OPTIMIZATION-BASED FRAMEWORKS FOR THE
ANALYSIS AND REDESIGN OF METABOLIC NETWORKS

A Thesis in
Chemical Engineering

by

Anthony P. Burgard

© 2004 Anthony P. Burgard

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2004
The thesis of Anthony P. Burgard was reviewed and approved* by the following:

Costas D. Maranas  
Associate Professor of Chemical Engineering  
Thesis Advisor  
Chair of Committee

Darrell Velegol  
Assistant Professor of Chemical Engineering

Reka Albert  
Assistant Professor of Physics

Andrew L. Zydney  
Interim Department Head  
Endowed Bio-Chair and Professor of Chemical Engineering

* Signatures are on file in the Graduate School.
Abstract

The central theme of this thesis is the development of computational tools for the analysis and redesign of metabolic networks. The key questions that are addressed in this work are: how can one (i) select the optimal set(s) of functions to recombine into microbial strains to enhance their maximum theoretical production capabilities, (2) identify the smallest set of reactions capable of supporting minimal metabolism under various conditions, (3) test whether hypothesized intracellular objectives are consistent with experimental measurements, (4) shape the connectivity of a metabolic network in such a way that a targeted biochemical becomes an obligatory byproduct of cell growth, and (5) elucidate the topological and structural features of metabolic reconstructions at the genome-scale.

In the first part of this thesis, a computational protocol is introduced and applied to select the optimal foreign reactions from a genomic database encompassing multiple species for increasing the maximum theoretical yields of the twenty amino acids in *Escherichia coli*. Here it was found that the production capabilities of seven amino acids could be enhanced by the addition of only one or two foreign functionalities. Next, the minimal reaction network framework revealed that 224 or 229 metabolic reactions are required to support *E. coli* growth on glucose or acetate, respectively, while a cell cultured on a rich optimally engineered medium could theoretically support growth with as few as 122 metabolic reactions. Thus the minimal reaction sets were revealed to be highly dependent on the imposed uptake environment and the growth requirements. In
the following section, the ObjFind procedure is applied to two sets of experimental \textit{E. coli} flux data (i.e., anaerobic and aerobic growth) revealing that a common metabolic objective, biomass production, is the most explanatory. The OptKnock framework is then introduced for pinpointing reactions in a metabolic network for removal to force growth-coupled biochemical production. Computational results for the overproduction of lactic acid, succinic acid, and 1,3-propanediol are consistent with both intuitive and non-intuitive knockout strategies published in the literature. Lastly, the flux coupling finder procedure is applied to three genome-scale metabolic models enabling the exhaustive identification of various types of reaction coupling.
# Table of Contents

List of Figures vii  
List of Tables x  
Acknowledgements xi  
1 Introduction 1  
  1.1 Literature Review  
  1.2 Stoichiometric Modeling  
  1.3 *Escherichia coli* Stoichiometric Models  
  1.4 Thesis Overview  
2 Augmenting Metabolic Networks with Non-native Functionalities 15  
  2.1 Background  
  2.2 Modeling and Computational Protocol  
  2.3 Amino Acid Maximum Theoretical Yield Optimizations  
  2.4 Conclusions  
3 Identifying Minimal Metabolic Reactions Sets for Cell Growth 31  
  3.1 Background  
  3.2 Modeling and Computational Protocol  
  3.3 Minimal Reaction Sets under Various Uptake Conditions  
  3.4 Conclusions  
4 Inferring and Testing Hypothesized Metabolic Imperatives 50  
  4.1 Background
4.2 The ObjFind Bilevel Optimization Framework
4.3 Aerobic and Anaerobic *Escherichia coli* Metabolic Objectives
4.4 Conclusions

5 Gene Knockout Strategies for Microbial Strain Optimization  70

5.1 Background
5.2 The OptKnock Bilevel Optimization Framework
5.3 Succinate and Lactate Knockouts Strategies
5.4 1,3-Propanediol Knockout Strategies
5.5 Minimization of Metabolic Adjustment Comparison
5.6 Conclusions

6 Flux Coupling Analysis of Genome-Scale Metabolic Reconstructions  96

6.1 Background
6.2 Exhaustive Identification of Blocked and Coupled Reactions
6.3 Flux Coupling Analysis Algorithm
6.4 Blocked Reaction Identification
6.5 Coupled Reaction Sets
6.6 Directional Coupling of Metabolic Reactions
6.7 Conclusions
6.8 APPENDIX: Flux Ratio Maximization/Minimization Transformation Proof

7 Synopsis  135

Bibliography  139
List of Figures

1.1 Given the stoichiometry of the reaction network, cellular composition information, and the uptake rate of substrate, an optimization model is formulated to predict the ideal intracellular flux distribution and cellular growth rate. (gDW = grams of dry weight of cells) ............... 14

2.1 Optimal wild-type *E. coli* and recombinant arginine production pathways for growth on glucose. .................................................. 26

2.2 Optimal wild-type *E. coli* and recombinant arginine production pathways for growth on acetate .......................................................... 27

2.3 Optimal asparagine production pathways in *E. coli* for two modes of glucose utilization: glucokinase and the phosphotransferase system .... 28

2.4 Optimal recombinant asparagine production pathway for glucose growth. 29

2.5 Optimal wild-type *E. coli* and recombinant histidine production pathways for growth on acetate .......................................................... 30

3.1 Number of reactions in each minimal set as a function of the imposed growth demands for (A) a glucose or acetate uptake environment and (B, C) an uptake environment containing multiple organic substrates .... 49

4.1 Pictorial depiction of weak and strong duality ........................................ 65

4.2 Reaction fluxes allowed to assumed nonzero coefficients are shown in color. The magnitudes of the CoIs are similar for both aerobic and anaerobic growth conditions. .................................................. 66

4.3 The values of the coefficients of importance for the aerobic and anaerobic experimental flux distributions ................................. 67

4.4 The maximum allowable increase of the coefficients of importance for the (a) aerobic and (b) anaerobic experimental flux distributions when solution optimality is relaxed. MIN = minimum. ...................... 68

4.5 The values of the coefficients of importance for the aerobic and anaerobic experimental flux distributions with an artificial biomass flux. . 69
5.1 The bilevel optimization structure of OptKnock ........................................ 90

5.2 The flux distributions of the (A) wild-type E. coli, (B) succinate mutant B, (C) succinate mutant C, and (D) lactate mutant C networks that maximize biomass yield under anaerobic conditions. ............................. 91

5.3 (A) Succinate or (B) lactate production limits under anaerobic conditions for mutant A, mutant B, mutant C and the wild-type E. coli network ......................................................... 92

5.4 The aerobic flux distributions of the (A) wild-type E. coli, (B) mutant A, and (C) mutant B networks that maximize biomass yield. Results for mutants A and B assume the reactions responsible for 1,3-propanediol production are available .................................................. 93

5.5 1,3-propanediol (PDO) production limits under aerobic conditions for mutant A, mutant B, and the wild-type E. coli network ................. 94

5.6 Projection of the multidimensional flux space onto two dimensions .... 95

6.1 Directionally, partially, fully, and un-coupled reaction pairs ............... 124

6.2 Examples of blocked reactions (dashed lines) and a fully coupled enzyme subset (heavy lines) .......................................................... 125

6.3 Examples of affected reaction sets and equivalent knockouts for reaction \( v^* \) ................................................................. 126

6.4 Total numbers and percentages of blocked reactions for the H. pylori, E. coli, and S. cerevisiae networks under different growth conditions ....... 127

6.5 Percentage of reactions contained in coupled sets in the H. pylori, E. coli, and S. cerevisiae metabolic networks for growth on either a complex or glucose minimal media (with and without a biomass reaction). ..... 128

6.6 Coupled reaction set identified for purine biosynthesis in E. coli on a glucose minimal media assuming a constant biomass composition .... 129

6.7 Reactions coupled to biomass formation for aerobic S. cerevisiae growth on a glucose minimal media. This enzyme subset is decomposed into numerous subsystems, indicated by different colored arrows, if the biomass reaction is replaced with drains on the various precursors .... 130

6.8 Comparison of the FCF-identified coupled reaction sets for H. pylori with the enzyme subsets identified by Schilling et al. (2002) .......... 131
6.9 The complete reaction coupling relationships in *E. coli* central metabolism for aerobic growth on glucose. .......................... 132

6.10 The number of reactions N(k) implying k other reactions are plotted as a function of k for *H. pylori*, *E. coli*, and *S. cerevisiae* growth on a glucose minimal media. .......................... 133

6.11 Genome-wide coupling for *E. coli* growth on a glucose minimal medium (a) with or (b) without the presence of a biomass reaction. ................. 134
List of Tables

2.1 Model predictions of maximum theoretical yields of amino acids for growth on glucose and acetate ........................................ 24
2.2 Model selections of enzymatic reactions that will enhance the amino acid production capabilities of Escherichia coli ......................... 25
3.1 Evolution of minimal reaction sets for case (i) under decreasing growth requirements. ......................................................... 44
3.2 Metabolites uptaken or secreted at each target growth rate on an optimally engineered medium. ............................................. 45
3.3 Evolution of minimal reaction sets for case (ii) under decreasing growth requirements. ......................................................... 46
3.4 Functional classification of minimal network reactions for growth on an optimally engineered medium. ....................................... 47
3.5 Comparison of minimal metabolic gene/reaction sets based on functional classification. .......................................................... 48
4.1 Experimentally (Schmidt et al. 1999a) determined flux distributions for aerobic and anaerobic conditions. ................................. 64
5.1 Biomass and chemical yields for various gene knockout strategies identified by OptKnock. The reactions and corresponding enzymes for each knockout strategy are listed. ........................................ 89
6.1 Reaction coupling statistics for the H. pylori, E. coli, and S. cerevisiae metabolic networks. ......................................................... 123
Acknowledgements

It is simply not possible to overstate my gratitude towards my Ph.D. thesis advisor, Costas Maranas. In addition to his constant guidance throughout the course of this research, there was never any question that Dr. Maranas always had my best interests at heart. I have learned more from him than any other person and I would not have it any other way. I must surely thank Andrew Zydney, Darrell Velegol, and Reka Albert for serving on my doctoral committee and contributing valuable insights to my thesis work. I would next like to express my sincerest gratitude to all of my past and present group members at Penn State, particularly, Priti Pharkya and Evgeni Nikolaev for their stimulating discussions and contributions to many of the ideas discussed in this thesis. I would also like to express thanks to Greg Moore, Anshuman Gupta, and Shankar Vaidyaraman for exemplifying how successful graduate students conduct research and for their help and patience in getting me started. Speaking of patience, I would certainly like to thank my wife, Alysa, for her encouragement, love, and understanding during my studies. Finally, I would like to dedicate this thesis to my parents, Anita and George, for their emotional, financial, but most importantly, unwavering support throughout my life. There are not two better people in the world and there is no one that I respect more.
Chapter 1

Introduction

1.1 Literature Review

The analysis and modification of metabolic pathways, known as metabolic engineering, has attracted significant interest in recent years catalyzed by the rapidly increasing number of sequenced microbial genomes. As of September 2004, 167 prokaryotic and 32 eukaryotic genomes have been completely sequenced, while another 523 prokaryotic and 441 eukaryotic sequencing projects are ongoing (Bernal et al. 2001; Liolios et al. 2004). This flood of genomic information coupled with gene annotation and metabolic reconstruction efforts is poised to revolutionize the understanding of microbial metabolic networks of biotechnological and biomedical importance. This research centers upon using this newly acquired information in conjunction with recent models of microbial metabolism to develop algorithmic and discrete optimization techniques for enhancing their predictive capabilities. Broadly, metabolism is defined as the totality of chemical reactions occurring in an organism. Each reaction is catalyzed by one or more enzymes coded for by one or more genes in an organism’s DNA. Metabolic models are particularly useful for investigating how raw materials (i.e., glucose, oxygen, etc.) are converted to products (i.e., desired biochemicals or cellular biomass precursors – amino acids, nucleotides, lipids, etc.) in individual cells. The particular path by which a raw material is converted to a product is typically described in terms of metabolic fluxes
which are defined as the amount of a particular metabolite (i.e., cellular compound) consumed by a given reaction per unit time.

Mathematical models of cellular metabolism fall into two distinct categories, ones that incorporate kinetic and regulatory information and others that include only the stoichiometry of the reaction pathways. The first class of models matches cellular behavior at an original steady state and then employs kinetic and regulatory relations to examine how the cell behaves away from this steady state in the presence of small perturbations brought about by environmental changes or enzyme engineering. Metabolic control analysis (Kacser and Burns 1973; Heinrich and Rapoport 1974) (MCA) introduced in the 1970’s, enabled the systematic evaluation and description of many metabolic factors governing the control of flux such as enzyme activities and effector concentrations. For example, the application of an MCA based (log)linear kinetic model (Hatzimanikatis and Bailey 1997) within a metabolic optimization framework pointed out experimentally verified ways for enhancing ethanol production in metabolically engineered \textit{E. coli} cells (Hatzimanikatis et al. 1996a; Hatzimanikatis et al. 1998).

Alternatively, synergetic or S-Systems representation models based on the Biochemical Systems Theory formalism (Savageau 1969a; Savageau 1969b; Savageau 1970) introduced in the late 1960s cast each metabolite concentration balance as two competing power-law functions describing aggregation and consumption. Upon a logarithmic transformation, this provided the framework for steady-state enzyme level optimization through linear programming (Voit 1992; Regan et al. 1993; Torres et al. 1996). Nonlinear kinetic models (e.g., Michaelis-Menten) have been used in place of the linear MCA or S-systems models whenever detailed kinetic expressions for each reaction step are known.
Both linear and nonlinear kinetic modeling approaches lend well to optimization strategies for exploring possible changes in enzymes activities and/or regulatory structure that optimize a given metabolic objective (Hatzimanikatis et al. 1996a; Hatzimanikatis et al. 1996b).

The key shortcoming of kinetics-based models, however, is that they require rare detailed kinetic and concentration information about numerous enzymes and cofactors which is difficult to obtain. Thus kinetic models have so far been largely limited to restricted parts of a metabolic network, such as glycolysis (Rizzi et al. 1997), the pentose phosphate pathway (Vaseghi et al. 1999), central nitrogen metabolism (van Riel et al. 1998) in *Saccharomyces cerevisiae*, and the penicillin biosynthetic pathway in *Penicillium chrysogenum* (Nielsen and Jorgensen 1996; Pissara et al. 1996; Theilgaard and Nielsen 1999), to name a few, due to the lack of kinetic parameters for complete metabolic networks (Gombert and Nielsen 2000). In fact, with the exception of the human red blood cell (Joshi and Palsson 1989a; Joshi and Palsson 1989b; Joshi and Palsson 1990b; Joshi and Palsson 1990a; Jamshidi et al. 2001), we still lack the necessary information to describe the metabolism of a single cell in mathematical detail despite the recent explosion of generated biological data.

The second class of models, on the other hand, utilizes only the stoichiometric mass balances of a steady-state metabolic network to generate the broadest set of flux distributions potentially available to the cell. By requiring only the stoichiometry of biochemical pathways and cellular composition information, stoichiometric or flux balance analysis (FBA) models can be used to construct boundaries for the metabolic flux distributions in the absence of detailed kinetic and thermodynamic data. In other words,
instead of predicting exactly how a metabolic network behaves, these models narrow the range of possible phenotypes (i.e., cellular behaviors) these systems can display. This versatility, however, comes at the expense of perhaps suggesting flux distributions that are unattainable due to kinetic or regulatory constraints. Therefore, FBA predictions must be treated as theoretical limits to the performance of metabolic networks. Stoichiometric models have been widely used to successfully study many metabolic networks including Penicillin production by *Penicillium chrysogenum* (Jorgensen et al. 1995; Henriksen et al. 1996), growth of *Corynebacterium melassecola* on glucose and fructose (Pons et al. 1996), and to improve large-scale cultivation of mammalian cells (Xie and Wang 1994a; Xie and Wang 1994e; Xie and Wang 1994d; Xie and Wang 1994b; Xie and Wang 1994c; Xie and Wang 1997). In addition, the analysis of metabolic flux boundaries has provided valuable physiological insight through the determination of maximum theoretical yields (Varma et al. 1993a) and the identification of key principle nodes controlling flux split ratios around branch points (Vallino and Stephanopoulos 1994a; Vallino and Stephanopoulos 1994b).

