-Doudoroff pathway mutants. The error bars represent standard deviation of three replicates.

### A.4.2 Characteristics of *E. coli* Δpgi mutant

The gene *pgi* encodes the enzyme phosphoglucone isomerase (EC 5.3.1.9), which is the first enzyme of
glycolysis and converts glucose-6-phosphate into fructose-6-phosphate. *E coli* Δpgi mutant has been char-
acterized in great detail for its growth characteristics (Fong and Palsson, 2004), metabolism (Canonaco
et al., 2001), and genetic changes over adaptive evolution (Charusanti et al., 2010). Major reported
consequences of *pgi* inactivation in *E. coli* are: 1) severely decreased growth rate due to lowered gly-
colytic flux (Fong et al., 2006), 2) accumulation of NADPH pools due to increased flux through the PPP
(Canonaco et al., 2001), 3) transient activation of the Entner-Doudoroff pathway and the glyoxylate
Table A.1: Growth rates of the glucose-6-phosphate node mutants.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\mu_{exp}$</th>
<th>$\mu_{pred}^{a,b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(h$^{-1}$)</td>
<td>(h$^{-1}$)</td>
</tr>
<tr>
<td>E. coli wild-type</td>
<td>0.35</td>
<td>0.82</td>
</tr>
<tr>
<td>E. coli ∆pgi</td>
<td>0.16</td>
<td>0.81</td>
</tr>
<tr>
<td>E. coli ∆eda</td>
<td>0.06</td>
<td>0.82</td>
</tr>
<tr>
<td>E. coli ∆aceA</td>
<td>0.37</td>
<td>0.82</td>
</tr>
<tr>
<td>E. coli ∆eda$^1$ ∆pgi$^2$</td>
<td>0.14</td>
<td>0.81</td>
</tr>
<tr>
<td>E. coli ∆pgi$^1$ ∆eda$^2$</td>
<td>0.13</td>
<td>0.81</td>
</tr>
<tr>
<td>E. coli ∆aceA$^1$ ∆pgi$^2$</td>
<td>0.13</td>
<td>0.81</td>
</tr>
</tbody>
</table>

$^a$ Glucose exchange rate = 8.9 mmol/gDW/h for all mutants.
$^b$ Growth rates of the mutants calculated using MOMA.

Growth characteristics and glucose consumption characteristics of the wild-type E. coli and E. coli ∆pgi are summarized in Table A.2. The characteristics of E. coli ∆pgi were found to be largely in agreement with those reported in the previous studies. The growth rate of E. coli ∆pgi was around 45% of the wild-type and the biomass yield was found to be slightly higher than the wild-type (Table A.2). The drop in the growth rate of E. coli ∆pgi has been reported to be as low as <20% of the wild-type (Charusanti et al., 2010). The comparatively higher growth rate observed in our study could be a result of the elaborate medium composition. As anticipated, no acetate production was detected in E. coli ∆pgi, in agreement with the previous studies (Fong et al., 2006).

A.4.3 Characteristics of E. coli ∆aceA ∆pgi mutants

Among the physiological changes in E. coli ∆pgi mentioned above, NADPH imbalance is a major consequence, perturbing a significant portion of metabolic network (Charusanti et al., 2010). The NADPH imbalance is caused due to high flux that is channeled through the PPP in absence of pgi, causing accumulation of NADPH. E. coli counters the NADPH imbalance by accumulating mutations in NADH/NADPH transhydrogenases encoded by the genes udhA and pntAB (Charusanti et al., 2010), and activating the...
Table A.2: Growth characteristics of pgi and aceA mutants

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\mu$ (h$^{-1}$)</th>
<th>$q_s$ (mmol/gDW/h)</th>
<th>$q_a$ (mmol/gDW/h)</th>
<th>$Y_{x/s}$ (gDW/mmol)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> wild-type</td>
<td>0.351 ± 0.004</td>
<td>8.91 ± 1.80</td>
<td>3.78 ± 0.90</td>
<td>0.043</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgi$</td>
<td>0.159 ± 0.001</td>
<td>3.89 ± 0.24</td>
<td>0.00 ± 0.00</td>
<td>0.047</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgi^1 \Delta aceA^2$</td>
<td>0.139 ± 0.000</td>
<td>2.73 ± 0.07</td>
<td>2.42 ± 0.04</td>
<td>0.051</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta aceA^1 \Delta pgi^2$</td>
<td>0.128 ± 0.004</td>
<td>2.42 ± 0.19</td>
<td>3.16 ± 0.51</td>
<td>0.053</td>
</tr>
</tbody>
</table>

*The errors represent standard deviation between three experimental results.*

$\mu$ = growth rate; $q_s$ = glucose uptake rate; $q_a$ = acetate secretion rate; $Y_{x/s}$ = biomass yield

Glyoxylate shunt (Fischer and Sauer, 2003). Activation of glyoxylate shunt alleviates the NADPH imbalance by diverting the metabolic flux away from NADPH producing isocitrate dehydrogenase (*icdA*) (Fischer and Sauer, 2003). As glyoxylate shunt is slightly less efficient for ATP production than the TCA cycle, activation of glyoxylate shunt can be considered as a suboptimal response to *pgi* deletion.

To investigate whether the absence of glyoxylate shunt would improve the growth rate of *E. coli* $\Delta pgi$, we constructed a pair of mutants with deletion of *aceA* and *pgi* in different order (*E. coli* $\Delta pgi^1 \Delta aceA^2$ and *E. coli* $\Delta aceA^1 \Delta pgi^2$). By characterizing the growth rates and substrate uptake rates of these mutants, we investigated the effect of glyoxylate-shunt on the suboptimal growth phase of *E. coli* $\Delta pgi$ mutants.