A brief description of the mathematics behind the stoichiometric modeling approach is provided next followed by a summary of the existing stoichiometric models of *E. coli* metabolism used as a benchmark system for frameworks developed in this research. The subsequent section will provide an overview to the remaining chapters of the thesis where optimization-based frameworks are introduced to tackle a wide variety of modeling challenges encountered in the analysis and redesign of metabolic networks.
1.2 Stoichiometric Modeling

Flux balance analysis or stoichiometric modeling requires only the stoichiometry of biochemical pathways and cellular composition information to identify boundaries for the flux distributions available to the cell. Although microorganisms have evolved highly complex control structures which eventually collapse these available boundaries into single points, stoichiometric models are still valuable in setting upper bounds for performance targets and in identifying “ideal” flux distributions. The underlying principle of FBA is mass balances on the metabolites of interest. For a metabolic network comprised of $N$ metabolites and $M$ metabolic reactions we have,

$$\sum_{j=1}^{M} S_{ij} v_j = b_i, \quad \forall i$$

where $S_{ij}$ is the stoichiometric coefficient of metabolite $i$ in reaction $j$, $v_j$ represents the flux of reaction $j$, and $b_i$ (i.e., $dC_i/dt$) quantifies the network’s uptake (if negative) or secretion (if positive) of metabolite $i$. For all internal metabolites, $b_i$ is zero. Reversible reactions are defined simply as two irreversible reactions in opposite directions, constraining all fluxes to positive values.

Typically, the resulting flux balance system of equations is underdetermined as the number of reactions exceeds the number of metabolites, and additional information is required to solve for the reaction fluxes. Several researchers have measured external fluxes (i.e., rates of glucose uptake, CO$_2$ evolution, acetate secretion, etc.) to add as constraints to their underdetermined models, rendering them completely determined or over-determined (Papoutsakis and Meyer 1985a; Papoutsakis and Meyer 1985b; Vallino and Stephanopoulos 1993; Jorgensen et al. 1995; Pons et al. 1996). However, additional
assumptions such as removing reaction pathways are often needed before external flux measurements can completely define a system, and neglecting potentially active pathways to render a system completely defined may cause large changes in calculated fluxes (Pramanik and Keasling 1997). A popular technique for investigating metabolic flux distributions is linear optimization (Varma and Palsson 1994a). The key conjecture is that the cell is capable of spanning all flux combinations allowable by the stoichiometric constraints and thus achieving any flux distributions that maximize a given metabolic objective. Furthermore, recent experimental evidence suggests that organisms have developed control structures to ensure optimal growth (i.e., maximum biomass production) in response to certain environmental conditions (Edwards and Palsson 2000a; Edwards et al. 2001). The general flux balance analysis model for maximizing biomass production is:

\[
\text{Maximize } Z = v_{\text{biomass}} \\
\text{subject to } \sum_{j=1}^{M} S_{ij}v_j = b_i \quad \forall \ i \\
\quad b_i \in \mathbb{R} \quad \forall \ i \\
\quad v_j \in \mathbb{R}^+ \quad \forall \ j
\]

where \(v_{\text{biomass}}\) is a flux drain comprised of all necessary components of biomass (i.e., amino acids, nucleotides, etc.) in their appropriate biological ratios (Neidhardt and Curtiss 1996). A variety of such cellular objectives has been proposed such as the maximization of ATP production (Majewski and Domach 1990; Ramakrishna et al. 2001), minimization of nutrient uptake (Savinell and Palsson 1992), minimization of metabolic adjustment (MOMA) (Segre et al. 2002), and the maximization of the cellular biomass yield (Varma and Palsson 1993b; Varma and Palsson 1994a). Stoichiometric
models of *Escherichia coli* metabolism utilizing the maximization of biomass objective have under certain conditions been successful at (i) the qualitative prediction of the outcomes of gene knockout experiments (Edwards and Palsson 2000a; Badarinarayana et al. 2001), (ii) the correct sequence of byproduct secretion under increasingly anaerobic conditions (Varma et al. 1993b), and (iii) the quantitative prediction of cellular growth rates (Edwards et al. 2001). Interestingly, recent work suggests that even when FBA predictions under the biomass maximization assumption fail, metabolic networks can be evolved towards maximum growth (i.e., biomass yield) through adaptive evolution (Ibarra et al. 2002). A pictorial representation of the flux balance analysis modeling approach is provided in Figure 1.1.

1.3 *Escherichia coli* stoichiometric models

The complete sequencing of the *E. coli* genome (Blattner et al. 1997) makes it a model organism for this research because extensive knowledge regarding its biochemical pathways is readily available. The first detailed stoichiometric *E. coli* model capable of predicting experimental observations included 95 reversible reactions utilizing 107 metabolites for simulating glucose catabolism and macromolecule biosynthesis (Varma and Palsson 1993b). This model was used to investigate byproduct secretion of *E. coli* at increasingly anaerobic conditions and was able to predict the correct sequence of byproduct secretion consistent with experimental findings: first acetate at slightly anaerobic conditions, then formate, and finally ethanol at highly anaerobic conditions (Varma et al. 1993b). Building on the previous model, Pramanik and Keasling (1997) introduced a model that incorporated 126 reversible reactions (including 12 reversible transport reactions) and 174 irreversible reactions, as well as 289 metabolites. Pramanik
and Keasling correlated the macromolecule composition of *E. coli* as a function of growth rate, and verified their model with experimental data. The model successfully predicted several levels of genetic control such as the glyoxylate shunt closing for growth on glucose and the PEP carboxykinase flux tending towards oxaloacetate. Furthermore, the glyoxylate shunt was active during growth on acetate while the flux through PEP carboxykinase was toward phosphoenolpyruvate.

The most recently published *E. coli* model includes the metabolic enzymes of 695 genes identified either directly from open reading frame assignments (650 of the 695 genes) or indirectly based on biochemical data suggesting the existence of these gene products (Edwards and Palsson 2000a). This flux balance analysis model incorporates 454 metabolites and 720 reactions including the glycolysis, tricarboxylic acid (TCA) cycle, pentose phosphate pathway (PPP), and respiration pathways along with synthesis routes for the amino acids, nucleotides, and lipids. Remarkably, this model was able to quantitatively predict the relationship between the uptake rate of the primary carbon source (acetate or succinate), the oxygen uptake rate, and the cellular growth rate suggesting the *E. coli* network is indeed optimized for maximum growth under certain conditions (Edwards et al. 2001). In addition, an *in silico* single gene deletion study performed with this model was able to predict the growth characteristics of 68 out of 79 (86 %) *E. coli* mutants (Edwards and Palsson 2000a; Edwards and Palsson 2000b). The accurate growth predictions of all *E. coli* mutants were prevented by three failure modes: (i) failure to incorporate regulatory events, (ii) failure to specify metabolic network demands correctly, and (iii) failure to account for the buildup of toxic intermediates. Current flux balance modeling efforts involving the incorporation of transcriptional
regulatory events have had success in circumventing the first failure mode (Covert et al. 2001b; Covert and Palsson 2002), while the second failure mode is just a matter of fine-tuning the definition of required biomass components (i.e., certain biomass components may not be vital to cellular growth) (Edwards et al. 2002). The third failure mode, however, is still un-addressable by flux balance analysis because this approach avoids the calculation of metabolite concentrations and cannot predict the accumulation of toxic metabolic byproducts caused by genetic alterations.

1.4 Thesis Overview

The following chapters in this thesis are focused on the development of optimization-based approaches aimed at a variety of challenges in metabolic pathway engineering. Chapter 2 introduces a systematic method for examining how the performance limits of a metabolic network can be expanded through the addition of non-native functionalities. Specifically, a selection procedure founded upon mixed-integer programming is introduced for choosing the optimal foreign reactions to incorporate into the *E. coli* metabolic network to enhance the maximum yield capabilities of the twenty amino acids. A universal stoichiometric model, consisting of 3,400 reactions catalogued from various non-*E. coli* species, is constructed and integrated into a flux balance analysis (FBA) model of *E. coli* metabolism. The approach utilizes binary 0-1 variables that attain a value of 1 if the foreign reaction is selected, and zero otherwise. Genetically, the selected non-native reactions could be added into the *E. coli* network via recombination techniques that introduce the foreign gene(s) responsible for encoding the appropriate enzyme(s). The chapter highlights the key results presented in:
Chapter 3 makes use of the discrete modeling approach to identify the smallest required sets of growth-sustaining metabolic reactions under a variety of environmental conditions. Here the binary variables act as switches to turn the various *E. coli* metabolic reactions on or off. The problem of minimizing the number of active metabolic reactions required to maintain various biomass formation rates is shown to assume the mathematical structure of a generalized network flow problem where the nodes denote metabolites and the arcs represent reactions. The minimal reaction networks are derived under two contrasting uptake environments: (i) restricting the uptake of organic material to a single component and (ii) allowing the uptake of any organic metabolite with a corresponding transport mechanism into the cell. These two extremes ensure the maximum and minimum reliance on internal metabolism for the generation of all biomass precursors. The results provide insight into the questions of what is minimal metabolism and how is it shaped by the environment. The complete description of the work can also be found in:


Chapter 4 attempts to provide perspective into the question: has cellular behavior been programmed through evolution to optimize a particular objective (e.g., growth, ATP production)? Although the existence of a fitness function driving an organism’s evolution is widely accepted, whether or not its fingerprint is evident in the flux distributions of
primary metabolism has been the subject of constant scientific debate throughout the course of this thesis work. In this chapter, the bilevel optimization formulation ObjFind is introduced for examining whether the maximization of a weighted combination of fluxes could explain a set of experimentally measured metabolic flux data. A solution strategy founded on linear programming duality theory is invoked and the framework is applied to both an aerobic and anaerobic set of *E. coli* flux data derived from isotopomer analysis. We examine (i) what is the metabolic objective function, if any, that is most consistent with the flux data, (2) is this objective unique, and (3) how do the driving forces governing metabolism adapt to different environmental conditions. The peer-reviewed publication associated with this work is:


The fifth chapter combines the discrete and bilevel programming approaches of the previous chapters into a unified optimization framework for microbial strain design. Specifically, the OptKnock framework is introduced and applied for identifying multiple gene knockout strategies leading to the overproduction of biochemicals in *E. coli*. The approach recognizes that the internal cellular objectives driving metabolism are often in direct competition with biochemical overproduction goals. The selected knockout strategies work to counteract this effect by shaping the connectivity of the metabolic network in such a way that the targeted biochemical becomes an obligatory byproduct of cell growth. Computational results for gene deletions for succinate, lactate, and 1,3-propanediol production are in good agreement with mutant strains published in the
literature. Furthermore, strains designed on the premise of growth-coupled biochemical production will lend well to growth selection/adaptation systems for enhancing product formation. Chapter 5 provides a complete description of the work provided in:


Though not discussed in this thesis, a modified version of OptKnock has been applied to the production of amino acids in *E. coli*:


In Chapter 6, the flux coupling finder (FCF) algorithm is introduced for extracting useful information from the structural connectivity of metabolic networks. The approach circumvents the combinatorial explosion concerns associated with the application of previously developed network analysis methods to genome-scale metabolic models with many hundreds of reactions. Specifically, a series of linear programming problems is solved to obtain the maximum and minimum ratios between every pair of metabolic fluxes in the network. Each reaction pair is placed into one of four categories: directionally coupled, partially coupled, fully coupled, or uncoupled. The procedure enables the identification of reactions that must be simultaneously active, is capable of suggesting multiple metabolic manipulations likely to have the same effect, and allows for the computational evaluation of the consequences of various gene knockouts. The FCF procedure is applied to three genome-scale metabolic reconstructions of increasing
size and complexity: (1) *Helicobacter pylori*, (2) *Escherichia coli*, and (3) *Saccharomyces cerevisiae*. The reference corresponding to this work is:


A similar framework for elucidating the conservation relationships for metabolite concentrations in genome-scale metabolic models is applied in


Chapter 7 provides a synopsis of the key contributions of the preceding chapters. A thorough description of how the frameworks presented in this thesis could be utilized in an integrated fashion to shape the gene content of microbial production strains can be found in:

Given:
(1) Stoichiometry of the network
(2) Cellular composition information
(3) Substrate uptake rate

Given:
Biomass Composition

\[
\begin{array}{c|c}
\text{comp} & \text{mmol} / \text{gDW} \\
\hline
B & 2 \\
C & 3 \\
D & 4 \\
\end{array}
\]

Optimization Model:
Maximize \( v_{\text{bio}} \)
subject to

\[
\begin{pmatrix}
-1 & 0 & 0 & 0 & 10 \\
1 & -1 & -1 & -2 & 0 \\
0 & 1 & 0 & -3 & 0 \\
0 & 0 & 1 & -4 & 0 \\
\end{pmatrix}
\begin{pmatrix}
v_1 \\
v_2 \\
v_3 \\
v_{\text{bio}} \\
b_D \\
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
0 \\
0 \\
\end{pmatrix}
\]

\( v_1, v_2, v_3, v_{\text{bio}}, b_D \geq 0 \)

Model Predictions:
Growth rate: \( v_{\text{bio}} = 1.111 \text{ hr}^{-1} \)

Figure 1.1: Given the stoichiometry of the reaction network, cellular composition information, and the uptake rate of substrate, an optimization model is formulated to predict the ideal intracellular flux distribution and cellular growth rate. (gDW = grams of dry weight of cells)
Chapter 2

Augmenting Metabolic Networks with Non-native Functionalities

2.1 Background

The utility of stoichiometric models coupled with the explosion of newly available genetic data motivates the need for a systematic procedure utilizing flux balance analysis to determine the most desirable recombinatory choices to undertake. Until now, recombinant DNA technology has been used to add straightforward conversion pathways which produce new and desirable cellular functions. For instance, *E. coli* strains producing biotin (Sabatie et al. 1991) and indigo (Ensley 1985) were designed by introducing foreign genes obtained directly from the biotin and indigo producing pathways of *Bacillus sphaericus* and *Pseudomonas putida*, respectively. Other success stories include enhancing *E. coli*'s ability to over express heterologous proteins (Aristidou et al. 1994; Chou et al. 1994; Dedhia et al. 1994) and engineering microorganisms to biodegrade pollutants such as heavy metals (Xie and Wang 1997), phosphates (Keasling et al. 1998; Van Dien and Keasling 1998), and trichloroethylene (Winter et al. 1989). It is the objective of this chapter to utilize flux balance analysis and mixed-integer programming tools to select the mathematically optimal genes for recombination into *E. coli* from a metabolic database encompassing many genes from multiple species. The resulting pathways need not lie directly on main production pathways, as they may enhance production indirectly by either redirecting metabolic fluxes into the production pathways or by increasing the energy efficiency of the present pathways.
A FBA model of the cellular metabolism of *Escherichia coli* is constructed incorporating the reaction pathways provided earlier (Pramanik and Keasling 1997) along with modifications based on more recent data (Karp et al. 2002). The modifications are either small molecule corrections (e.g., ATP in place of GTP for succinate thiokinase) based on more recent metabolic information or the removal of certain pathways now known to be absent from the *E. coli* network. This modified version of the Pramanik and Keasling model contains 289 metabolites, 178 irreversible, 111 reversible and 12 transport reactions. In addition, a stoichiometric matrix containing all metabolic reactions from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al. 2002) is compiled and incorporated into the model. This multi-species stoichiometric model containing 3,700 total reactions will heretofore be referred to as the Universal model. The mathematical framework for selecting the optimal gene candidates for recombination from this universal matrix is presented next followed by its application to improving the maximum theoretical yields of the amino acid production pathways in *E. coli*.

### 2.2 Modeling and Computational Protocol

The goal of this section is to formulate a mathematical framework to pinpoint the mathematically optimal foreign reactions from the Universal data set capable of increasing the maximum theoretical yields of the twenty amino acids in *E. coli*. The corresponding gene counterparts to these foreign reactions could be identified through current gene annotation efforts, although a one-to-one correspondence between genes and reactions is assumed in this thesis for simplicity of presentation. First, the set of reactions is partitioned into two subsets \( E \) and \( NE \) where subset \( E \) represents reactions present in *E. coli* and subset \( NE \) represents reactions present only in non-*E. coli* species. Subsequently,
the set of binary variables $y_j$ is introduced to describe the presence or absence of each reaction $j$:

$$y_j = \begin{cases} 0 & \text{if reaction } j \text{ is absent} \\ 1 & \text{if reaction } j \text{ is present} \end{cases}$$

Thus the constraint, $0 \leq v_j \leq y_j \cdot v_j^{\text{max}}$, ensures that reaction flux $v_j$ is set to zero only if the binary variable $y_j$ is set to zero. Alternatively, when $y_j$ is equal to one, $v_j$ is free to assume any value between zero and an upper bound $v_j^{\text{max}}$. In this study, $v_j^{\text{max}}$ was set to an arbitrarily high value.

The problem of maximizing the theoretical yield, $b_{\text{naa}}^{\text{UNV}}$, of a particular amino acid in the *E. coli* metabolic network by judiciously incorporating foreign reactions from the Universal network is formulated as:

Maximize $b_{\text{naa}}^{\text{UNV}}$

subject to

$$\sum_{j=1}^{M} S_{ij} v_j = b_i \quad \forall \ i$$

$$0 \leq v_j \leq y_j \cdot v_j^{\text{max}} \quad \forall \ j$$

$$y_j = 1 \quad \forall \ j \in E$$

$$v_j \in \mathbb{R}^+ \quad \forall \ j$$

$$b_i \in \mathbb{R} \quad \forall \ i$$

$$y_j \in \{0,1\} \quad \forall \ j$$

Note that this formulation ensures that all *E. coli* reactions are available (i.e., $y_j = 1$) and allows for the selection of any number of reactions from the multi-species reaction list. Foreign reactions chosen by the model but absent in *E. coli* (i.e., all non-zero $y_j$ elements of $NE$) provide routes for manipulating the cellular metabolism through recombinant DNA technology. These problems were solved using CPLEX 6.6 accessed via the
commercial software package GAMS. Problems with up to 3700 binary variables were solved on the order of seconds on an IBM RS6000-270 workstation.