The growth profiles of the two mutants are shown in Figure A.3, and their properties are summarized in Table A.2. Slight differences were observed between the growth rates, biomass yields, acetate production rates, and glucose uptake rates of the mutants (Table A.2). These results clearly indicated that order of gene deletions affect the growth behaviour of the mutants in the suboptimal growth phase. As the metabolic network of both the mutants was exactly same (as both had the same genes deleted), the differences in the properties of the mutants were caused due to additional properties such as altered metabolite pools. The most likely reason for *E. coli* $\Delta pgi^1 \Delta aceA^2$ to have a higher growth rate compared to *E. coli* $\Delta aceA^1 \Delta pgi^2$ was the difference in the glucose uptake rate. As glucose uptake rate in *E. coli* is controlled by the intracellular pyruvate and PEP pools, changes in the concentration of these metabolites most likely caused the differences in the phenotypes of *E. coli* $\Delta pgi^1 \Delta aceA^2$ and *E. coli* $\Delta aceA^1 \Delta pgi^2$.

Different order of gene deletions also caused differences in acetate production rates. *E. coli* $\Delta aceA^1 \Delta pgi^2$ mutant produced more acetate than *E. coli* $\Delta pgi^1 \Delta aceA^2$. With *pgi* deleted before *aceA*, the mutant *E. coli* $\Delta pgi^1 \Delta aceA^2$ had lowered glycolytic flux and hence no overflow metabolism. Additionally, due to the activation of glyoxylate shunt immediately after the deletion of *pgi*, any acetyl-CoA produced was being consumed by the glyoxylate shunt. Deletion of *aceA* after *pgi*, which disrupted the glyoxylate
Figure A.3: Growth plots of two double mutants with different order of gene deletions, a) E. coli Δpgi\(^1\) ΔaceA\(^2\), and b) E. coli ΔaceA\(^1\) Δpgi\(^2\). The symbol, ○ represents biomass, ● represents glucose concentration, and ■ represents acetate concentration. The error bars represent standard deviation of three replicates.

shunt, caused the mutant E. coli Δpgi\(^1\) ΔaceA\(^2\) to produce some acetate which was observed (Fig. A.3a). Whereas, in the double mutant E. coli ΔaceA\(^1\) Δpgi\(^2\), the glycolytic flux was same as the wild-type with high overflow metabolism even after deletion of the first gene aceA. With deletion of pgi after deletion of aceA, the glycolytic flux was lowered; however, acetyl-CoA that had most likely accumulated from high initial glycolytic flux was secreted as acetate.

Based on the characterization of the two mutants, it was established that the mutant E. coli Δpgi ΔaceA can exist in two distinct metabolic states during the suboptimal growth phase.

A.5 Model simulations of the mutant phenotypes

FBA can be used to predict the maximum growth rates of the gene knockout mutants at a given substrate uptake rate. To compare the predicted growth rates with the experimentally observed results, we set the glucose uptake rates as found in the experiments. The results are summarized in Table A.3. The growth rates predicted by FBA were much higher than the experimentally observed growth rates for the wild-type and all the mutants. The growth rates calculated by FBA are the maximum possible growth rates for a given substrate uptake rate and thus represent the growth rates of fully evolved strains (Fong and Palsson, 2004). As the mutants characterized in our study were not adaptively evolved, the growth
Table A.3: FBA and MOMA simulations for the higher-order mutants.

<table>
<thead>
<tr>
<th>Mutants</th>
<th>$q_s$ (mmol/gDW/h)</th>
<th>$\mu$ (h$^{-1}$)</th>
<th>$q_x$ (mmol/gDW/h)</th>
<th>$Y_{x/s}$ (gDW/mmol)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> wild-type</td>
<td>8.91</td>
<td>0.82</td>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgI$</td>
<td>3.89</td>
<td>0.34</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgI^{1}$ $\Delta aceA^{2}$</td>
<td>2.73</td>
<td>0.23</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta aceA^{1}$ $\Delta pgI^{2}$</td>
<td>2.42</td>
<td>0.20</td>
<td>0</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>FBA</th>
<th>MOMA $^a$</th>
<th>FBA</th>
<th>MOMA</th>
<th>FBA</th>
<th>MOMA</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em> wild-type</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgI$</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta pgI^{1}$ $\Delta aceA^{2}$</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td><em>E. coli</em> $\Delta aceA^{1}$ $\Delta pgI^{2}$</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
<td>0</td>
<td>0.08</td>
</tr>
</tbody>
</table>

$^a$Glucose uptake rate = 3.89 mmol/gDW/h taken as the reference point.

*MOMA simulations for the higher-order mutants.*

rates did not match the FBA predicted growth rates. Additionally, FBA could not capture the observed acetate production caused due to overflow metabolism. FBA cannot be used to simulated growth rate of the mutants in different order of gene deletions as it does not account for intermittent growth rate changes on gene deletions. Hence, the growth rates calculated by FBA for the mutants *E. coli* $\Delta pgI^{1}$ $\Delta aceA^{2}$ and *E. coli* $\Delta aceA^{1}$ $\Delta pgI^{2}$ were exactly same.

MOMA is a constraint-based method that can simulate growth rates of gene deletion mutants in suboptimal growth phase. As the objective function used for MOMA is not biomass-maximization, rather, minimization of metabolic adjustment (Segrè et al., 2002), MOMA calculates the growth rate change after a gene deletion from a given reference state. MOMA can thus be used to simulate sequential changes to the growth rate with each gene deletion, and hence can conceptually distinguish between two mutants with different order of gene deletions. We first used MOMA to find the flux distribution of *E. coli* $\Delta pgI$ using the wild-type steady-state flux as the reference. However, MOMA could not predict the activation of either the Entner-Doudoroff pathway or the glyoxylate shunt in response to $pgI$ knockout.

To check whether MOMA could predict growth rate differences between the two mutants *E. coli* $\Delta pgI^{1}$ $\Delta aceA^{2}$ and *E. coli* $\Delta aceA^{1}$ $\Delta pgI^{2}$, we simulated the growth of *E. coli* $\Delta pgI^{1}$ $\Delta aceA^{2}$ using *E. coli* $\Delta pgI$ as the reference flux, and the growth of *E. coli* $\Delta aceA^{1}$ $\Delta pgI^{2}$ using *E. coli* $\Delta aceA$ as the reference flux. MOMA results obtained showed slight differences in the growth rates of the two mutants (Table A.3). However, since $aceA$ activation was not predicted to be active after $pgI$ deletion by MOMA, the difference in the growth rates was not caused due to presence or absence of $aceA$. 
A.6 Conclusions

This chapter focuses on characterization of the higher-order mutants around the glucose-6-phosphate node of *E. coli*. Two main target phenomena were investigated: 1) synthetic recovery, and 2) effect of order of gene deletion in higher-order mutants. We observed that deletion of genes from either the Entner-Doudoroff pathway or the glyoxylate shunt could not recover the reduced growth rate in *E. coli* Δ*pgi* mutant, suggesting that these latent pathways are not suboptimal, and play an essential role in adaptation of *E. coli* Δ*pgi*. Secondly, we found differences in the physiologies of the higher-order mutants with different orders of gene deletions. The differences in the phenotypes could not be explained using metabolic models alone, suggesting a role of regulation and/or metabolite concentrations in determining the physiology of higher-order mutants. The results indicate a requirement of more elaborate models that include kinetic information for prediction of phenotypes of higher-order mutants.
Appendix B

Protocol for Solving SimUp

The SimUp algorithm has been coded in AMPL. Following are the details on the SimUp code and the metabolic model of *Escherichia coli* that was used to solve the glucose-xylose co-utilization problem.

1. The code is divided into multiple files. Table B.1 lists all the files required to solve SimUp. The three AMPL files “code.run”, “model.dat”, and “exec.run” should be stored in the same folder to run the code.

   Table B.1: Files used to solve SimUp.

<table>
<thead>
<tr>
<th>File name</th>
<th>File type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>model.xml</td>
<td>SBML file</td>
<td>metabolic model file</td>
</tr>
<tr>
<td>segmatrix.m</td>
<td>MATLAB file</td>
<td>generates model.dat file</td>
</tr>
<tr>
<td>code.run</td>
<td>AMPL file</td>
<td>stores the code for SimUp</td>
</tr>
<tr>
<td>model.dat</td>
<td>AMPL file</td>
<td>stores the model information</td>
</tr>
<tr>
<td>exec.run</td>
<td>AMPL file</td>
<td>executes the code.run and model.dat files</td>
</tr>
<tr>
<td>results.txt</td>
<td>AMPL file</td>
<td>stores the results generated</td>
</tr>
</tbody>
</table>

2. All the files required to run the code have been stored in the shared folder named SimUp on the Google Drive, which can be accessed using the login lmse.ut. AMPL and CPLEX 11.2 have been installed on the servers FIRE and ICE, and can be accessed remotely.

3. Following steps are required to solve SimUp:

   **Step 1:** The first step is to convert the stoichiometric model of the organism into an AMPL-
readable file format. AMPL reads data from a text file with an extension “.dat”. The stoichiometric model of organism is generally available in SBML file format, which can be read using COBRA Toolbox in MATLAB environment. The SBML file of the organism is first loaded in MATLAB using <readCbModel> command in COBRA Toolbox. The model file is then converted into an AMPL-readable “model.dat” file using the MATLAB script stored in the file named “segmatrix.m”. The “model.dat” file contains the stoichiometric matrix, upper and lower bounds for the reaction fluxes, reaction names, metabolite names, and the objective function coefficient vector. The script “segmatrix.m” also segregates the stoichiometric matrix, the upper and lower bounds for the reaction fluxes, and the objective coefficient vector into reversible and irreversible equivalents.

**Step 2:** The values of additional parameters required to solve SimUp code are manually added to the “model.dat” file. These parameters include the growth rate of the wild-type on glucose, the growth rate of the wild-type of xylose, and the growth rate of the wild-type on glucose and xylose. These values are obtained by solving FBA for each of these cases (using COBRA Toolbox in MATLAB). Values of additional parameters listed in Table B.2 are also added to the “model.dat” file. After the “model.dat” file is ready to process, the SimUp code can be executed using “exec.run” file.

**Step 3:** The “exec.run” file requires the value of the parameter $K$, defining the limit on the number of knockouts. The value of $K$ can either be fixed or iteratively increased to find multiple solutions. Integrality tolerance for CPLEX solver is declared in “exec.run” file. The recommended value for integrality tolerance is $10^{-9}$.