2.3 Amino Acid Maximum Theoretical Yield Optimizations

We explored the theoretically optimal formation of all twenty amino acids. Each optimization run was performed for two cases: (i) including only the reactions present in *E. coli*, and (ii) allowing the model to select all reactions from the Universal stoichiometric matrix. The theoretical amino acid production capabilities of the *E. coli* metabolic network, with and without the additional reactions from the Universal matrix, are shown in Table 2.1 for growth on glucose and acetate. It must be noted that it is the structural pathway changes predicted by the model that are more meaningful than the exact numerical values because these are theoretical maximum yield calculations. Predictions by the Varma and Palsson (1993) model are shown for comparison. As expected, the maximum production capabilities by the Varma and Palsson (1993) model are slightly below the predictions of the more complex employed model due to the additional metabolic routes available for production.

The results show that improvements to seven amino acid production pathways of *Escherichia coli* are theoretically attainable with the addition of genes from various organisms. Manipulation of the arginine pathway shows the most promise, with 8.75% and 9.05% increases with additional genes for growth on glucose and acetate, respectively. The optimal recombinant asparagine pathway shows 5.77% and 5.45% increases over current *E. coli* growth on glucose and acetate, while cysteine production can be raised 3.57% and 3.80%, respectively. The histidine production pathway is revealed as another encouraging target for DNA recombination with 0.23% and 4.53%
improvements available as well. The isoleucine, methionine, and tryptophan formation pathways offer the final three genetic objectives for enhancing production.

The enzymes responsible for introducing these various improvements to the *E. coli* amino acid production pathways are shown in Table 2.2. In most cases, the addition of only one or two genes to the original amino acid production pathway results in an increased maximum theoretical yield even though the complete list of 3,400 reactions was available for selection. For example, introducing foreign genes coding for carbamate kinase and the pyrophosphate dependant version of 6-phosphofructokinase further optimizes arginine production for growth on glucose, while adding carbamate kinase and another gene coding for acetate kinase renders the arginine production pathway on acetate stoichiometrically optimal. Expressing the genes coding for aspartate-ammonia ligase and sulfate adenylyltransferase in *E. coli* results in the increased mentioned earlier in asparagine and cysteine productions, respectively. Only the production of isoleucine on glucose and acetate substrates and the production of methionine on acetate require over two additional enzymes to reach optimality according to the model.

Careful examination of these amino acid pathways reveals how these additional enzymes improve the energetic efficiency of the original routes. The original and Universal arginine production pathways for growth on glucose are shown in Figure 2.1. The two pathways differ in only two reactions – the pyrophosphate dependant analog of 6-phosphofructokinase in the Universal model replaces the ATP dependent version present in *E. coli*, and carbamate kinase in the Universal model replaces carbamoyl phosphate synthetase from the original *E. coli* model. The first improvement to energy utilization occurs because the Universal model 6-phosphofructokinase uses
pyrophosphate formed from Argininosuccinate synthase reaction instead of ATP to transfer a phosphate group to fructose-6-phosphate in the third step in glycolysis. The *E. coli* model, which sends this pyrophosphate through pyrophosphatase for hydrolytic cleavage, in effect wastes the energy from this energy-rich phosphoanhydride bond. By recapturing this otherwise wasted energy, the pyrophosphate version of 6-phosphofructokinase requires one less ATP phosphoanhydride bond per arginine molecule produced.

The second form of cellular energy savings is realized by the replacement of carbamoyl phosphate synthetase. The native carbamoyl phosphate synthetase creates one mole of carbamoyl phosphate from carbon dioxide at the expense of two ATP phosphoanhydride bonds. This reaction also requires an amino group of one glutamine molecule, which subsequently forms glutamate. Reforming glutamine from glutamate requires yet another ATP; thus each unit flux through carbamoyl phosphate synthetase requires three ATP. Carbamate kinase, incorporated in the Universal model, forms carbamoyl phosphate from carbon dioxide and ammonia at the expense of only one ATP. Therefore, carbamate kinase requires two less ATP bonds per unit flux of carbamoyl phosphate formed. Overall, the additional genes used by the Universal model save the original pathway three net ATP bonds increasing arginine production by 8.75%. A similar analysis can be performed on native and Universal arginine production routes from acetate substrate depicted in Figure 2.2.

The *E. coli* asparagine production pathway is shown in Figure 2.3 for two modes of glucose entry into the metabolic network – glucokinase and the phosphotransferase system. Interestingly, the *E. coli* model prefers glucokinase to the more common
phosphotransferase system for glucose entry during optimal asparagine production. Although glucokinase is known to play a minor role in glucose metabolism under normal conditions, replacement of the phosphotransferase system by this reaction increases asparagine production from 1.560 mol/mol glucose to 1.818 mol/mol glucose. Glucose entry via the phosphotransferase system requires substantial flux through phosphoenolpyruvate (PEP) synthase to regenerate PEP from pyruvate carrying the net expense of one ADP phosphoanhydride bond. Thus either over-expressing glucokinase in *E. coli* or adding a more active recombinant glucokinase enzyme may improve asparagine production. Figure 2.4 illustrates the optimal Universal route for asparagine production on glucose. By choosing the ADP-forming aspartate-ammonia ligase enzyme over the AMP-forming version present in *E. coli*, the energy efficiency of this pathway is improved. Presently no pathways for the conservation of the pyrophosphate bond energy have been identified in *E. coli*, thus the formation of AMP uses the equivalent of two ATP phosphoanhydride bonds. In contrast, by forming ADP, the Universal pathway requires the breakage of only one phosphoanhydride bond per unit flux. In fact, the energy efficiency of the Universal model is such that the formation of asparagine does not require ATP formation from the trans-membrane proton gradient. This gradient is used solely to transport inorganic phosphate into the cell. This mechanism improves asparagine production 5.77% for growth on glucose and 5.45% for growth on acetate.

The optimal histidine production pathways of the *E. coli* and Universal models for growth on acetate are shown in Figure 2.5. Again, the Universal model selects a reaction to conserve the phosphoanhydride bond energy of pyrophosphate generated in this case by both ATP phosphoribosyltransferase and phosphoribosyl-ATP pyrophosphatase. Thus
the Universal model is at least 2 ATP more efficient than the *E. coli* model per histidine molecule produced. In addition, the addition of glycine dehydrogenase to the *E. coli* model improves the carbon conversion of the native histidine pathway. Under optimal histidine production conditions in native *E. coli*, intracellular glycine is converted to carbon dioxide and ammonia by the glycine cleavage system. In this process, only one of glycine’s carbons is conserved by its transfer to tetrahydrofolate. The Universal model, on the other hand, conserves both carbons by converting glycine to glyoxylate which subsequently is pumped back into the glyoxylate shunt. Both mechanisms improve the maximum theoretical yield of histidine 4.53%.

### 2.4 Conclusions

The proposed optimization framework provided the quantitative means to study metabolic network performance in response to gene additions. Metabolic network performance relates to either robustness in the face of gene deletions or flux enhancements through foreign gene recombination from an ever-expanding database of available genes. Although complete gene-enzyme relationships are not available for all microorganisms, the formulation enables the incorporation of this information as it becomes available. The studies revealed that adding additional options to the *Escherichia coli* genotype by DNA recombination provided improvements to the maximum theoretical productions of seven amino acids. These improvements occur by one of two mechanisms: (i) by improving the energy efficiency or (ii) by increasing the carbon conversion efficiency of the production route.

The reliance of flux balance analysis strictly on stoichiometric characteristics is its greatest strength but also can be its most prominent weakness. The flux distributions
within the cell are ultimately uniquely determined by the regulatory mechanisms within the cell, the kinetic characteristics of cellular enzymes, and the expression of these enzymes. Assuming cells operate in a stoichiometrically optimal fashion yields a wider boundary of metabolic flux distributions than is available to the cell. Furthermore, internal cellular objectives such as growth or ATP production are often in direct competition with biochemical production for the utilization of resources. Thus the flux distributions of metabolic networks rarely resemble those obtained through maximizing the theoretical yield of a desired biochemical. The next challenge, addressed in more detail in Chapter 4, becomes how can we direct the flux distributions of central metabolism towards this “best-case” scenario calculated from the maximum theoretical yield optimizations. Nevertheless, the framework introduced here provides a valuable initial step for selecting the most optimal gene additions for enhancing the biochemical production capabilities of an organism.
Table 2.1: Model predictions of maximum theoretical yields of amino acids for growth on glucose and acetate

<table>
<thead>
<tr>
<th></th>
<th>Maximum Theoretical Yield (mmol / per 10 mmol Glucose)</th>
<th>Maximum Theoretical Yield (mmol / per 10 mmol Acetate)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Palsson '93</td>
<td>Modified Keasling '97</td>
</tr>
<tr>
<td>Alanine</td>
<td>20.00</td>
<td>20.00</td>
</tr>
<tr>
<td>Arginine</td>
<td>7.74</td>
<td>9.26</td>
</tr>
<tr>
<td>Asparagine</td>
<td>15.60</td>
<td>18.18</td>
</tr>
<tr>
<td>Aspartate</td>
<td>18.20</td>
<td>20.00</td>
</tr>
<tr>
<td>Cysteine</td>
<td>9.75</td>
<td>11.49</td>
</tr>
<tr>
<td>Glutamate</td>
<td>10.00</td>
<td>13.33</td>
</tr>
<tr>
<td>Glutamine</td>
<td>10.00</td>
<td>13.33</td>
</tr>
<tr>
<td>Glycine</td>
<td>20.00</td>
<td>35.33</td>
</tr>
<tr>
<td>Histidine</td>
<td>7.30</td>
<td>9.77</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>7.34</td>
<td>8.00</td>
</tr>
<tr>
<td>Leucine</td>
<td>6.67</td>
<td>8.00</td>
</tr>
<tr>
<td>Lysine</td>
<td>7.84</td>
<td>8.45</td>
</tr>
<tr>
<td>Methionine</td>
<td>5.74</td>
<td>7.04</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>5.29</td>
<td>5.76</td>
</tr>
<tr>
<td>Proline</td>
<td>10.00</td>
<td>10.91</td>
</tr>
<tr>
<td>Serine</td>
<td>20.00</td>
<td>23.04</td>
</tr>
<tr>
<td>Threonine</td>
<td>12.30</td>
<td>15.00</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>4.14</td>
<td>4.67</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>5.48</td>
<td>6.03</td>
</tr>
<tr>
<td>Valine</td>
<td>10.00</td>
<td>10.00</td>
</tr>
</tbody>
</table>

Palsson '93: \textit{E. coli} model proposed by Palsson (1993)
Modified Keasling '97: Modified Keasling (1997) \textit{E. coli} model as described in text
Universal Model: Modified Keasling (1997) \textit{E. coli} model augmented with non-\textit{E. coli} reactions compiled by the Kyoto Encyclopedia of Genes and Genomes
% Increase: Between the modified Keasling (1997) model and the Universal model
Table 2.2: Model selections of enzymatic reactions that will enhance the amino acid production capabilities of Escherichia coli

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Substrate</th>
<th>EC#</th>
<th>Enzyme</th>
<th>Reaction Catalyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arginine</td>
<td>Glucose: 2.7.1.90</td>
<td>6-Phosphofructokinase (pyrophosphate)</td>
<td>Fructose-6-P + PPI → Fructose-1,6-Bisphosphate + Pi</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.2.2</td>
<td>Carbamate kinase</td>
<td>ATP + NH3 + CO2 → ADP + Carbamoyl Phosphate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acetate: 2.7.2.2</td>
<td>Carbamate kinase</td>
<td>ATP + NH3 + CO2 → ADP + Carbamoyl Phosphate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.2.12</td>
<td>Acetate kinase (pyrophosphate)</td>
<td>Acetate + PPI → Pi + Acetyl-Phosphate</td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>Glucose/ Acetate: 6.3.1.4</td>
<td>Aspartate—a minima ligase (ADP-forming)</td>
<td>ATP + NH3 + L-Aspartate → Pi + ADP + L-Asparagine</td>
<td></td>
</tr>
<tr>
<td>Cysteine</td>
<td>Glucose/ Acetate: 2.7.7.5</td>
<td>Sulfate adenylyltransferase (ADP)</td>
<td>Sulfate + ADP → Pi + Adenylly-Sulfate</td>
<td></td>
</tr>
<tr>
<td>Histidine</td>
<td>Glucose: 1.4.1.10</td>
<td>Glycine dehydrogenase</td>
<td>NAD + glycine → glyoxylate + NADH + NH3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acetate: 1.4.1.10</td>
<td>Glycine dehydrogenase</td>
<td>NAD + glycine → glyoxylate + NADH + NH3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.1.1.38</td>
<td>Phosphoenolpyruvate carboxykinase (pyrophosphate)</td>
<td>PPI + Oxaloacetate → CO2 + Pi + PEP</td>
<td></td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Glucose: many</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methionine</td>
<td>Glucose: 2.7.7.5</td>
<td>Sulfate adenylyltransferase (ADP)</td>
<td>Sulfate + ADP → Pi + Adenylly-Sulfate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Acetate: 1.4.1.10</td>
<td>Glycine dehydrogenase</td>
<td>NAD + glycine → glyoxylate + NADH + NH3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.7.5</td>
<td>Sulfate adenylyltransferase (ADP)</td>
<td>Sulfate + ADP → Pi + Adenylly-Sulfate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.9.1</td>
<td>Pyruvate,phosphate dikinase</td>
<td>Pyruvate + Pi + ATP → AMP + PPI + PEP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.1.1.38</td>
<td>Phosphoenolpyruvate carboxykinase (pyrophosphate)</td>
<td>PPI + Oxaloacetate → CO2 + Pi + PEP</td>
<td></td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Glucose: 2.7.1.90</td>
<td>6-Phosphofructokinase (pyrophosphate)</td>
<td>Fructose-6-P + PPI → Fructose-1,6-Bisphosphate + Pi</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.9.1</td>
<td>Pyruvate,phosphate dikinase</td>
<td>Pyruvate + Pi + ATP → AMP + PPI + PEP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7.9.1</td>
<td>Pyruvate,phosphate dikinase</td>
<td>Pyruvate + Pi + ATP → AMP + PPI + PEP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.1.1.38</td>
<td>Phosphoenolpyruvate carboxykinase (pyrophosphate)</td>
<td>PPI + Oxaloacetate → CO2 + Pi + PEP</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.1: Optimal wild-type (non-underlined) *E. coli* and recombinant (underlined) arginine production pathways for growth on glucose.
Figure 2.2: Optimal wild-type (non-underlined) *E. coli* and recombinant (underlined) arginine production pathways for growth on acetate.
Figure 2.3: Optimal asparagine production pathways in *E. coli* for two modes of glucose utilization: glucokinase (underlined) and the phosphotransferase system (non-underlined).
**Figure 2.4:** Optimal recombinant asparagine production pathway for glucose growth.
Figure 2.5: Optimal wild-type (non-underlined) *E. coli* and recombinant (underlined) histidine production pathways for growth on acetate
Chapter 3

Identifying Minimal Metabolic Reactions Sets for Cell Growth

3.1 Background

The recent explosion of fully sequenced genomes has brought significant attention to the question of how many genes are necessary for sustaining cellular life. A minimal genome is generally defined as the smallest set of genes that allows for replication and growth in a particular environment (Cho et al. 1999). Attempts to uncover this minimal gene set include both experimental and theoretical approaches. Global transposon mutagenesis was used by Hutchison and coworkers to determine that 265 to 350 of the 480 protein-coding genes of *Mycoplasma genitalium*, the smallest known cellular genome (580 kb), are essential for survival under laboratory growth conditions (Hutchison et al. 1999). Additional experimental work (Goebl and Petes 1986; Itaya 1995) revealed that only 12% and 9% respectively of the yeast and *Bacillus subtilis* genomes are essential for cellular growth and replication. Theoretical methods stem from the assumption that genes conserved across large evolutionary boundaries are vital to cellular survival. Based on this hypothesis, a minimal set of 256 genes was compiled by Mushegian and Koonin by assuming genes common to *M. genitalium* and *Haemophilus influenzae* must be members of a minimal genome (Mushegian and Koonin 1996). Interestingly, only 6 out of 26 *Escherichia coli* open reading frames of unknown function conserved in *M. genitalium* were deemed essential to species survival (Arigoni et al. 1998). The existence of multiple, quite different, species and environment specific minimal genomes has long been speculated (Huynen 2000).
Here we describe a computational procedure for testing this claim by estimating the minimum required growth-sustaining core of metabolic reactions under different uptake conditions. The stoichiometric model of \textit{E. coli} metabolism proposed by Palsson and coworkers (Edwards and Palsson 2000a) is employed to identify the smallest set of enzymatic reactions capable of supporting given targets on the growth rate for either a glucose, an acetate, or a complex substrate. This flux balance analysis (FBA) model incorporates 454 metabolites and 720 reactions including the glycolysis, tricarboxylic acid (TCA) cycle, pentose phosphate pathway (PPP), and respiration pathways along with synthesis routes for the amino acids, nucleotides, and lipids. Growth is quantified by adding an additional reaction to the model simulating a drain on the various components of \textit{E. coli} biomass in their appropriate biological ratios. By associating a gene to each metabolic reaction in the network, gene activations and inactivations are incorporated into the FBA model using logic 0-1 binary variables. The problem of minimizing the number of active metabolic reactions required to meet specific metabolic objectives (\textit{i.e.}, growth rates) is shown to assume the mathematical structure of a generalized network flow problem where nodes denote metabolites and connecting arcs represent reactions. Alternatively, instead of a biomass target, minimum levels of ATP production or lowest allowable levels of key components/metabolites could readily be incorporated in the model. A mixed-integer linear programming (MILP) solver, CPLEX 6.5 (Brooke et al. 1998b) accessed via GAMS (Brooke et al. 1998a), is employed to solve the resulting large-scale combinatorial problems with CPU times ranging from minutes to days.