**Step 4:** The executable file “exec.run” is used to run the code. This file has instructions to first read the “model.dat” file, and then execute the “code.run” file. For execution of the file, the AMPL command <include filename> is used.

**Step 5:** The results are stored in the text file named “results.txt”. This file stores the unique sets of gene knockout strategies and the corresponding values of the objective function.

**Step 6:** The results obtained from AMPL can be individually verified using COBRA Toolbox in MATLAB.

---

4. The central model of *E. coli* used to solve the glucose-xylose co-utilization problem using SimUp was originally published by Schilling et al. (2000). This model accounts for central carbon metabolism including the Embden-Meyerhof-Parnas pathway, the pentose phosphate pathway, the citric acid cycle, the Entner-Doudoroff pathway, and the electron transport chain. The reactions
Table B.2: Parameter values used for SimUp.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_R^L$</td>
<td>lower bound for reversible flux</td>
<td>-1000</td>
</tr>
<tr>
<td>$v_R^U$</td>
<td>upper bound for reversible flux</td>
<td>1000</td>
</tr>
<tr>
<td>$v_I^L$</td>
<td>lower bound for irreversible flux</td>
<td>0</td>
</tr>
<tr>
<td>$v_I^U$</td>
<td>upper bound for irreversible flux</td>
<td>1000</td>
</tr>
<tr>
<td>$\lambda_{min}$</td>
<td>lower bound for dual variable</td>
<td>-500</td>
</tr>
<tr>
<td>$\lambda_{max}$</td>
<td>upper bound for dual variable</td>
<td>500</td>
</tr>
<tr>
<td>$(\lambda_R^L)^{min}$</td>
<td>glover transform parameter</td>
<td>-500</td>
</tr>
<tr>
<td>$(\lambda_R^L)^{max}$</td>
<td>glover transform parameter</td>
<td>500</td>
</tr>
<tr>
<td>$(\lambda_R^U)^{min}$</td>
<td>glover transform parameter</td>
<td>-500</td>
</tr>
<tr>
<td>$(\lambda_R^U)^{max}$</td>
<td>glover transform parameter</td>
<td>500</td>
</tr>
<tr>
<td>$(\lambda_I^U)^{min}$</td>
<td>glover transform parameter</td>
<td>-500</td>
</tr>
<tr>
<td>$(\lambda_I^U)^{max}$</td>
<td>glover transform parameter</td>
<td>500</td>
</tr>
</tbody>
</table>

for xylose transport and metabolism have been added to the model. The newly discovered central metabolic reactions described in Nakahigashi et al. (2009) have also been added to the model. Based on the results obtained from characterization of the mutant LMSE₁, glucose-6-phosphate was not found to be a biomass component. The biomass composition in the original model has accordingly been modified.
Appendix C

Primers

Table C.1 lists all the primers used in the studies reported in this thesis. All the long primers used for construction of plasmids are listed separately at the end of the table.

Table C.1: Primers.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primers used in <strong>Chapter 4</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pgi forward</td>
<td>check primer</td>
<td>CTGTGACTGCGCGCTACAATC</td>
</tr>
<tr>
<td>pgi reverse</td>
<td>check primer</td>
<td>CTTATCCGCGCTACATATCGAC</td>
</tr>
<tr>
<td>rpe forward</td>
<td>check primer</td>
<td>CCATCGAACATCTGCGCCGTTACC</td>
</tr>
<tr>
<td>rpe reverse</td>
<td>check primer</td>
<td>CGATACGATGTTAACGCGTGAGTGG</td>
</tr>
<tr>
<td>eda forward</td>
<td>check primer</td>
<td>CTGCAGCCTCTGCGCTTATTCTGC</td>
</tr>
<tr>
<td>eda reverse</td>
<td>check primer</td>
<td>CGTAACACCAGAAGCCTACG</td>
</tr>
<tr>
<td>gnd forward</td>
<td>check primer</td>
<td>CTACGTTTTTGGATGGTCAAAG</td>
</tr>
<tr>
<td>gnd reverse</td>
<td>check primer</td>
<td>CAAATCGCAACTTTTGATCG</td>
</tr>
<tr>
<td>pfkB forward</td>
<td>check primer</td>
<td>GATGAGCAAGCTGTCGAGAAG</td>
</tr>
<tr>
<td>pfkB reverse</td>
<td>check primer</td>
<td>GGATCAAAGATTAGCGTCCCTGG</td>
</tr>
<tr>
<td>fbp forward</td>
<td>check primer</td>
<td>GCCGTAAGCGTGCTGCGTCG</td>
</tr>
<tr>
<td>fbp reverse</td>
<td>check primer</td>
<td>CAATGTGCGCTTTTGATC</td>
</tr>
</tbody>
</table>

*continued to the next page...*
<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tpi forward</td>
<td>check primer</td>
<td>CGTTATTACACCGAGACAGAAGG</td>
</tr>
<tr>
<td>tpi reverse</td>
<td>check primer</td>
<td>GCTGGCAGAAGAAGTACCTGAG</td>
</tr>
<tr>
<td>kt (kanR forward)</td>
<td>check primer</td>
<td>CGGCCACAGTCGATGAAATCC</td>
</tr>
<tr>
<td>k2 (kanR reverse)</td>
<td>check primer</td>
<td>CGGTGCCCTGAATGAACCTGC</td>
</tr>
</tbody>
</table>