Based on the \textit{E. coli} model, the minimal reaction network is explored for different growth requirements under two contrasting uptake environments (i) restricting the uptake of organic material to a single organic component and (ii) allowing the uptake of any
organic metabolite with a corresponding transport reaction. These two extreme uptake scenarios were chosen to model maximum and minimum reliance on internal metabolism for component synthesis respectively, and probe their effect on the minimum reaction set required. Previous attempts utilized reductionist methodologies to extract the set of essential genes through a series of gene knockouts. Here we propose an efficient computational procedure for selecting the minimal set by simultaneously considering the effect of all reactions on cell growth. A minimal gene set can then be inferred by mapping the enzyme(s) catalyzing these reactions to the corresponding coding genes. While the obtained results are, in principle, dependent on the specifics of the employed flux balance *E. coli* model (Edwards and Palsson 2000a), they still provide valuable insight and perspective to the questions of what is the minimal genome and how is it shaped by the environment.

2.2 Modeling and Computational Protocol

The mixed-integer linear programming problem of minimizing the total number of functional reactions in the network capable of meeting a target for biomass production $v_{\text{target}}^{\text{biomass}}$ is as follows:
Minimize \[ Z = \sum_{j=1}^{M} y_j \]

subject to \[ \sum_{j=1}^{M} S_{ij} y_j = b_i, \quad i = 1, ..., N \]

\[ v_{\text{biomass}} \geq v_{\text{target}} \]

\[ v_{\text{min}} \cdot y_j \leq v_j \leq v_{\text{max}} \cdot y_j, \quad j = 1, ..., M \]

\[ y_j \in \{0,1\}, \quad j = 1, ..., M \]

\[ v_j \in \mathbb{R}^+, \quad j = 1, ..., M \]

\[ b_i \in \mathbb{R}, \quad i = 1, ..., N \]

The above MILP belongs to the class of generalized network problems (Ahuja et al. 1993). Here each metabolite constitutes a node and each reaction represents an arc in the network. The presence/absence of reactions is described by the binary variables \( y_j \) which assume a value of one if reaction \( j \) is active and a value of zero if it is inactive.

The presence of over one thousand binary variables causes the problem to become computationally intractable for some instances. In particular, the computational burden increases for lower biomass targets and it is much greater for case (ii) than case (i) due to the added complexity associated with multiple uptakes. To alleviate the computational burden, four preprocessing techniques are employed: (i) isoenzyme grouping, (ii) futile cycle exclusion, (iii) flux bounds generation, and (iv) connectivity constraint addition. Isoenzyme grouping refers to the aggregation of the 171 reactions catalyzed by isoenzymes. Reactions differing only in the catalyzing enzyme (i.e., isoenzymes) are grouped together treating all isoenzymes as a single reaction. This reduces complexity by pruning the total number of binary variables. Futile cycle exclusion addresses the removal of sets of reactions (2 or more) which collectively recycle fluxes in a loop...
without any net effect on metabolism. A special case is reversible reactions with nonzero fluxes for both directions. In general, a set $\mathcal{K}$ composed of $K$ reactions forms a futile cycle if

$$\sum_{j \in \mathcal{K}} S_{ij} = 0, \quad i = 1, \ldots, N$$

The following constraint ensures that at least one of them will be inactive breaking the cycle:

$$\sum_{j \in \mathcal{K}} v_j \leq K - 1$$

Overall, 346 futile cycles were identified and eliminated from the model. Most of the futile cycles involved simply reversible reactions.

The solution time of the resulting MILP problems is highly dependent on the tightness of the imposed lower $v_j^{\min}$ and upper $v_j^{\max}$ bounds on the fluxes $v_j$. Tight bounds $v_j^{\min}$ and $v_j^{\max}$ are obtained by minimizing and maximizing respectively, every single reaction flux $v_j$ subject to the flux balance constraints and the biomass target specification.

Maximize/Minimize $v_j$,
subject to

$$\sum_{j=1}^{M} S_{ij} v_j = b_i, \quad i = 1, \ldots, N$$

$$v^{\text{biomass}} \geq v^{\text{target}}$$

$$v_j \in \mathbb{R}^+, \quad j = 1, \ldots, M$$

$$b_i \in \mathbb{R}, \quad i = 1, \ldots, N$$

This is a linear programming (LP) problem (no binary variables) and is quickly solved (i.e., less than a few seconds) for all cases. Note that different bounds are generated for different biomass targets, and the higher the biomass target is, the tighter the obtained bounds are.
Connectivity constraints are also added to ensure that if a reaction producing an intracellular metabolite is active, then at least one reaction consuming this metabolite must be active and vice versa. In addition, if a reaction transporting an extracellular metabolite into the cell is active, then at least one intracellular reaction consuming this metabolite must be active and vice versa. These relations are incorporated in the model as follows after partitioning the reaction set $J$ into two subsets: $I_{int}$ representing intracellular reactions and $I_{trans}$ representing reactions transporting metabolites to and from the cell. The metabolite set $I$ is also partitioned into two subsets with $I_{int}$ and $I_{ext}$ representing intracellular and extracellular metabolites respectively.

\[
\sum_{i \in I_{int}, j \in J} S_{ij} y_{ij} \leq \sum_{j \in J} y_j \quad \forall j \in \{j \mid S_{ij} > 0\}
\]

\[
\sum_{i \in I_{int}, j \in J} S_{ij} y_{ij} \geq \sum_{j \in J} y_j \quad \forall j \in \{j \mid S_{ij} < 0\}
\]

\[
\sum_{i \in I_{ext}, j \in J_{trans}} S_{ij} y_{ij} \leq \sum_{j \in J_{trans}} y_j \quad \forall j \in \{j \mid S_{ij} > 0\}
\]

\[
\sum_{i \in I_{ext}, j \in J_{trans}} S_{ij} y_{ij} \geq \sum_{j \in J_{trans}} y_j \quad \forall j \in \{j \mid S_{ij} < 0\}
\]

These connectivity constraints are also employed to identify the smallest set of reactions capable of ensuring adequate connectivity between the external metabolites and the components of biomass. This problem involves minimizing $\sum_j y_j$ subject the above connectivity constraints with an active biomass reaction, $y_{biomass} = 1$. An algorithm for solving two similar network connectivity problems is presented by Romero and Karp (Romero and Karp 2001) and applied to the EcoCyc database (Karp et al. 2000).
The iterative generation of the multiple minimal reaction sets is achieved by accumulating integer cuts and resolving the MILP formulation. Each integer cut excludes one previously found solution. For example, solution \( y_j^* \) is excluded from consideration by adding the following integer cut:

\[
\sum_{j : y_j = 1} y_j + \sum_{j : y_j = 0} (1 - y_j) \leq M - 1
\]

All optimization problems are solved using CPLEX 6.5 (Brooke et al. 1998b) accessed through the modeling environment GAMS (Brooke et al. 1998a) on an IBM RS6000-270 workstation. The total cumulative CPU expended for this study was in the order of 400 hours.

3.3 Minimal Reaction Sets under Various Uptake Conditions

The first case study involves identifying the minimal reaction set supporting \( E. coli \) growth on a glucose substrate. A constrained amount of glucose (< 10 mmol/gDW-hr), along with unconstrained uptake routes for inorganic phosphate, oxygen, sulfate, and ammonia are enabled to fuel the metabolic network. Secretion routes for every metabolite capable of exiting the cell are also provided. Under these conditions, the FBA model predicts that the \( E. coli \) reaction network is capable of achieving a maximum theoretical growth rate of 0.966 g biomass/gDW-hr, which we will refer to as the maximum growth rate (MGR). By requiring the reaction network to match the MGR we determined that at least 234 reactions out of 720 are required for maximum growth on glucose.

The growth demands are then relaxed in subsequent studies to identify the minimal number of metabolic reactions required to meet various sub-maximal growth demands (% of MGR). Interestingly, the number of necessary metabolic reactions decreases only
mildly with the falling growth demands imposed on the network as indicated by Figure 3.1A. While a reaction set comprised of 234 reactions is needed for maximum growth, the minimal reaction set corresponding to growth rates of 30% and lower involves only 224 reactions. The same minimal reaction set persists even for growth rates as low as 0.1% of the MGR. In general, the reaction set reductions are attained by successively eliminating energy producing reactions occurring in (i) glycolysis, (ii) the TCA cycle, and (iii) the pentose phosphate pathway as the growth demands are lessened. However, certain reactions absent at higher growth rates enter the minimal sets at lower growth rates suggesting a much more complex mechanism of flux redirection than successive reaction elimination. A detailed description of the reactions entering/leaving the minimal reaction set as the imposed growth requirements are lowered is provided in Table 3.1.

For comparison, a similar study enabling a constrained amount of acetate (< 30 mmol/gDW·hr) to enter the network instead of glucose was performed (see Figure 3.1A). Here the network is much less tolerant of reaction set reductions than in glucose study. While for a glucose substrate the minimal network sizes decrease from 234 to 224 reactions as the growth demands are lowered, for an acetate substrate the network sizes reduce only from 231 to 229 reactions. This implies that the minimal reaction set size is not only dependent on the imposed biomass production requirements, but also on the specific choice for the single substrate.

It is important to note that neither the minimal reaction sets nor their corresponding reaction fluxes are unique. For example, for the 30% glucose uptake case we identified over 100 different minimal reaction sets containing exactly 224 enzymatic reactions without even counting the multiplicities associated with the 171 isoenzymes present in
the network. Among most of these multiple minimal reaction sets, the activity and flux directions of the major pathways differ very little. Most variations are concentrated on the catabolic parts of the networks. For instance, while some minimal reaction sets secrete carbon dioxide, acetate, and fumarate as the only metabolic byproducts, other sets may also secrete varying amounts of formate, glycerol, and the amino acids phenylalanine and tyrosine. These results provide a computational confirmation of the astounding redundancy and flux redirection versatility of the *E. coli* network. More importantly, all minimal reactions sets identified include 11 of 12 reactions whose corresponding gene deletions were determined experimentally to be lethal for growth on glucose. Earlier analyses (Edwards and Palsson 2000a) based on single gene deletions conducted with this model using linear optimization identified only 7 of 12 lethal gene deletions motivating the importance of considering simultaneous gene deletions within an MILP framework.

In the second case study, the uptake or secretion of any organic metabolite is enabled. The amount of organic material entering the network is kept consistent with the first case study by allowing the uptake of a constrained amount of carbon atoms (< 60 mmol/gDW·hr). Unconstrained uptake routes for oxygen, inorganic phosphate, sulfate, and ammonia are also provided as in the first study. Under these "ideal" uptake conditions, we find that a maximum growth rate (MGR) of 1.341 g biomass/gDW·hr is attainable requiring at least 201 metabolic reactions. The fact that only five amino acids are imported under maximum growth (i.e., MGR) conditions indicates that it is stoichiometrically more favorable to produce most amino acids internally rather than transport them into the cell from the medium.
This trend, however, is quickly reversed as the growth rate requirement is reduced. This reversal yields a corresponding sharp decrease in the total number of required reactions as a direct result of the importation of an increasing number of metabolites at sub-maximum target growth demands. Table 3.2 lists the metabolites uptaken or secreted at each target growth rate, while Figure 3.1B (100% - 90% of MGR) and Figure 3.1C (100% - 1% of MGR) illustrate the number of required metabolic reactions needed to attain various target growth demands. The rapid reduction in size of the minimal reaction sets by importing an increasing number of metabolites as the biomass demands are lessened (see Table 3.2) continues until the growth demands are reduced to about 90% from the MGR. Below this growth target (see Figure 3.1C) additional but modest reductions are achieved primarily through flux redirections. Table 3.3 summarizes the reactions which are being removed or added to the minimal reaction set as the growth target is successively lowered. The smallest minimal reaction network for the second case study, comprised of 122 reactions, is reached when the target growth demands are lowered to 10% of the MGR. This minimal network is comprised mostly of cell envelope and membrane lipid biosynthetic reactions, along with a number of transport and salvage pathway reactions, as shown in Table 3.4. As in the glucose-only study, multiple minimal reaction sets for multi-organic uptake case are expected.

By solving a related problem, we find that only 91 reactions are required to provide sufficient network connectivity between the available external metabolites and constituents of biomass for the multiple organic uptake study. However, because this minimal set is based strictly on network connectivity it inherently neglects the specific stoichiometry of each reaction thus underestimating the minimal reaction set size.
Available as supplementary material to this work (Burgard et al. 2001) are (i) the flux distributions associated with the minimal reaction sets under different growth targets and the corresponding genes coding for the catalyzing enzymes; (ii) a partial list comprised of 15 distinct alternate minimal reaction sets for the glucose-only uptake case study.

2.4 Conclusions

In this study, we have identified the minimum number of *E. coli* metabolic reactions capable of supporting growth under two different uptake environments (i) a glucose or acetate-only uptake environment and (ii) free uptake or secretion of any organic metabolite involving a corresponding transport reaction. The obtained results quantitatively demonstrate that minimal reaction sets and thus corresponding minimal gene sets are strongly dependent on the uptake opportunities afforded by the growth medium. While an *E. coli* cell grown on a medium containing only glucose or acetate requires at least 224 or 229 metabolic reactions respectively to support growth, a cell cultured on a rich optimally engineered medium could theoretically support growth with as few as 122 metabolic reactions. In addition, the choice of the single substrate affects the minimal reaction set size and composition. As expected, the minimal reaction set becomes larger by increasing the required growth rate. However, the magnitude of this increase is quite different for the examined cases. While in case (i) the minimal reaction set increases only from 224 to 234 to meet the maximum growth rate on glucose and from 229 to 231 for acetate growth, in case (ii) the minimal reaction set almost doubles going from 122 to 201. Another significant observation is the large redundancy of the *E. coli* metabolic network, which is capable of supporting growth utilizing only 31% of the available metabolic reactions for growth on glucose, and only 17% of the available
reactions for growth on a complex medium. Even these reduced minimal reaction network sets exhibit large multiplicities. Specifically, a non-exhaustive list of 100 alternative minimal reaction sets were identified for the glucose-only uptake case.

It must be noted that our analysis provides a species-specific minimal metabolic reaction set, which is a subset of the complete \textit{E. coli} minimal genome. This is a consequence of the adopted reaction-based analysis which cannot account for genes associated with translation, replication, recombination, repair, transcription, and genes of unknown function. A comparison of our minimal metabolic reaction set with the essential gene set of Hutchison \textit{et al.} (1999) and the minimal gene set proposed by Mushegian and Koonin (1996) in their studies with \textit{Mycoplasma genitalium} is provided in Table 3.5. The obtained results agree conceptually with the finding of Hutchison and coworkers that limited metabolic capacity can be compensated for by a proportionately greater dependence on the importation of nucleosides, amino acids, and other metabolites. Although a complete genome-based reconstruction of the \textit{M. genitalium} metabolic network is currently unavailable preventing a reaction-by-reaction comparison, the distributions of metabolic genes/reactions among the various functional classifications in the three studies are quite similar. Thus, perhaps the simultaneous reaction removal strategy applied to \textit{E. coli} in this work parallels the evolutionary pressures placed on \textit{M. genitalium} to reduce its genome size. The minimal reaction set size overestimation in our analysis may be largely due to its species-specific nature. Whereas the cellular envelope of \textit{E. coli} contains a cell wall made up largely of peptidoglycan, the cellular envelope of mycoplasmas lacks a cell wall. Thus many of the cellular envelope reactions necessary for \textit{E. coli} survival are not included in the genes sets of Hutchison \textit{et al.} (1999) and
Mushegian and Koonin (1996). Another contributing factor is that we assign a different reaction/gene to the uptake or secretion of each metabolite although similar metabolites can be transported by mechanisms associated with a single gene. Furthermore, since our analysis is based on the E. coli model, more efficient reaction combinations, perhaps occurring in non-E. coli species, could further reduce the minimal gene set lowering the discrepancy.