**Primers used in Chapter 5**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ptsG forward</td>
<td>sequencing primer</td>
<td>ATAACTTCGCCCGCTCTGTTTC</td>
</tr>
<tr>
<td>ptsG reverse</td>
<td>sequencing primer</td>
<td>GTTCGACAAACACCTACGATACC</td>
</tr>
<tr>
<td>ptsG mid-forward</td>
<td>sequencing primer</td>
<td>GCCTATGGGCATCATGGTTAAAACC</td>
</tr>
<tr>
<td>ptsH forward</td>
<td>sequencing primer</td>
<td>GCTTTACCAACAAGAATATTGTGG</td>
</tr>
<tr>
<td>ptsH reverse</td>
<td>sequencing primer</td>
<td>TTAICTCGAGTCCGGCCCATCAG</td>
</tr>
<tr>
<td>ptsI forward</td>
<td>sequencing primer</td>
<td>AAGACGAGACGAAGAAGCGTTG</td>
</tr>
<tr>
<td>ptsI reverse</td>
<td>sequencing primer</td>
<td>ACCAATGTTGCGTCTACTTGG</td>
</tr>
<tr>
<td>ptsI mid-forward1</td>
<td>sequencing primer</td>
<td>GGAACCTACCTGCATCGTGGGG</td>
</tr>
<tr>
<td>ptsI mid-forward2</td>
<td>sequencing primer</td>
<td>GGAACCTGCAGCAGAAGGTAA</td>
</tr>
<tr>
<td>ptsI mid-reverse</td>
<td>sequencing primer</td>
<td>GATGAACCCACGACCTTTCTTCAG</td>
</tr>
<tr>
<td>crr forward</td>
<td>sequencing primer</td>
<td>ACAACGGACGAGTTAATGACG</td>
</tr>
<tr>
<td>crr reverse</td>
<td>sequencing primer</td>
<td>TTAACCGATGCAGTGCACCGAG</td>
</tr>
<tr>
<td>crp forward</td>
<td>sequencing primer</td>
<td>TGCGRCTTTTCAGGTCGAGRA</td>
</tr>
<tr>
<td>crp reverse</td>
<td>sequencing primer</td>
<td>CAGGCAGGACGAGGGAAGAG</td>
</tr>
<tr>
<td>crp mid-forward</td>
<td>sequencing primer</td>
<td>CGACTCTCGAATGTTTCTTTGTC</td>
</tr>
<tr>
<td>mgsA forward</td>
<td>sequencing primer</td>
<td>TACCCAGCTCATCAACCAGG</td>
</tr>
<tr>
<td>mgsA reverse</td>
<td>sequencing primer</td>
<td>GCCACCGGTTGAGGATTCTCC</td>
</tr>
<tr>
<td>xylF forward</td>
<td>sequencing primer</td>
<td>GAACGCAGTCGAGCTTGGT</td>
</tr>
<tr>
<td>xylF reverse</td>
<td>sequencing primer</td>
<td>CAGCAGTCTCCATCGACGTGAG</td>
</tr>
<tr>
<td>xylG forward</td>
<td>sequencing primer</td>
<td>TCCACAAAGAGAGCGAGCTG</td>
</tr>
<tr>
<td>xylG reverse</td>
<td>sequencing primer</td>
<td>CAGACGGATTGTTTGGAC</td>
</tr>
<tr>
<td>xylG mid-reverse</td>
<td>sequencing primer</td>
<td>CCAGTTGTTGTTGCCCAAGCC</td>
</tr>
<tr>
<td>xylH forward</td>
<td>sequencing primer</td>
<td>CGTGACTCTGTTGATGCAAGGA</td>
</tr>
<tr>
<td>xylH reverse</td>
<td>sequencing primer</td>
<td>GAAAATATCCCATCCGATTG</td>
</tr>
<tr>
<td>xylH mid-reverse</td>
<td>sequencing primer</td>
<td>CACCATCGTTCCAGGAC</td>
</tr>
</tbody>
</table>
## Appendix C. Primers

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>xylR forward</td>
<td>sequencing primer</td>
<td>CGCAACAAACGCCCGTTTC</td>
</tr>
<tr>
<td>xylR reverse</td>
<td>sequencing primer</td>
<td>ATGATTCACAAGCTGAAGGGGC</td>
</tr>
<tr>
<td>xylR mid-forward</td>
<td>sequencing primer</td>
<td>GGATCTCGCGCTCCAATCTT</td>
</tr>
<tr>
<td>xylR mid-reverse</td>
<td>sequencing primer</td>
<td>GGCAATGTAATGAACGGGTGG</td>
</tr>
<tr>
<td>xylE forward</td>
<td>sequencing primer</td>
<td>GCAATCTGAATCGTGGCC</td>
</tr>
<tr>
<td>xylE reverse</td>
<td>sequencing primer</td>
<td>GCATATGGATACCGCCGCC</td>
</tr>
<tr>
<td>xylE mid-forward</td>
<td>sequencing primer</td>
<td>GCAATGTATATCGGGAACCTG</td>
</tr>
<tr>
<td>xylE mid-reverse</td>
<td>sequencing primer</td>
<td>CAGTTCGCAATATACATTGGGC</td>
</tr>
</tbody>
</table>