This framework can be utilized to construct minimal reaction sets for additional species. By contrasting these minimal sets it could be inferred how minimal reaction sets (metabolic gene sets) compare along different evolutionary branches. Specifically, minimal reaction sets for M. genitalium and H. influenza could be determined and benchmarked with earlier studies (Mushegian and Koonin 1996). Additionally, a species independent minimal metabolic reaction set can be pursued by lumping reactions occurring in many different species (Kanehisa and Goto 2000; Karp et al. 2000; Maranas and Burgard 2001) within a Universal stoichiometric matrix (Schilling et al. 1999; Burgard and Maranas 2001). As more elaborate models are developed describing elementary functions of minimal cells, such as the work of Michael Shuler for the initiation of DNA replication (Browning and Shuler 2001) and purine and pyrimidine transport and metabolism (Castellanos et al. 2004), more detail can be added to the model. Even though the proposed computational procedure is dependent upon the assumptions of the adopted stoichiometric model, it affords the versatility to study different uptake/secretion environments as well as encompass reaction sets from multiple species in the search for the minimal genome.
Table 3.1: Evolution of minimal reaction sets for case (i) under decreasing growth requirements.

<table>
<thead>
<tr>
<th>Target % Maximum Growth Rate</th>
<th>Minimal Reaction Set (# Reactions)</th>
<th>Key Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>234</td>
<td>The glycolysis, tricarboxylic acid cycle, and pentose phosphate pathways are all operating in their forward directions, optimally generating the energy cofactors ATP, NADH, and NADPH required for cell growth. All available glucose is oxidized into the cell's only secreted byproduct, carbon dioxide.</td>
</tr>
<tr>
<td>90%</td>
<td>229</td>
<td>The fluxes through two TCA cycle reactions 2-ketoglutarate dehydrogenase and succinate dehydrogenase are zero while succinyl-CoA synthetase operates in its reverse direction suggesting a less demanding energetic state under the sub-maximal growth demands. Acetate is now secreted as a byproduct along with carbon dioxide.</td>
</tr>
<tr>
<td>80%</td>
<td>228</td>
<td>Fluxes through two additional TCA cycle reactions, fumarase and malate dehydrogenase, are eliminated while a reaction secreting fumarate is added.</td>
</tr>
<tr>
<td>70%</td>
<td>226</td>
<td>The pentose phosphate pathway operates solely for nucleotide biosynthesis with the reaction fluxes through ribulose phosphate 3-epimerase, transketolase I, transketolase II, and transaldolase B all operating in reverse. Fluxes through glucose-6-phosphate dehydrogenase, lactonase, and 6-phosphogluconate dehydrogenase are absent in this case, replaced by pyridine nucleotide transhydrogenase which meets the cellular NADPH needs. In addition, formate is now secreted along with acetate, fumarate, and carbon dioxide.</td>
</tr>
<tr>
<td>60%, 50%, 40%</td>
<td>225</td>
<td>Acetate is no longer secreted as a metabolic byproduct, but is converted to acetyl-CoA by acetyl-CoA synthetase.</td>
</tr>
<tr>
<td>30%, 20%, 10%, 1%</td>
<td>224</td>
<td>Three glycolytic reactions, phosphoglycerate mutase, enolase, and pyruvate kinase are eliminated, but both serine deaminase and phosphoenolpyruvate synthase are added to supply the cell with phosphoenolpyruvate.</td>
</tr>
</tbody>
</table>
**Table 3.2:** Metabolites uptaken or secreted at each target growth rate on an optimally engineered medium.  
U – denotes metabolite uptake  
S – denotes metabolite secretion

| Metabolite                  | Percentage of 100% Biomass Generation Required | 100% | 99.5% | 99% | 98% | 97% | 96% | 95% | 90% | 85% | 80% | 70% | 60% | 10% |
|-----------------------------|-----------------------------------------------|------|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Acetate                     |                                               |      |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Acetaldehyde                |                                               |      |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Adenine                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Adenosine                   |                                               | U    | U     |     |     |     |     |     |     |     |     |     |     |     |     |
| Alanine                     |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Arginine                    |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Asparagine                  |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Aspartate                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Carbon dioxide              |                                               | S    | S     | S   | S   | S   | S   | S   | S   |     |     |     |     |     |     |
| Cysteine                    |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| D-Alanine                   |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Thymidine                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Ethanol                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Glycerol                    |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Glycerol-3-phosphate        |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Glutamine                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Glutamate                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Glycine                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Guanine                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Guanosine                   |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Histidine                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Isoleucine                  |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Leucine                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Lysine                      |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Meso-diaminopimelate        |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Methionine                  |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Mannitol                    |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Ammonia                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Oxygen                      |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Phenylalanine               |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Phosphate                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Proline                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Putrescine                  |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Pyruvate                    |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Ribose                      |                                               | U    |       |     |     |     |     |     |     |     |     |     |     |     |     |
| Serine                      |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Spermidine                  |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Threonine                   |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Tryptophan                  |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Tyrosine                    |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Uracil                      |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Uridine                     |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| Valine                      |                                               | U    | U     | U   | U   | U   | U   | U   | U   |     |     |     |     |     |     |
| **# Metabolites Uptaken**   |                                               | 12   | 17    | 19   | 21  | 22  | 24  | 26  | 28  | 29  | 31  | 29  | 34  | 34  |
**Table 3.3:** Evolution of minimal reaction sets for case (ii) under decreasing growth requirements.

<table>
<thead>
<tr>
<th>Target % Maximum Growth Rate</th>
<th>Minimal Reaction Set (# Reactions)</th>
<th>Key Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% 201</td>
<td></td>
<td>The organic material transported into the cell includes ethanol and glycerol-3-phosphate which fuel glycolysis, the TCA cycle, and PPP. The flux directions of the glycolysis pathway are split with all reaction fluxes preceding glyceraldehyde-3-phosphate (G3P) dehydrogenase operating in reverse, and all fluxes following and including G3P dehydrogenase operate in their forward directions. Putrescine, spermidine, and five amino acids are transported into the network eliminating the need for biosynthetic pathways for these components.</td>
</tr>
<tr>
<td>90% 132</td>
<td></td>
<td>While the PPP and TCA cycle reactions are still functional, the network no longer utilizes the five glycolytic reactions from glyceraldehyde-3-phosphate dehydrogenase to pyruvate kinase. Consequently, the TCA cycle is completely fueled by imported ethanol and acetate rather than flux from the glycolysis pathway.</td>
</tr>
<tr>
<td>80% 125</td>
<td></td>
<td>This network tolerates the complete elimination of the TCA cycle and glyoxylate shunt. As a result, the function of the pentose phosphate pathway reactions is no longer restricted to nucleotide biosynthesis, but now includes the formation of cellular NADPH. Most of this NADPH is subsequently converted to NADH by pyridine nucleotide transhydrogenase to replace the cellular reducing power lost from the inactivity of the TCA cycle.</td>
</tr>
<tr>
<td>70% 124</td>
<td></td>
<td>A slightly less efficient set of internal metabolic reactions enables the growth demands to be met with the importation of one less metabolite (i.e. one less transport reaction) than its 80% counterpart.</td>
</tr>
<tr>
<td>60%, 50%, 40%, 30%, 20% 123</td>
<td></td>
<td>Neither the TCA cycle nor PPP are utilized for reducing power. Most of the cellular reducing capabilities are now generated from the uptake of ethanol and its subsequent conversion into acetyl-CoA.</td>
</tr>
<tr>
<td>10%, 1% 122</td>
<td></td>
<td>This minimal network is comprised mostly of cell envelope and membrane lipid biosynthetic reactions, along with a number of transport and salvage pathway reactions. Here, the three core metabolic routes, glycolysis, the TCA cycle, and the pentose phosphate pathway are almost completely dismantled with only one glycolytic and 4 PPP reactions remaining.</td>
</tr>
</tbody>
</table>
Table 3.4: Functional classification of minimal network reactions for growth on an optimally engineered medium.

<table>
<thead>
<tr>
<th>Functional Classification</th>
<th># rxns</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA Isomerization</td>
<td>1</td>
</tr>
<tr>
<td>Alternative Carbon Source</td>
<td>7</td>
</tr>
<tr>
<td>Anaplerotic Reactions</td>
<td>1</td>
</tr>
<tr>
<td>Cell Envelope Biosynthesis</td>
<td>29</td>
</tr>
<tr>
<td>EMP Pathway</td>
<td>5</td>
</tr>
<tr>
<td>Membrane Lipid Biosynthesis</td>
<td>16</td>
</tr>
<tr>
<td>Pentose Phosphate Pathway</td>
<td>4</td>
</tr>
<tr>
<td>Pyrimidine Biosynthesis</td>
<td>1</td>
</tr>
<tr>
<td>Respiration</td>
<td>5</td>
</tr>
<tr>
<td>Salvage Pathways</td>
<td>17</td>
</tr>
<tr>
<td>Transport</td>
<td>36</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>122</strong></td>
</tr>
</tbody>
</table>
Table 3.5: Comparison of minimal metabolic gene/reaction sets based on functional classification

<table>
<thead>
<tr>
<th>Metabolic Function</th>
<th>Essential Gene Set&lt;sup&gt;+&lt;/sup&gt; &lt;br&gt;Ref. (2)</th>
<th>Minimal Gene Set &lt;br&gt;Ref. (5)</th>
<th>Minimal Reaction Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Genes</td>
<td># Genes</td>
<td># Reactions</td>
</tr>
<tr>
<td>Amino acid biosynthesis</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Biosynthesis of cofactors, prosthetic groups, and carriers</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Cell envelope</td>
<td>2</td>
<td>11</td>
<td>29</td>
</tr>
<tr>
<td>Central intermediary metabolism</td>
<td>7</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Energy metabolism</td>
<td>31</td>
<td>32</td>
<td>21</td>
</tr>
<tr>
<td>Fatty acid and phospholipid metabolism</td>
<td>5</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>Purines, pyrimidines, nucleosides, and nucleotides</td>
<td>17</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Transport and binding proteins</td>
<td>17</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td><strong>83</strong></td>
<td><strong>99</strong></td>
<td><strong>122</strong></td>
</tr>
</tbody>
</table>

* Gene functions based on the current TIGR (20) version of the M. genitalium annotation
+ Obtained from disrupting the expression of individual genes and determining which are essential for cellular survival
Figure 3.1: Number of reactions in each minimal set as a function of the imposed growth demands for (A) a glucose or acetate-only uptake environment and (B, C) an uptake environment allowing multiple organic uptakes.
Chapter 4

Inferring and Testing Hypothesized Metabolic Imperatives

4.1 Background

Living organisms have evolved to maximize their chances for survival (Darwin 1899). This is manifested at the level of metabolism with the presence of multiple redundant pathways leading to and from key intermediates so that the removal of a single enzyme will (likely) not prevent an organism’s ability to produce key components (Edwards and Palsson 2000c; Schilling and Palsson 2000; Price et al. 2002). Furthermore, experimental evidence suggests that organisms have developed control structures to ensure optimal growth in response to environmental constraints (Edwards et al. 2001). While the existence of a fitness function driving an organism’s evolution is widely accepted, it is unclear whether its fingerprint can be detected in the flux distributions of primary metabolism. Specifically, are metabolic networks driven to evolve as optimal biomass producers, maximum ATP generators, or optimal consumers of available substrates? In this chapter, we address the question of whether such a fitness function, or in optimization language objective function, can be identified from experimentally determined metabolic fluxes. We also examine how this fitness surrogate varies as environmental conditions change.

There exist two classes of metabolic modeling frameworks that inherently account for the presence of a fitness function that drives the metabolic machinery towards optimal
survivability. First, the cybernetic modeling approach assumes that an organism is an optimal strategist utilizing all available resources with maximum efficiency (Ramkrishna et al. 1987). The expression and activity of the enzymes that catalyze network functionality are regulated by cybernetic control variables obtained from the solution of a constrained optimization problem (Kompala et al. 1984; Dhurati et al. 1985). This framework also contends that even genetically altered systems have the same underlying goal of optimal resource allocation although the set of competing physiological choices open to the network expands or contracts depending upon the nature of the alteration (Varner and Ramkrishna 1999a; Varner and Ramkrishna 1999b). Cybernetic models have been successful in modeling the growth dynamics of yeast in batch and continuous cultures (Jones and Kompala 1999), diauxic growth patterns and simultaneous consumption of substrates of *Escherichia coli* (Ramakrishna et al. 1996) and *Klebsiella oxytoca* (Kompala et al. 1986), and the time evolution of the aspartate family of amino acids in *Corynebacterium lactofermentum* (Varner and Ramkrishna 1999c).

Stoichiometric or flux balance analysis (FBA) models, on the other hand, employ this optimality principal in a slightly different fashion (Edwards et al. 2002). They use only the stoichiometric mass balances of the metabolic network and cellular composition information to generate the broadest set of reaction flux distributions potentially available to the cell. This assumes that a metabolic network is capable of spanning all flux combinations allowable by the stoichiometric constraints with the maximization of biomass production typically postulated as the objective function. Stoichiometric models have in some cases been successful in predicting the phenotypical characteristics of cells such as growth rates (Pons et al. 1996; Edwards et al. 2001), metabolic byproduct

While many hypotheses have been put forward as surrogates for cellular fitness functions, substantially less work has been conducted towards systematically validating them with experimentally derived flux distributions of metabolic networks. This is due in part to fact that the complete quantification of fluxes throughout the central metabolic pathways was intractable until recently. Traditionally, the only observables have been the growth rate, the uptake and secretion rates of substrates and products, and those reaction fluxes that can be directly calculated based on the uptake or secretion of these external compounds. The development of complete isotopomer models (Zupke et al. 1997; Klapa et al. 1999; Park et al. 1999; Christensen and Nielsen 2000; Forbes et al. 2001), however, enables the calculation of the amount of reaction flux passing through every reaction of central metabolism. Briefly, isotopomer analysis involves the uptake of a \(^{13}\text{C}\) labeled substrate and the subsequent analysis of the labeling state by NMR and/or GC/MS measurements (Szyperski 1995; Wiechert and de Graaf 1996; Christensen and Nielsen 1999; Schmidt et al. 1999a; Schmidt et al. 1999b). This allows for a more accurate quantification of intracellular flux distributions providing an additional set of observables to test the various fitness function hypotheses.

In this chapter, we introduce a mathematically rigorous framework for testing whether experimental flux data are consistent with different hypothesized objective functions. Rather than starting by postulating such an objective function, or even
accepting that one exists, we introduce a quantitative framework akin to inverse optimization (Tarantola 1987) for inferring or disproving the consistency of different hypotheses. Specifically, we examine whether the maximization of a weighted combination of fluxes can explain a set of observed experimental data. For example, the driving force governing cellular metabolism may be a compromise between ATP, redox, and biomass production fluxes depending on the state of the system. FBA-based modeling typically assumes that the objective consists of a unit vector in the direction of the growth flux comprised of all necessary components of biomass in their corresponding biological ratios (Ingraham 1983). Other postulated objective functions include (i) the maximization of metabolite (Varma and Palsson 1993a) or ATP (Majewski and Domach 1990; Ramakrishna et al. 2001) production, and (ii) the minimization of the Euclidean norm (i.e., sum of the fluxes) (Bonarius et al. 1996), nutrient uptake or redox production (Savinell and Palsson 1992). The approach proposed here, referred to as ObjFind, requires the solution of a bilevel optimization problem that minimizes the squared deviations of identified fluxes from experimental data while ensuring that the identified fluxes are the product of an optimization problem. A solution strategy founded upon duality theory concepts is discussed in detail in the following section. Based on this framework, we examine (i) what is the objective function (if any) of a metabolic network that is the most consistent with experimental flux data, (ii) whether this objective function is unique, and (iii) how the driving forces governing cellular metabolism vary under different environmental conditions.
4.2 The ObjFind Bilevel Optimization Framework

Mathematically, deciphering this balance requires the identification of the weights or coefficients $c_j$ that accept a set of experimental fluxes $v_j^{\text{exp}}$ as an optimal solution to the following linear programming (LP) problem called the Primal

\[
\text{maximize} \quad Z_p = \sum_{j \in P} c_j v_j \quad \text{(Primal)}
\]

\[
\text{subject to} \quad \sum_{j=1}^{M} S_{ij} v_j = 0, \quad \forall \ i \in N
\]

\[
v_{\text{GLC}} = \text{upt. rate}, \quad \forall \ j \in \text{glc. upt.}
\]

\[
v_j \geq 0, \quad \forall \ j \in M
\]

where $S_{ij}$ is the stoichiometric coefficient of metabolite $i$ in reaction $j$, $v_j$ represents the flux of reaction $j$, $v_{\text{GLC}}$ is the uptake rate of glucose, and $c_j$ is a weight associated with importance of the reaction flux $v_j$ referred to as a coefficient of importance. The coefficients of importance (CoIs) are scaled so as their sum is equal to one. Intuitively, the CoIs quantify the fraction of the additive contribution of a given flux $v_j$ to the fitness function whose optimization explains the experimental flux data. A high value for $c_j$ implies that the experimental flux data are consistent with the hypothesis that the flux $v_j$ is driven towards its maximum allowable value whereas a low value implies the converse. The vector $v$ includes both internal and external fluxes and reversible reactions are defined as two irreversible reactions in opposite directions, constraining all fluxes to positive values.

Given a set of fluxes $v_j^{\text{exp}}$ which have been experimentally determined, it is possible to determine which linear combinations of reaction fluxes weighted by $c_j$ are maximized such that the experimental fluxes are optimal with respect to the linear programming
problem described by the Primal. This requires the solution of the following bilevel optimization problem

\[
\begin{align*}
\text{minimize} & \quad \sum_{j \in E} (v_j - v^{\exp}_j)^2 \\
\text{subject to} & \quad \sum_{j \in P} c_j v_j \\
\text{subject to} & \quad \sum_{j=1}^M S_g v_j = 0, \quad \forall \ i \in N \\
& \quad v_{\text{GLC}} = \text{upt. rate}, \quad \forall \ j \in \text{glc. upt.} \\
& \quad v_j \geq 0, \quad \forall \ j \in M \\
& \quad \sum_{j \in P} c_j = 1 \\
& \quad c_j \geq 0, \quad \forall \ j \in P
\end{align*}
\]

in which the coefficient of importance \((c_j)\) values for the inner problem are adjusted by the outer problem so that the sum-squared difference between the experimental fluxes and the optimal solution \(v_j\) for the inner problem is minimized. Note that the ObjFind problem includes the optimality of the Primal problem as a constraint giving rise to two nested optimization problems.

We propose an efficient solution approach borrowing from LP duality theory which shows that for every linear programming problem (primal) there exists a unique optimization problem (dual) whose optimal objective value is equal to that of the primal problem. The dual problem associated with the \textbf{Primal} linear programming problem (Bertsimas and Tsitsiklis 1997) is
minimize \[ Z_D = (\text{upt. rate}) \cdot g \] \hspace{1cm} \text{(Dual)}

subject to

\[ \sum_{i=1}^{N} u_i S_{ij} \geq c_j, \quad \forall \ j \in P \]

\[ \sum_{i=1}^{N} u_i S_{ij} \geq 0, \quad \forall \ j \not\in P, \text{glc. upt.} \]

\[ \sum_{i=1}^{N} u_i S_{ij} + g = 0, \quad \forall \ j \in \text{glc. upt.} \]

where \( u_i \) is the dual variable associated with the first set of constraints in the Primal and \( g \) is the dual variable associated with the glucose uptake constraint. The dual variables, \( u_i \) and \( g \), indicate the change in the optimal value of \( Z_P \) per unit change in the right hand side of their associated constraint. Likewise, the reaction fluxes, \( v_j \), are the dual variables associated with the constraints of the Dual problem.

The concept of strong duality (see Figure 4.1) implies that if the primal has an optimal solution, so does the dual, and their respective optimal objective values are equal. Furthermore, the primal and dual problems can be simultaneously feasible only at their respective optimal solutions. Therefore by constructing an optimization problem formulation that includes both the Primal and Dual constraints along with an equality constraint forcing their respective objective function values to be equal to each other, we ensure that any feasible solution \((v_j, g, u_i)\) will be optimal to both the Primal and Dual problems. Therefore, the solution of the following single level nonlinear optimization problem
minimize $\sum_{j \in E} (v_j - v_j^{\text{exp}})^2$ \hspace{1cm} (ObjFind)

subject to $\sum_{j \in P} c_j v_j = (\text{upt. rate}) \cdot g$

$\sum_{j \in P} c_j = 1$

$\sum_{j=1}^{M} S_y v_j = 0, \quad \forall i \in N$

$v_{GLC} = \text{upt. rate}, \quad \forall j \in \text{glc. upt.}$

$\sum_{i=1}^{N} u_i S_{ij} \geq c_j, \quad \forall j \in P$

$\sum_{i=1}^{N} u_i S_{ij} \geq 0, \quad \forall j \notin P, \text{ glc. upt.}$

$\sum_{i=1}^{N} u_i S_{ij} + g = 0, \quad \forall j \in \text{glc. upt.}$

$v_j \geq 0, \quad \forall j \in M$

$c_j \geq 0, \quad \forall j \in P$

systematically characterizes the set of all possible $c_j$ values consistent with the minimization of sum-squared difference between a subset of observed fluxes $v_j^{\text{exp}}$ and an optimal solution to the Primal. Note that any problems resulting from the presence of multiple optima to the primal are circumvented by including the flux variables directly in the dual minimization. By utilizing LP duality concepts, a method is introduced for transforming the original intractable two-stage optimization problem into a single-stage optimization problem.

A related problem of maximizing the value of a particular coefficient $c_j^*$ subject to the sum of the squared deviations being below a target value is
maximize \( c_j^* \)

subject to \( \sum_{j \in P} c_j v_j = (\text{upt. rate}) \cdot g \)

\( \sum_{j \in E} (v_j - v_j^{\exp})^2 \leq \text{target} \)

\( \sum_{j \in P} c_j = 1 \)

\( \sum_{j=1}^{M} S_j v_j = 0, \quad \forall i \in N \)

\( v_{GLC} = \text{upt. rate}, \quad \forall j \in \text{glc. upt.} \)

\( \sum_{i=1}^{N} u_i S_{ij} \geq c_j, \quad \forall j \in P \)

\( \sum_{i=1}^{N} u_i S_{ij} \geq 0, \quad \forall j \notin P, \text{ glc. upt.} \)

\( \sum_{i=1}^{N} u_i S_{ij} + g = 0, \quad \forall j \in \text{glc. upt.} \)

\( v_j \geq 0, \quad \forall j \in M \)

\( c_j \geq 0, \quad \forall j \in P \)

where \( c_j^* \) can be the weight associated with any potential cellular objective (i.e., biomass formation, energy production, etc.). It should be noted that the constraint \((Z_P = Z_D)\) is nonconvex due to the bilinear \( c_j v_j \) terms. Therefore, multiple starting points were used to identify multiple optimal solutions in each case. We observed that after over one hundred restarts, only a handful of multiple optima were identified. Problems containing as many as 200 variables were solved in seconds using MINOS 5.0 accessed via the GAMS modeling environment on an IBM RS6000-270 workstation.
4.3 Aerobic and Anaerobic Escherichia coli Metabolic Objectives

The ObjFind procedure is applied to the central metabolic network of *E. coli*. Experimental flux values *v_j^exp* determined from an isotopomer analysis (Schmidt et al. 1999a) study for both aerobic and anaerobic growth conditions are used in conjunction with a stoichiometric model of *E. coli* central metabolism (Palsson 2002) in an effort to pinpoint which underlying driving forces are governing the network’s operation. This model, comprised of 62 reactions and 48 metabolites, includes all reactions of glycolysis, the TCA cycle, and the pentose phosphate pathway as well as a number of respiration reactions. Coefficients of importance are assigned to each reaction flux associated with a metabolite drain, energy dissipation, or redox potential dissipation. In other words, an assignment is made for every flux that consumes, by either draining or dissipating, a resource in the network. The reaction fluxes associated with these coefficients are shown with colored arrows in Figure 4.2. Note that previously postulated objective functions are encompassed here as linear combinations of these reaction fluxes.

In our first case study, we identify Col’s consistent with the experimental fluxes *v_j^exp* being optimal to the LP problem maximizing $\sum_j c_j v_j$ subject to the network stoichiometry. Consequently, if this LP problem is solved using the coefficients *c_j* identified by ObjFind, an optimal solution *v_j^* exists, though not necessarily unique, such that $\sum_j (v_j^* - v_j^\text{exp})^2$ is minimized. The minimum sums of the squared flux deviations from the experimental data for the aerobic and anaerobic fluxes were found to be 0.016 (mmol/gDW·hr)$^2$ and 0.797 (mmol/gDW·hr)$^2$, respectively, which were well within the experimental error. The identified Col’s consistent with the aerobic and anaerobic
experimental flux distributions are superimposed in Figure 4.3. Remarkably, the CoI’s for both growth conditions are strikingly similar even though the flux distributions (see Table 4.1) for the two cases are quite different. This unexpected convergence is consistent with the presence of a single metabolic objective driving the flux distributions in both cases. This objective is exemplified by the values of the coefficients of importance. It appears that fluxes with similar CoI’s cluster within groups that are both topologically and functionally related. Specifically, seven fluxes shown in purple are clustered where the glycolysis pathway meets the TCA cycle. In addition, the fluxes with the largest coefficients, shown in red in Figure 4.2, are associated with drains of metabolites not far from glucose in the metabolic network, while the smallest coefficients, shown in blue, are associated with ATP and NADH dissipation. The most notable differences between the two sets of coefficients of importance are associated with the secretion of acetate, ethanol, and succinate where the anaerobic CoI’s are much larger. This is consistent with the fact that these metabolic byproducts are secreted only under anaerobic conditions.

Next we investigate the effect of deviations in the flux distributions from the experimental ones on the robustness of the values identified for the coefficients of importance. Specifically, the sum of the squared deviations between the identified and experimental fluxes is allowed to increase from its minimum value by (i) zero, (ii) one, (iii) ten, and (iv) one hundred squared flux units (mmol/gDW⋅hr)²; while at the same time each CoI is maximized individually to identify its sensitivity to experimental errors. The deviations of the coefficients of importance from their nominal values for both aerobic and anaerobic case are shown in Figure 4.4. Clearly, while some changes are present, the maximum extent possible is quite modest for small experimental errors implying a
substantial robustness in the assignment of values to the coefficients of importance with respect to experimental error. In addition, these deviations seem to be proportional to the nominal values of the coefficients of importance and are approximately 95% larger for the anaerobic versus the aerobic case.

After identifying the coefficients of importance and verifying their robustness to experimental error we turn our attention towards deciphering the biological significance of their values. Specifically, we examine how close these coefficients of importance track the biomass maximization hypothesis. A biomass reaction flux (Varma and Palsson 1993b), complete with energy and reducing power requirements, is added to the network to drain metabolic precursors in their appropriate ratios as proposed by (Ingraham 1983) for biomass formation. A coefficient of importance is assigned to this aggregate biomass flux. Note that an infinite number of solutions exist for the CoI’s because the biomass flux is comprised of a linear combination of the other drain fluxes. We thus identify its maximum value capable of explaining the flux distributions for the aerobic and anaerobic cases, respectively. The value for the coefficient of importance for biomass can then be interpreted as the maximum fraction of cellular resources that are diverted to biomass formation. We find that the maximum possible values of the biomass CoI’s for the aerobic and anaerobic cases are 0.58 and 0.68, respectively, as shown by Figure 4.5. This means that biomass maximization appears to be an important descriptor of the observed flux distributions, but not the unique one, given that 0.58 and 0.68 are not equal to one. No other flux has a coefficient of importance nearly as high as the one identified for biomass formation. Interestingly, the relative magnitude of the coefficients for all other reaction fluxes remain similar to their original values except for the metabolic byproduct
coefficients, \( xsuc \) and \( xeth \), which are equal to zero for the aerobic case, and the coefficients for \( xace \) and \( nadh \) dissipation which drop out under both growth conditions.

It is important to note, however, that while for a given flux distribution the range of allowable values for the coefficients of importance is rather narrow, the converse is not true. In fact, the maximization of the sum of the CoI based weighted fluxes (i.e., \( \sum_j c_j v_j \) subject to stoichiometric balances accepts many different flux distributions as optimal solutions as a consequence of the degeneracy of the LP optimization problem. Therefore, while the experimental fluxes do constitute optimal solutions to the optimization problem, many other alternate optimal solutions exist with or without biological meaning. This implies that the original experimental flux distribution cannot be unambiguously recovered based solely on the values of the coefficients of importance. The same holds whenever the maximization of biomass, ATP, or any other resource is adopted a priori. Also note that simple inspection of the ratio \( \frac{v^{\text{exp}}}{v^{\text{max}}} \), where \( v^{\text{max}} \) is the maximum theoretical value of the specified flux subject to the stoichiometric constraints, does not reveal the trends that are uncovered with the CoI’s. Specifically, the ratio \( \frac{v_{\text{biomass}}^{\text{exp}}}{v_{\text{biomass}}^{\text{max}}} \), where \( v_{\text{biomass}}^{\text{exp}} \) is the maximum biomass formation with all fluxes equal to their experimental values, is only 0.058 and 0.120 for the aerobic and anaerobic cases, respectively, whereas the CoI’s for biomass clearly reveals their importance. This implies that the proposed framework is more robust to deviations in the biomass composition from the Ingraham (1983) approximation.
4.4 Conclusions

In this work, a quantitative framework termed ObjFind based on a bilevel optimization procedure is developed for testing the consistency of different hypothesized objective functions with experimentally determined flux distributions. This method was applied to identify the coefficients of importance for *E. coli* flux distributions under aerobic and anaerobic growth conditions. These coefficients were remarkably similar indicating a single cellular driving force governing the distribution of metabolic fluxes. In addition, surprisingly little flexibility was present in the CoI’s for both cases among alternate optimal solutions identified when the sum-squared deviation was equal to its minimum value, although some flexibility was observed in these coefficients as the sum-squared deviation was allowed to increase. We also found that the maximization of the aggregate biomass flux is consistent with the observed experimental flux values. Thus the maximization of cellular biomass appears to be an important descriptor, though not the unique one, in explaining the observed fluxes. Finally, significant degeneracy was found among optimal solutions to the linear programming problem maximizing $\sum_j c_j v_j$ subject to the network stoichiometry. This implied that the flux distributions through the network cannot be uniquely defined based solely on the identified CoI’s although they do provide insight as to which fluxes, when maximized, are consistent with the experimental flux data. It should be noted that while one can never prove the existence of a universal objective function, the ObjFind procedure provides an unbiased framework for researchers to test the validity of different hypotheses leading to a better characterization of the underlying driving forces of cellular metabolism.
Table 4.1: Experimentally (Schmidt et al. 1999a) determined flux distributions for aerobic and anaerobic conditions.

| Reaction* | Aerobic | Anaerobic† | | difference |
|-----------|---------|------------|-----------------|
| pts       | 115.0   | 115.0      | 0.0%            |
| hxi       | 61.0    | 30.0       | 50.8%           |
| gdh       | 53.1    | 84.9       | 59.8%           |
| ald       | 89.1    | 84.9       | 4.8%            |
| tk1       | 18.9    | 28.2       | 49.1%           |
| tal       | 18.9    | 28.2       | 49.1%           |
| tk2       | 9.6     | 26.9       | 179.9%          |
| eno       | 181.2   | 194.6      | 7.4%            |
| pyk       | 26.6    | 69.1       | 159.9%          |
| pdh       | 126.1   | 111.4      | 11.7%           |
| cis       | 52.0    | 4.8        | 90.8%           |
| akd       | 45.0    | 4.7        | 89.6%           |
| ppc       | 21.4    | 6.1        | 71.4%           |
| xace      | 0.0     | 34.1       | -               |
| xeth      | 0.0     | 65.3       | -               |
| xlac      | 0.0     | 69.8       | -               |
| xsuc      | 0.0     | 3.9        | -               |
| coo‡      | 254.9   | 199.6      | 21.7%           |
| baca‡     | 74.1    | 5.9        | 91.3%           |
| bakg‡     | 7.0     | 0.1        | 98.4%           |
| be4p‡     | 9.3     | 1.3        | 84.7%           |
| bf6p‡     | 0.3     | 0.2        | 27.2%           |
| bg6p‡     | 0.9     | 0.3        | 63.6%           |
| bgap‡     | 6.5     | 1.7        | 71.4%           |
| boaa‡     | 14.4    | 2          | 84.8%           |
| bpep‡     | 18.1    | 3.8        | 77.1%           |
| bpyr‡     | 15.3    | 2.8        | 80.0%           |
| br5p‡     | 5.7     | 1.5        | 71.3%           |

* Reactions corresponding to abbreviations are provided in Figure 4.2.
† Flux distributions for the anaerobic case are scaled so that glucose uptake rate is identical to the aerobic case.
‡ Fluxes were obtained from completely defined mass balances around the metabolites of interest.
Figure 4.1: Weak duality states that any feasible dual solution has an objective value that is greater than the optimal primal objective, while any feasible primal solution has an objective value that is less than the optimal dual objective. Strong duality states that if the primal has an optimal solution, so does the dual, and their respective optimal objective function values are equal.
Figure 4.2: The reaction fluxes allowed to assume non-zero coefficients of importance are shown in color. The reactions with similar CoI magnitudes are denoted by the same colors. Note the magnitudes of the CoI’s are similar for both the aerobic and anaerobic growth conditions.
Figure 4.3: The values of the coefficients of importance for the aerobic (■) and anaerobic (□) experimental flux distributions.
Figure 4A

Figure 4B

**Figure 4.4:** The maximum allowable increase of the coefficients of importance (CoI’s) for the (A) aerobic and (B) anaerobic experimental flux distributions when solution optimality is relaxed. MIN = the minimum
Figure 4.5: The coefficients of importance for the aerobic (■) and anaerobic (□) experimental flux distributions with the addition of a biomass flux.
Chapter 5

Gene Knockout Strategies for Microbial Strain Optimization

5.1 Background

The systematic development of engineered microbial strains for optimizing the production of chemicals or biochemicals is an overarching challenge in biotechnology (Stephanopoulos et al. 1998). However, in the absence of metabolic and genetic engineering interventions, the product yields of many microorganisms are often far below their theoretical maximums. This is expected because cellular metabolism is primed, through natural selection, for the maximum responsiveness to the history of selective pressures rather than for the overproduction of specific chemical compounds. Not surprisingly, the behavior of metabolic networks is governed by internal cellular objectives which are often in direct competition with chemical overproduction targets. In this chapter, a bilevel optimization framework termed OptKnock is developed for suggesting gene knockout strategies for biochemical overproduction while recognizing that metabolic flux distributions are governed by internal cellular objectives. Here we explore two such objectives specifically the maximization of biomass yield and the minimization of metabolic adjustment (MOMA).