**Primers used in Chapter 6**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>yidA forward</td>
<td>sequencing primer</td>
<td>ATGGCTATTAACCTCATGGCT</td>
</tr>
<tr>
<td>yidA reverse</td>
<td>sequencing primer</td>
<td>TTAATTCAGCACATACTCTCTCA</td>
</tr>
<tr>
<td>yidA mid-forward</td>
<td>sequencing primer</td>
<td>GCATGAATCCTTCGGCCAC</td>
</tr>
<tr>
<td>yidA mid-reverse</td>
<td>sequencing primer</td>
<td>GTGCAACAGGAAGATTCATGC</td>
</tr>
<tr>
<td>ybiV forward</td>
<td>sequencing primer</td>
<td>ATGAGCGTAAAAGTTATCGT</td>
</tr>
<tr>
<td>ybiV reverse</td>
<td>sequencing primer</td>
<td>TACAGCTGTTAAAAGGGGATGTG</td>
</tr>
<tr>
<td>ybiV mid-forward</td>
<td>sequencing primer</td>
<td>AAAACACTACCATCGGACTGA</td>
</tr>
<tr>
<td>ybiV mid-reverse</td>
<td>sequencing primer</td>
<td>TTCAGGCGATGTTAGTGTTC</td>
</tr>
</tbody>
</table>

**Primers used in Appendix A**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>aceA forward</td>
<td>check primer</td>
<td>CGAAGAAGCTGTTCTCCAGG</td>
</tr>
<tr>
<td>aceA reverse</td>
<td>check primer</td>
<td>CGTAAACGCTTTATCCAGG</td>
</tr>
</tbody>
</table>

**Primers used for plasmid construction**

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Description</th>
<th>Sequence (5’–3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ptsG* EcoRI forward</td>
<td></td>
<td>AGTCATTGAAATTCGGTTGAATGAGTTTTTTAAAAAGC</td>
</tr>
<tr>
<td>ptsG* HindIII reverse</td>
<td></td>
<td>GGATCGTAAAGCTTTTAGTGTTACGGAGTGACTCA</td>
</tr>
<tr>
<td>yidA KpnI forward</td>
<td></td>
<td>AGTCATTGCTACAGGAGTTAAACATATGGCTATTAACCTTCTTCTT</td>
</tr>
<tr>
<td>yidA HindIII reverse</td>
<td></td>
<td>GGATCGTAAAGCTTTAATTTCAGCACATACTTTCTC</td>
</tr>
<tr>
<td>ybiV KpnI forward</td>
<td></td>
<td>AGTCATTGCTACAGGAGTTAAACATATGAGCGTAAAAGTTATCGT</td>
</tr>
<tr>
<td>ybiV HindIII reverse</td>
<td></td>
<td>GGATCGTAAAGCTTTACGCTGTTAAAAGGGGATGTG</td>
</tr>
</tbody>
</table>
Appendix D

Mutants and Plasmids

Table D.1 lists the mutants and the plasmids constructed in this thesis. All the mutants are listed first, followed by the list of the plasmids.

The naming convention for the LMSE-series mutants was decided by the consensus of the research group to standardize the naming practices. The convention changed multiple times during the course of this thesis. Originally, the mutants were labeled as LMSE1, LMSE2, et cetera, which was changed to the names with subscripts, such as LMSE_{1}, LMSE_{2}, et cetera (used throughout this thesis). The currently followed convention uses the names LMSE-001, LMSE-002, et cetera.

Table D.1: Mutants and plasmids used in this thesis.

<table>
<thead>
<tr>
<th>Mutant/Plasmid</th>
<th>Description</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>LMSE(_1) (LMSE-001)</td>
<td>\textit{E. coli} (\Delta\text{eda} \ \Delta\text{pgi} \ \Delta\text{gnd}) :kanR</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>LMSE(_2) (LMSE-002)</td>
<td>\textit{E. coli} (\Delta\text{pgi} \ \Delta\text{eda} \ \Delta\text{rpe}) :kanR</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>LMSE(_3) (LMSE-003)</td>
<td>\textit{E. coli} (\Delta\text{pgi} \ \Delta\text{gnd} \ \Delta\text{fbaB}) :kanR</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>LMSE(_4) (LMSE-004)</td>
<td>\textit{E. coli} (\Delta\text{pgi} \ \Delta\text{gnd} \ \Delta\text{tpi}) :kanR</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>LMSE(_5) (LMSE-005)</td>
<td>\textit{E. coli} (\Delta\text{pgi} \ \Delta\text{gnd} \ \Delta\text{pfkA} \ \Delta\text{pfkB} \ \Delta\text{fbp}) :kanR</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>SQ-001 (LMSE-006)</td>
<td>\textit{E. coli} (\Delta\text{ptsG})</td>
<td>Chapter 4</td>
</tr>
<tr>
<td>SQ-002</td>
<td>LMSE-006 (\text{pBT4.ptsG}^*)</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>SQ-003 (LMSE-007)</td>
<td>\textit{E. coli} (\Delta\text{icdA})</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>SQ-004</td>
<td>LMSE-007 (\text{pBT4.icdA}^*)</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>RB-001</td>
<td>\textit{E. coli} (\text{pTrc99A})</td>
<td>Chapter 6</td>
</tr>
</tbody>
</table>