The recent explosion of annotated sequence information along with a wealth of chemical literature has enabled the reconstruction of genome-scale metabolic networks for many microorganisms (Edwards and Palsson 2000a; Schilling and Palsson 2000; Schilling et al. 2002; Forster et al. 2003). This information, used in the context of the flux
balance analysis (FBA) modeling framework (Varma and Palsson 1993b), has been employed extensively to explore the integrated functions of metabolic networks (Burgard and Maranas 2001; Burgard et al. 2001; Papin et al. 2003; Price et al. 2003). FBA models typically invoke the optimization of a particular cellular objective (e.g., ATP production (Majewski and Domach 1990; Ramakrishna et al. 2001), biomass formation (Varma and Palsson 1993b; Varma and Palsson 1994a), minimization of metabolic adjustment (Segre et al. 2002)), subject to network stoichiometry, to suggest a likely flux distribution. Stoichiometric models of *Escherichia coli* metabolism utilizing the biomass maximization hypothesis have been in some cases successful at (i) predicting the lethality of gene knockouts (Edwards and Palsson 2000a; Badarinarayana et al. 2001), (ii) identifying the correct sequence of byproduct secretion under increasingly anaerobic conditions (Varma et al. 1993b), and (iii) quantitatively predicting cellular growth rates under certain conditions (Edwards et al. 2001). Interestingly, recent work suggests that even when FBA predictions under the biomass maximization assumption seem to fail, metabolic networks can be evolved, for certain cases, towards maximum growth (i.e., biomass yield) through adaptive evolution (Ibarra et al. 2002).

The ability to investigate the metabolism of single-cellular organisms at a genomic scale, and thus systemic level, motivates the need for novel computational methods aimed at identifying strain engineering strategies. In this chapter, we introduce the OptKnock framework for suggesting gene deletion strategies leading to the overproduction of specific chemical compounds in *E. coli*. This is accomplished by ensuring that the production of the desired chemical becomes an obligatory byproduct of growth by “shaping” the connectivity of the metabolic network. In other words, OptKnock identifies
and subsequently removes metabolic reactions that are capable of uncoupling cellular growth from chemical production. The computational procedure is designed to identify not just straightforward but also non-intuitive knockout strategies by simultaneously considering the entire *E. coli* metabolic network as abstracted in the *in silico* *E. coli* model of Palsson and coworkers (Edwards and Palsson 2000a). The complexity and built-in redundancy of this network (e.g., the *E. coli* model encompasses 720 reactions) necessitates a systematic and efficient search approach to combat the combinatorial explosion of candidate gene knockout strategies.

The nested optimization framework shown in Figure 5.1 is developed to identify multiple gene deletion combinations that maximally couple cellular growth objectives with externally imposed chemical production targets. This multi-layered optimization structure involving two competing optimal strategists (i.e., cellular objective and chemical production) is referred to as a bilevel optimization problem (Bard 1998). Problem formulation specifics along with an elegant solution procedure drawing upon linear programming (LP) duality theory are described in the following section. The OptKnock procedure is applied to succinate, lactate, and 1,3-propanediol (PDO) production in *E. coli* with the maximization of the biomass yield for a fixed amount of uptaken glucose employed as the cellular objective. The obtained results are also contrasted against using the minimization of metabolic adjustment (Segre et al. 2002) as the cellular objective. Based on the OptKnock framework, we identify the most promising gene knockout strategies and their corresponding allowable envelopes of chemical versus biomass production in the context of succinate, lactate, and PDO production in *E. coli*. 
5.2 The OptKnock Bilevel Optimization Framework

The maximization of a cellular objective quantified as an aggregate reaction flux for a steady state metabolic network comprising a set \( N = \{1, \ldots, N\} \) of metabolites and a set \( M = \{1, \ldots, M\} \) of metabolic reactions fueled by a glucose substrate is expressed mathematically as follows,

\[
\text{maximize} \quad v_{\text{cellular objective}} \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \q
from the network are members of $M_{\text{secr\_only}}$. Note also that the complete set of reactions $M$ is subdivided into reversible $M_{\text{rev}}$ and irreversible $M_{\text{irrev}}$ reactions. The cellular objective is often assumed to be a drain of biosynthetic precursors in the ratios required for biomass formation (Neidhardt and Curtiss 1996). The fluxes are reported per $1\ \text{g}DW\cdot\text{hr}$ such that biomass formation is expressed as $g\ \text{biomass produced}/\text{g}DW\cdot\text{hr}$ or $1/\text{hr}$.

The modeling of gene deletions, and thus reaction elimination, first requires the incorporation of binary variables into the flux balance analysis framework (Burgard and Maranas 2001; Burgard et al. 2001). These binary variables,

$$y_j = \begin{cases} 
1 & \text{if reaction flux } v_j \text{ is active} \\
0 & \text{if reaction flux } v_j \text{ is not active}
\end{cases}, \forall j \in M$$

assume a value of one if reaction $j$ is active and a value of zero if it is inactive. The following constraint,

$$v_j^{\text{min}} \cdot y_j \leq v_j \leq v_j^{\text{max}} \cdot y_j, \forall j \in M$$

ensures that reaction flux $v_j$ is set to zero only if variable $y_j$ is equal to zero. Alternatively, when $y_j$ is equal to one, $v_j$ is free to assume any value between a lower $v_j^{\text{min}}$ and an upper $v_j^{\text{max}}$ bound. In this study, $v_j^{\text{min}}$ and $v_j^{\text{max}}$ are identified by minimizing and subsequently maximizing every reaction flux subject to the constraints from the Primal problem.

The identification of optimal gene/reaction knockouts requires the solution of a bilevel optimization problem that chooses the set of reactions that can be accessed ($y_j = 1$) so as the optimization of the cellular objective indirectly leads to the overproduction of the chemical or biochemical of interest (see also Figure 5.1). Using biomass formation as
the cellular objective, this is expressed mathematically as the following bilevel mixed-integer optimization problem

\[
\begin{align*}
\text{maximize} & \quad v_{\text{chemical}} (\text{OptKnock}) \\
\text{subject to} & \quad \begin{cases} 
\text{maximize} & \quad v_{\text{biomass}} (\text{Primal}) \\
& \text{subject to } \sum_{j=1}^{M} S_g v_j = 0, \quad \forall i \in \mathbb{N} \\
& v_{\text{pts}} + v_{\text{glk}} = v_{\text{glc}_{\text{uptake}}} \\
& v_{\text{atp}} \geq v_{\text{atp}_{\text{main}}} \\
& v_{\text{biomass}} \geq v_{\text{biomass}}^{\text{target}} \\
& v_j^{\text{min}} \cdot y_j \leq v_j \leq v_j^{\text{max}} \cdot y_j, \quad \forall j \in M \\
\end{cases}
\end{align*}
\]

\[y_j = \{0,1\}, \quad \forall j \in M\]

\[
\sum_{j \in M} (1 - y_j) \leq K
\]

where \(K\) is the number of allowable knockouts. The final constraint ensures that the resulting network meets a minimum biomass yield, \(v_{\text{biomass}}^{\text{target}}\).

The direct solution of this two-stage optimization problem is intractable given the high dimensionality of the flux space (i.e., over 700 reactions) and the presence of two nested optimization problems. To remedy this, we develop an efficient solution approach borrowing from LP duality theory which shows that for every linear programming problem (primal) there exists a unique optimization problem (dual) whose optimal objective value is equal to that of the primal problem. A similar strategy was employed by (Burgard and Maranas 2003) for identifying/testing metabolic objective functions from metabolic flux data. The dual problem (Ignizio and Cavalier 1994) associated with the OptKnock inner problem is
minimize \( v_{\text{atp.main}} \cdot \mu_{\text{atp}} + v_{\text{arg.et}}^{\text{biomass}} \cdot \mu_{\text{biomass}} + v_{\text{glc.uptake}} \cdot \text{glc} \quad \text{(Dual)} \)

subject to

\[
\sum_{i=1}^{N} \lambda_{i}^{\text{stoich}} S_{i,\text{glk}} + \mu_{\text{glk}} + \text{glc} = 0
\]

\[
\sum_{i=1}^{N} \lambda_{i}^{\text{stoich}} S_{i,\text{pts}} + \mu_{\text{pts}} + \text{glc} = 0
\]

\[
\sum_{i=1}^{N} \lambda_{i}^{\text{stoich}} S_{i,\text{biomass}} + \mu_{\text{biomass}} = 1
\]

\[
\sum_{i=1}^{N} \lambda_{i}^{\text{stoich}} S_{ij} + \mu_{j} = 0, \quad \forall j \in \mathbb{M}, j \neq \text{glk, pts, biomass}
\]

\[
\mu_{j}^{\text{min}} \cdot (1 - y_{j}) \leq \mu_{j} \leq \mu_{j}^{\text{max}} \cdot (1 - y_{j}), \quad \forall j \in \mathbb{M}_{\text{rev}} \text{ and } j \notin \mathbb{M}_{\text{secr_only}}
\]

\[
\mu_{j} \geq \mu_{j}^{\text{min}} \cdot (1 - y_{j}), \quad \forall j \in \mathbb{M}_{\text{rev}} \text{ and } j \notin \mathbb{M}_{\text{secr_only}}
\]

\[
\mu_{j} \leq \mu_{j}^{\text{max}} \cdot (1 - y_{j}), \quad \forall j \in \mathbb{M}_{\text{irrev}} \text{ and } j \notin \mathbb{M}_{\text{secr_only}}
\]

\[
\mu_{j} \in \mathbb{R}, \quad \forall j \in \mathbb{M}_{\text{irrev}} \text{ and } j \in \mathbb{M}_{\text{secr_only}}
\]

\[
\lambda_{i}^{\text{stoich}} \in \mathbb{R}, \quad \forall j \in \mathbb{N}
\]

\[
\text{glc} \in \mathbb{R}
\]

where \( \lambda_{i}^{\text{stoich}} \) is the dual variable associated with the stoichiometric constraints, \( \text{glc} \) is the dual variable associated with the glucose uptake constraint, and \( \mu_{j} \) is the dual variable associated with any other restrictions on its corresponding flux \( v_{j} \) in the Primal. Note that the dual variable \( \mu_{j} \) acquires unrestricted sign if its corresponding flux in the OptKnock inner problem is set to zero by enforcing \( y_{j} = 0 \). The parameters \( \mu_{j}^{\text{min}} \) and \( \mu_{j}^{\text{max}} \) are identified by minimizing and subsequently maximizing their values subject to the constraints of the Dual problem.

If the optimal solutions to the Primal and Dual problems are bounded, their objective function values must be equal to one another at optimality. This means that every optimal solution to both problems can be characterized by setting their objectives equal to one another and accumulating their respective constraints. Thus the bilevel formulation for
OptKnock shown previously can be transformed into the following single-level MILP

\[
\begin{align*}
\text{maximize} & \quad v_{\text{chemical}} & (\text{OptKnock}) \\
\text{subject to} & \quad v_{\text{biomass}} = v_{\text{atp}\_\text{main}} \cdot \mu_{\text{atp}} + v_{\text{arg}} \cdot \mu_{\text{arg}} + v_{\text{glc\_uptake}} \cdot glc \\
& \quad \sum_{j=1}^{M} S_j v_j = 0, \quad \forall i \in N \\
& \quad v_{\text{pts}} + v_{\text{glk}} = v_{\text{glc\_uptake}} \text{ mmol/gDW-hr} \\
& \quad v_{\text{atp}} \geq v_{\text{atp}\_\text{main}} \text{ mmol/gDW-hr} \\
& \quad \sum_{i=1}^{N} \lambda_i^{\text{stoich}} S_{i,\text{glk}} + \mu_{\text{glk}} + glc = 0 \\
& \quad \sum_{i=1}^{N} \lambda_i^{\text{stoich}} S_{i,\text{pts}} + \mu_{\text{pts}} + glc = 0 \\
& \quad \sum_{i=1}^{N} \lambda_i^{\text{stoich}} S_{i,\text{biomass}} + \mu_{\text{biomass}} = 1 \\
& \quad \sum_{j=1}^{N} \lambda_i^{\text{stoich}} S_{i,\text{glk}} + \mu_{j} = 0, \quad \forall j \in M, j \neq \text{glk, pts, biomass} \\
& \quad \sum_{j=1}^{N} (1 - y_j) \leq K \\
& \quad v_{\text{biomass}} \geq v_{\text{arg}}^{\text{stoich}} \\
& \quad \mu_j^{\text{min}} \cdot (1 - y_j) \leq \mu_j \leq \mu_j^{\text{max}} \cdot (1 - y_j), \quad \forall j \in M_{\text{rev}} \text{ and } j \notin M_{\text{secr\_only}} \\
& \quad \mu_j \geq \mu_j^{\text{min}} \cdot (1 - y_j), \quad \forall j \in M_{\text{rev}} \text{ and } M_{\text{secr\_only}} \\
& \quad \mu_j \leq \mu_j^{\text{max}} \cdot (1 - y_j), \quad \forall j \in M_{\text{irrev}} \text{ and } j \notin M_{\text{secr\_only}} \\
& \quad \mu_j \in \mathbb{R}, \quad \forall j \in M_{\text{irrev}} \text{ and } M_{\text{secr\_only}} \\
& \quad v_j^{\text{min}} \cdot y_j \leq v_j \leq v_j^{\text{max}} \cdot y_j, \quad \forall j \in M \\
& \quad \lambda_i^{\text{stoich}} \in \mathbb{R}, \quad \forall j \in N \\
& \quad glc \in \mathbb{R} \\
& \quad y_j = \{0,1\}, \quad \forall j \in M
\end{align*}
\]

An important feature of the above formulation is that if the problem is feasible, the optimal solution will always be found. In this thesis, the candidates for gene knockouts
included all reactions of glycolysis, the TCA cycle, the pentose phosphate pathway, respiration, and all anaplerotic reactions. This is accomplished by limiting the number of reactions included in the summation (i.e., \( \sum_{j \in \text{Central Metabolism}} (1 - y_j) = K \)). Problems containing as many as 100 binary variables were solved in the order of minutes to hours using CPLEX 7.0 accessed via the GAMS modeling environment on an IBM RS6000-270 workstation.

### 5.3 Succinate and Lactate Knockouts Strategies

In this section, we identify which reactions, if any, can be removed from the *E. coli* K-12 stoichiometric model (Edwards and Palsson 2000a) so as the remaining network produces succinate or lactate whenever biomass maximization is a good descriptor of flux allocation. For this study, a prespecified amount of glucose (10 mmol/gDW⋅hr), along with unconstrained uptake routes for inorganic phosphate, oxygen, sulfate, and ammonia are provided to fuel the metabolic network. The optimization step could opt for or against the phosphotransferase system, glucokinase, or both mechanisms for the uptake of glucose. Secretion routes for acetate, carbon dioxide, ethanol, formate, lactate and succinate are also enabled. Note that because the glucose uptake rate is fixed, the biomass and product yields are essentially equivalent to the rates of biomass and product production, respectively. In all cases, the OptKnock procedure eliminated the oxygen uptake reaction pointing at anaerobic growth conditions consistent with current succinate (Zeikus et al. 1999) and lactate (Datta et al. 1995) fermentative production strategies.

Table 5.1 summarizes three of the identified gene knockout strategies for succinate overproduction (*i.e.*, mutants A, B, and C). The anaerobic flux distributions at the maximum biomass yields for the complete *E. coli* network (*i.e.*, wild-type), mutant B,
and mutant C are illustrated in Figure 5.2A-C. The results for mutant A suggest that the removal of two reactions (i.e., pyruvate formate lyase and lactate dehydrogenase) from the network results in succinate production reaching 63% of its theoretical maximum at the maximum biomass yield. This knockout strategy is identical to the one employed by Stols and Donnelly (Stols and Donnelly 1997) in their succinate overproducing *E. coli* strain. Next, the envelope of allowable succinate versus biomass production is explored for the wild-type *E. coli* network and the three mutants listed in Table 5.1. Note that the succinate production limits, shown in Figure 5.3A, reveal that mutant A does not exhibit coupled succinate and biomass formation until the yield of biomass approaches 80% of the maximum. Mutant B, however, with the additional deletion of acetaldehyde dehydrogenase, results in a much earlier coupling of succinate with biomass yields.