\textit{continued to the next page...}
<table>
<thead>
<tr>
<th>Mutant/Plasmid</th>
<th>Description</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB-002</td>
<td><em>E. coli</em> pTrc99A.yidA</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>RB-003</td>
<td><em>E. coli</em> pTrc99A.ybiV</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>RB-004</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt; pTrc99A</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>RB-005</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt; pTrc99A.yidA</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>RB-006</td>
<td>LMSE&lt;sub&gt;2&lt;/sub&gt; pTrc99A.ybiV</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>LMSE-008</td>
<td><em>E. coli</em> Δpgi</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-009</td>
<td><em>E. coli</em> Δeda</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-010</td>
<td><em>E. coli</em> ΔaceA</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LP-001 (LMSE-011)</td>
<td><em>E. coli</em> ΔaceA Δpgi</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LP-002 (LMSE-011R)</td>
<td><em>E. coli</em> Δpgi ΔaceA</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LP-003 (LMSE-012)</td>
<td><em>E. coli</em> Δeda Δpgi</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LP-004 (LMSE-012R)</td>
<td><em>E. coli</em> Δpgi Δeda</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-009K</td>
<td><em>E. coli</em> Δeda::kanR</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-013</td>
<td><em>E. coli</em> Δgnd</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-014</td>
<td><em>E. coli</em> Δgnd Δpgi</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-015</td>
<td><em>E. coli</em> Δgnd Δeda</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-016K</td>
<td><em>E. coli</em> Δeda ΔaceA::kanR</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-017</td>
<td><em>E. coli</em> Δpgi Δgnd ΔfsaB</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-018</td>
<td><em>E. coli</em> Δpgi Δgnd Δfbp ΔpfkA</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-018K</td>
<td><em>E. coli</em> Δpgi Δgnd Δfbp ΔpfkA::kanR</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-019</td>
<td><em>E. coli</em> Δpgi Δgnd Δfbp</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-019K</td>
<td><em>E. coli</em> Δpgi Δgnd Δfbp::kanR</td>
<td>Appendix A</td>
</tr>
<tr>
<td>LMSE-020K</td>
<td><em>E. coli</em> Δpgi Δeda ΔaceA::kanR</td>
<td>Appendix A</td>
</tr>
</tbody>
</table>

**Plasmids**

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>pBT4</td>
<td>Empty vector</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>pTrc99A</td>
<td>Empty vector</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>pBZE-001</td>
<td>pBT4 expressing <em>ptsG</em> from <em>E. coli</em> wild-type</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>pBZE-002</td>
<td>pBT4 expressing <em>ptsG</em> from the mutant LMSE&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>pBZE-003</td>
<td>pBT4 expressing <em>icdA</em> from <em>E. coli</em> wild-type</td>
<td>Chapter 5</td>
</tr>
</tbody>
</table>

*continued to the next page...*
<table>
<thead>
<tr>
<th>Mutant/Plasmid</th>
<th>Description</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>pBZE-004</td>
<td>pBT4 expressing <em>icdA</em> from the mutant LMSE$_2$</td>
<td>Chapter 5</td>
</tr>
<tr>
<td>pBZE-005</td>
<td>pTrc99A expressing <em>yidA</em> from <em>E. coli</em> wild-type</td>
<td>Chapter 6</td>
</tr>
<tr>
<td>pBZE-006</td>
<td>pTrc99A expressing <em>ybiV</em> from <em>E. coli</em> wild-type.</td>
<td>Chapter 6</td>
</tr>
</tbody>
</table>
Appendix E

Minimal Medium Composition

The minimal medium composition used throughout this thesis was adopted from Causey et al. (2003). The minimal medium contained, per L: 3.5 g of KH$_2$PO$_4$, 5.0 g of K$_2$HPO$_4$, 3.5 g of (NH$_4$)$_2$HPO$_4$, 0.25 g of MgSO$_4$.7H$_2$O, 15 mg of CaCl$_2$.2H$_2$O, 0.5 mg of thiamine, and 1 mL of trace metal stock. The trace metal stock was prepared in 0.1 M HCl and consisted of, per L: 1.6 g of FeCl$_3$, 0.2 g of CoCl$_2$.6H$_2$O, 0.1 g of CuCl$_2$, 0.2 g of ZnCl$_2$.4H$_2$O, 0.2 g of NaMoO$_4$, and 0.05 g of H$_3$BO$_3$. Appropriate sugar concentration was used depending upon the mutant being cultured. 4-Morpholinopropanesulfonic acid (MOPS) (0.1 M) was added to the medium in flasks, where pH could not be controlled by base addition. Appropriate concentrations of antibiotics were added as required.

Sugar stock solution, MgSO$_4$.7H$_2$O stock solution (250 g/L), and CaCl$_2$.2H$_2$O stock solution (15 g/L) were autoclaved separately, and added to the appropriate volume of phosphate salts solution to obtain the desired concentrations. Trace metal solution, thiamine stock solution (500 mg/L), MOPS stock solution (1 M), and antibiotics stocks were sterilized separately by passing through 0.22 μm filters, and added to obtain the desired concentrations in the final medium.
Appendix F

Bioinformatics Pipeline

The bioinformatics analysis on the whole genome sequencing results was carried out by the McGill University and Génome Québec Innovation Center. The pipeline was executed on Compute Canada clusters via Unix bash commands, Perl scripts and open source software. The following document, provided by the McGill University and Génome Québec Innovation Center, lists the steps used in the genome sequencing analysis pipeline.

1. **Reads**: Around 50 to 150 million 100 base pair paired-end reads from the Illumina HiSeq 2000 sequencer were obtained. The base quality was encoded in Phred quality score.

2. **Read trimming and clipping**: Reads were trimmed from the 3’-end to have a Phred score of at least 30. Illumina sequencing adapters were removed from the reads, and all reads were required to have a length of at least 32 bps. Trimming and clipping were done with the Fastx software (http://hannonlab.cshl.edu/fastx_toolkit/).

3. **Aligning the reads**: The filtered reads were aligned to the reference genome of *E. coli* MG1655. The alignment was done per lane of sequencing, and then merged for a complete Binary Alignment Map file (.bam). The alignment software used was BWA (http://bio-bwa.sourceforge.net/), and the merging was done with the Picard software (http://picard.sourceforge.net/).

4. **Realigning insertions and deletions**: Insertion and deletion realignment was performed on regions where multiple base mismatches were preferred over indels by the aligner (since it can appear to be less costly by the algorithm). Such regions would introduce false positive variant calls which could be filtered out by realigning those regions properly. Realignment was done with the GATK software (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit).
5. **Fixing the read-mates**: Once the local regions were realigned, the read mate coordinates of the aligned reads were recalculated, since the reads are realigned at positions that differ from their original alignment. Fixing the read mate positions was done with Picard software.

6. **Marking duplicates**: Aligned reads are duplicates if they have the same 5' alignment positions (for both mates in the case of paired-end reads). All but the best pair (based on alignment score) were marked as a duplicate in the .bam file. Marking duplicates was done with Picard software.

7. **Generate metrics and coverage track**: Multiple metrics including,

   (a) number of raw reads,
   (b) number of filtered reads (after Step 2),
   (c) number of aligned reads (after Step 3),
   (d) number of duplicate reads (after Step 6),
   (e) duplicate rate (number of duplicate reads / number of raw reads. Good run max of 25%),
   (f) median, mean, and standard deviation of insert sizes of reads after alignment,
   (g) mean coverage (mean number of reads per base position),
   (h) percentage of bases covered at X reads (%_bases_above_50 means the % of bases which have at least 50 reads. A good run is typically around 50%),

were generated. A TDF (.tdf) coverage track was also generated at this step for easy visualization of coverage in the IGV browser.

8. **Variant calling**: Variants (SNPs and INDELs) were called using samtools mpileup and bcftools varfilter. The following options are given to mpileup to filter for low quality variants which could introduce false positive calls: 
   ```
   -L 1000 -E -q 1 -u -D -S
   ```
   where:

   - `-L INT` max per-sample depth for INDEL calling [250]
   - `-E` extended BAQ for higher sensitivity but lower specificity
   - `-q INT` skip alignments with mapQ smaller than INT [0]
   - `-u` generate uncompress BCF output
   - `-D` output per-sample DP in BCF (require -g/-u)
   - `-S` output per-sample strand bias P-value in BCF (require -g/-u)

   The output of mpileup was then fed to varfilter, which did an additional filtering of the variants and transformed the output into the VCF (.vcf) format. The arguments used were: `-d 2 -D 1200`
-Q 15 -1 0.0, where:

- **-d INT** minimum read depth [2]
- **-D INT** maximum read depth [10000000]
- **-Q INT** minimum RMS mapping quality for SNPs [10]
- **-1 FLOAT** min P-value for strand bias (given PV4) [0.0001]

The final .vcf files were filtered for long ‘N’ INDELs which are sometimes introduced and cause excessive memory usage by downstream tools.

9. **Mappability annotation**: An in-house database identified the regions in which reads were confidently mapped to the reference genome. Generally, low mappability corresponds nicely to RepeatMasker regions and increases substantially with read length. A region was identified as HC = coverage too high (>400), LC = low coverage (<50), MQ = to low mean mapQ (<20) and ND = no data at the position.

10. **dbSNP annotation**: The .vcf files were annotated for dbSNP using the software SnpSift (http://snpeff.sourceforge.net/SnpSift.html).

11. **Variant effect annotation**: The .vcf files were annotated for variant effects using the SnpEff software (http://snpeff.sourceforge.net/SnpSift.html). SnpEff annotates and predicts the effects of variants on genes (such as amino acid changes).

12. **Additional SVN annotations**: Additional information about SVN was obtained using numerous published databases.

   (a) Biomart: adds GO annotations based on gene information.

   (b) dbNSFP: an integrated database of functional annotations from multiple sources for the comprehensive collection of human non-synonymous SNPs. It compiles prediction scores from four prediction algorithms (SIFT, Polyphen2, LRT, and MutationTaster), three conservation scores (PhyloP, GERP++, and SiPhy) and other function annotations.

   (c) Cosmic (Catalogue of Somatic Mutations in Cancer): Annotates SVN which are known somatic mutations.
Appendix G

Coverage of the Deleted Regions in the Mutant LMSE$_2$

Figures G.1 and G.2 show the coverage of the regions corresponding to the genes $pgi$ and $rpe$, respectively, in the mutant LMSE$_2$ (sequenced), the mutant LMSE$_2$ (unsequenced), and the wild-type $E. coli$. The figures were provided by the Bioinformatics Facility of the McGill University and Génome Québec Innovation Center. The figures clearly show a low, non-continuous, and/or incomplete coverage of the genes $pgi$ and $rpe$ in both the variants of the mutant LMSE$_2$. 
Appendix G. Coverage of the Deleted Regions in the Mutant LMSE$_2$

Figure G.1: Coverage of the region encoding the gene pgf in the mutant LMSE$_2$ and the wild-type.
Figure G.2: Coverage of the region encoding the gene rpe in the mutant LMSE$_2$ and the wild-type.
Bibliography


Negrete, A., Ng, W., and Shiloach, J. (2010). Glucose uptake regulation in *E. coli* by the small RNA SgrS: comparative analysis of *E. coli* K-12 (JM109 and MG1655) and *E. coli* B (BL21). *Microbial Cell Factories*, 9(75).