A less intuitive strategy is identified for mutant C which focuses on inactivating two PEP consuming reactions rather than eliminating competing byproduct (i.e., ethanol, formate, and lactate) production mechanisms. First, the phosphotransferase system is disabled requiring the network to rely exclusively on glucokinase for the uptake of glucose. Next, pyruvate kinase is removed leaving PEP carboxykinase as the only central metabolic reaction capable of draining the significant amount of PEP supplied by glycolysis. This strategy, assuming that the maximum biomass yield could be attained, would result in a succinate yield approaching 88% of the theoretical maximum. In addition, Figure 5.3A reveals significant succinate production for every attainable biomass yield, while the maximum theoretical yield of succinate is the same as that for the wild-type strain.
The OptKnock framework was next applied to identify knockout strategies for coupling lactate and biomass production. Table 5.1 shows three of the identified gene knockout strategies (i.e., mutants A, B, and C) and the flux distribution of mutant C at the maximum biomass yield is shown in Figure 5.2D. Mutant A redirects flux toward lactate at the maximum biomass yield by blocking acetate and ethanol production. This result is consistent with previous work demonstrating that an \textit{adh, pta} mutant \textit{E. coli} strain could grow anaerobically on glucose by producing lactate (Gupta and Clark 1989). Mutant B provides an alternate strategy involving the removal of an initial glycolysis reaction along with the acetate production mechanism. This results in a lactate yield of 90% of its theoretical limit at the maximum biomass yield. The vertical red line for mutant B in Figure 5.3B indicates that the network could avoid producing lactate while maximizing biomass formation. This is due to the fact that OptKnock does not explicitly account for the “worst-case” alternate solution. We are in the process of developing an alternative formulation that safeguards against this. Note that upon the additional elimination of the glucokinase and ethanol production mechanisms, mutant C exhibits a tighter coupling between lactate and biomass production.

5.4 1,3-Propanediol Knockout Strategies

In addition to devising optimum gene knockout strategies, OptKnock can be used to design strains where gene additions are needed along with gene deletions such as in PDO production in \textit{E. coli}. Although microbial 1,3-propanediol (PDO) production methods have been developed utilizing glycerol as the primary carbon source (Hartlep et al. 2002; Zhu et al. 2002), the production of 1,3-propanediol directly from glucose in a single microorganism has recently attracted considerable interest (Cameron et al. 1998; Biebl et
al. 1999; Zeng and Biebl 2002). Because wild-type *E. coli* lacks the pathway necessary for PDO production, we first employed the gene addition framework (Burgard and Maranas 2001) to identify the additional reactions needed for producing PDO from glucose in *E. coli*. The gene addition framework identified a straightforward three-reaction pathway involving the conversion of glycerol-3-P to glycerol by glycerol phosphatase, followed by the conversion of glycerol to 1,3 propanediol by glycerol dehydratase and 1,3-propanediol oxidoreductase. These reactions are then added to the *E. coli* stoichiometric model and the OptKnock procedure is subsequently applied.

OptKnock reveals that there is neither a single nor a double deletion mutant with coupled PDO and biomass production. However, we identified one triple and multiple quadruple knockout strategies that can couple PDO production with biomass production. Two of these knockout strategies are shown in Table 5.1. The results suggest that the removal of certain key functionalities from the *E. coli* network results in PDO overproducing mutants for growth on glucose. Specifically, Table 5.1 reveals that the removal of two glycolytic reactions along with an additional knockout preventing the degradation of glycerol yields a network capable of reaching 72% of the theoretical maximum yield of PDO at the maximum biomass yield. Note that the glyceraldehyde-3-phosphate dehydrogenase (*gapA*) knockout was used by DuPont in their PDO-overproducing *E. coli* strain (Nakamura 2002). Mutant B reveals an alternative strategy, involving the removal of the triose phosphate isomerase (*tpi*) enzyme exhibiting a similar PDO yield and a 38% higher biomass yield. Interestingly, a yeast strain deficient in triose phosphate isomerase activity was recently reported to produce glycerol, a key precursor to PDO, at 80-90% of its maximum theoretical yield (Compagno et al. 1996).
The flux distributions of the wild-type *E. coli*, mutant A, and mutant B networks that maximize the biomass yield are available in Figure 5.4. Not surprisingly, further conversion of glycerol to glyceraldehyde is disrupted in both mutants A and B. For mutant A, the removal of two reactions from the top and bottom parts of glycolysis results in a nearly complete inactivation of the pentose phosphate and glycolysis (with the exception of triose phosphate isomerase) pathways. To compensate, the Entner-Doudoroff glycolysis pathway is activated to channel flux from glucose to pyruvate and glyceraldehyde-3-phosphate (GAP). GAP is then converted to glycerol which is subsequently converted to PDO. Energetic demands lost with the decrease in glycolytic fluxes from the wild-type *E. coli* network case, are now met by an increase in the TCA cycle fluxes. The knockouts suggested for mutant B redirect flux toward the production of PDO by a distinctly different mechanism. The removal of the initial pentose phosphate pathway reaction results in the complete flow of metabolic flux through the first steps of glycolysis. At the fructose bisphosphate aldolase junction, the flow is split into the two product metabolites: dihydroxyacetone-phosphate (DHAP) which is converted to PDO and GAP which continues through the second half of the glycolysis. The removal of the triose-phosphate isomerase reaction prevents any interconversion between DHAP and GAP. Interestingly, a fourth knockout is predicted to retain the coupling between biomass formation and chemical production. This knockout prevents the “leaking” of flux through a complex pathway involving 15 reactions that together convert ribose-5-phosphate (R5P) to acetate and GAP.

Next, the envelope of allowable PDO production versus biomass yield is explored for the two mutants listed in Table 5.1. The production limits of the mutants along with the
original *E. coli* network, illustrated in Figure 5.5, reveal that the wild-type *E. coli* network has no “incentive” to produce PDO if the biomass yield is to be maximized. On the other hand, both mutants A and B have to produce significant amounts of PDO if any amount of biomass is to be formed given the reduced functionalities of the network following the gene removals. Mutant A, by avoiding the *tpi* knockout that essentially sets the ratio of biomass to PDO production, is characterized by a higher maximum theoretical yield of PDO. The above described results hinge on the use of glycerol as a key intermediate to PDO. Next, we explore the possibility of utilizing an alternative to the glycerol conversion route for 1,3-propandediol production.

Based on a literature search we identified a pathway in *Chloroflexus aurantiacus* involving a two-step NADPH-dependant reduction of malonyl-CoA to generate 3-hydroxypropionic acid (3-HPA) (Menendez et al. 1999; Hugler et al. 2002). 3-HPA could then be subsequently converted chemically to 1,3 propanediol given that, to our knowledge, there is no biological functionality to achieve this transformation. This pathway offers a key advantage over PDO production through the glycerol route because its initial step (acetyl-CoA carboxylase) is a carbon fixing reaction. Accordingly, the maximum theoretical yield of 3-HPA (1.79 mmol/mmol glucose) is considerably higher than for PDO production through the glycerol conversion route (1.34 mmol/mmol glucose). The application of the OptKnock framework upon the addition of the 3-HPA production pathway reveals that many more knockouts are required before biomass formation is coupled with 3-HPA production. One of the most interesting strategies involves nine knockouts yielding 3-HPA production at 91% of its theoretical maximum at optimal growth. The first three knockouts are relatively straightforward as they involve
removal of competing acetate, lactate, and ethanol production mechanisms. In addition, the Entner-Doudoroff pathway (either phosphogluconate dehydratase or 2-keto-3-deoxy-6-phosphogluconate aldolase), four respiration reactions (i.e., NADH dehydrogenase I, NADH dehydrogenase II, glycerol-3-phosphate dehydrogenase, and the succinate dehydrogenase complex), and an initial glycolysis step (i.e., phosphoglucone isomerase) are disrupted. This strategy results in a 3-HPA yield that, assuming the maximum biomass yield, is 69% higher than the previously identified mutants utilizing the glycerol conversion route.

5.5 Minimization of Metabolic Adjustment Comparison

All results described in the previous section were obtained by invoking the maximization of biomass yield as the cellular objective that drives flux allocation. This hypothesis essentially assumes that the metabolic network could arbitrarily change and/or even rewire regulatory loops to maintain biomass yield maximality under changing environmental conditions (maximal response). Recent evidence suggests that this is sometimes achieved by the K-12 strain of *E. coli* after multiple cycles of growth selection (Ibarra et al. 2002). In this section, we examine a contrasting hypothesis (i.e., minimization of metabolic adjustment (MOMA) (Segre et al. 2002) that assumes a myopic (minimal) response by the metabolic network upon gene deletions. Specifically, the MOMA hypothesis suggests that the metabolic network will attempt to remain as close as possible to the original steady state of the system rendered unreachable by the gene deletion(s). This hypothesis has been shown to provide a more accurate description of flux allocation immediately after a gene deletion event (Segre et al. 2002). Figure 5.6 pictorially shows the two differing new steady states predicted by the two hypotheses,
respectively. For this study, we utilize the MOMA objective to predict the flux distributions in the mutant strains identified by OptKnock. The base case for the lactate and succinate simulations was assumed to be maximum biomass formation under anaerobic conditions, while the base case for the PDO simulations was maximum biomass formation under aerobic conditions. The results are shown in the last column of Table 5.1. In all cases, the suggested multiple gene knock-out strategy suggests only slightly lower chemical production yields for the MOMA case compared to the maximum biomass hypothesis. This implies that the OptKnock results are fairly robust with respect to the choice of cellular objective.

5.6 Conclusions

In this chapter, the OptKnock framework was described for suggesting gene deletions strategies that could lead to chemical production in *E. coli* by ensuring that the drain towards metabolites/compounds necessary for growth resources (i.e., carbons, redox potential, and energy) must be accompanied, due to stoichiometry, by the production of the desired chemical. Therefore, the production of the desired product becomes an obligatory byproduct of cellular growth. Specifically, OptKnock pinpoints which reactions to remove from a metabolic network, which can be realized by deleting the gene(s) associated with the identified functionality. The procedure was demonstrated based on succinate, lactate, and PDO production in *E. coli* K-12. The obtained results exhibit good agreement with strains published in the literature. While some of the suggested gene deletions are quite straightforward, as they essentially prune reaction pathways competing with the desired one, many others are at first quite non-intuitive reflecting the complexity and built-in redundancy of the metabolic network of *E. coli*. For
the succinate case, OptKnock correctly suggested anaerobic fermentation and the removal of the phosphotranferase glucose uptake mechanism as a consequence of the competition between the cellular and chemical production objectives, and not as a direct input to the problem. In the lactate study, the glucokinase-based glucose uptake mechanism was shown to decouple lactate and biomass production for certain knockout strategies. For the PDO case, results show that the Entner-Doudoroff pathway is more advantageous than EMP glycolysis despite the fact that it is substantially less energetically efficient. In addition, the so far popular tpi knockout was clearly shown to reduce the maximum yields of PDO while a complex network of 15 reactions was shown to be theoretically possible of "leaking" flux from the PPP pathway to the TCA cycle and thus decoupling PDO production from biomass formation. The obtained results also appeared to be quite robust with respect to the choice for the cellular objective.

It is important to note that the suggested gene deletion strategies must be interpreted carefully. For example, in many cases the deletion of a gene in one branch of a branched pathway is equivalent with the significant up-regulation in the other. In addition, inspection of the flux changes before and after the gene deletions provides insight as to which genes need to be up or down-regulated. Lastly, the problem of mapping the set of identified reactions targeted for removal to its corresponding gene counterpart is not always uniquely specified. Therefore, careful identification of the most economical gene set accounting for isozymes and multifunctional enzymes needs to be made.

Currently, in the OptKnock framework, the substrate uptake flux (i.e., glucose) is assumed to be 10 mmol/gDW-hr. Therefore, all reported chemical production and biomass formation values are based upon this postulated and not predicted uptake
scenario. Thus, it is quite possible that the suggested deletion mutants may involve substantially lower uptake efficiencies. However, because OptKnock essentially suggests mutants with coupled growth and chemical production, one could envision a growth selection system that will successively evolve mutants with improved uptake efficiencies and thus enhanced desired chemical production characteristics.

OptKnock so far can only suggests gene deletions as the sole mechanism for chemical overproduction as a consequence of the lack of any regulatory or kinetic information within the purely stoichiometric representation of the inner optimization problem that performs flux allocation. Clearly, the lack of any regulatory or kinetic information in the model is a simplification that may in some cases suggest unrealistic flux distributions. We expect to remedy this limitation by importing regulated *E. coli* models currently under development (Covert et al. 2001b; Covert and Palsson 2002; Covert et al. 2004; Herrgard et al. 2004). The incorporation of regulatory information will not only enhance the quality of the suggested gene deletions by more appropriately resolving flux allocation, but also allow us to suggest regulatory modifications along with gene deletions as mechanisms for strain improvement. The use of alternate modeling approaches (e.g., cybernetic (Kompala et al. 1984; Ramakrishna et al. 1996; Varner and Ramkrishna 1999a), metabolic control analysis (Kacser and Burns 1973; Heinrich and Rapoport 1974; Hatzimanikatis et al. 1998)), if available, could also be incorporated within the OptKnock framework to more accurately estimate the metabolic flux distributions of gene-deleted metabolic networks. Nevertheless, despite its simplifications, OptKnock already provides useful suggestions for strain improvement
and more importantly establishes a systematic framework that will naturally encompass future improvements in metabolic and regulatory modeling frameworks.
Table 5.1: Biomass and chemical yields for various gene knockout strategies identified by OptKnock. The reactions and corresponding enzymes for each knockout strategy are listed. The maximum biomass and corresponding chemical yields are provided on a basis of 10 mmol/hr glucose fed and 1 gDW of cells. The rightmost column provides the chemical yields for the same basis assuming a minimal redistribution of metabolic fluxes from the wild-type (undeleted) E. coli network (MOMA assumption). For the 1,3-propanediol case, glycerol secretion was disabled for both knockout strategies.

### Succinate

<table>
<thead>
<tr>
<th>ID</th>
<th>Knockouts</th>
<th>Enzyme</th>
<th>( \text{max} v_{\text{biomass}} ) (1/hr)</th>
<th>( \text{max} v_{\text{succinate}} ) (mmol/hr)</th>
<th>( \text{min} \sqrt{\sum (v - y)^2} ) Succinate (mmol/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild</td>
<td>'Complete'</td>
<td></td>
<td>0.38</td>
<td>0.12</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1 COA + PYR ( \rightarrow ) ACCOA + FOR</td>
<td>Pymvate formate lyase</td>
<td>0.31</td>
<td>10.70</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>2 NADH + PYR ( \leftrightarrow ) LAC + NAD</td>
<td>Lactate dehydrogenase</td>
<td>0.31</td>
<td>10.70</td>
<td>4.79</td>
</tr>
<tr>
<td>B</td>
<td>1 COA + PYR ( \rightarrow ) ACCOA + FOR</td>
<td>Pymvate formate lyase</td>
<td>0.31</td>
<td>10.70</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>2 NADH + PYR ( \leftrightarrow ) LAC + NAD</td>
<td>Lactate dehydrogenase</td>
<td>0.31</td>
<td>10.70</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>3 ACCOA + 2 NADH ( \rightarrow ) COA + ETH + 2 NAD</td>
<td>Acetaldehyde dehydrogenase</td>
<td>0.16</td>
<td>15.15</td>
<td>6.21</td>
</tr>
</tbody>
</table>

### Lactate

<table>
<thead>
<tr>
<th>ID</th>
<th>Knockouts</th>
<th>Enzyme</th>
<th>( \text{max} v_{\text{biomass}} ) (1/hr)</th>
<th>( \text{max} v_{\text{lacate}} ) (mmol/hr)</th>
<th>( \text{min} \sqrt{\sum (v - y)^2} ) Lactate (mmol/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild</td>
<td>'Complete'</td>
<td></td>
<td>0.38</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1 ACTP + ADP ( \leftrightarrow ) AC + ATP</td>
<td>Acetate kinase</td>
<td>0.28</td>
<td>10.46</td>
<td>5.58</td>
</tr>
<tr>
<td></td>
<td>2 ACCOA + Pi ( \leftrightarrow ) ACTP + COA</td>
<td>Acetaldehyde dehydrogenase</td>
<td>0.28</td>
<td>10.46</td>
<td>5.58</td>
</tr>
<tr>
<td>B</td>
<td>1 ACTP + ADP ( \leftrightarrow ) AC + ATP</td>
<td>Acetate kinase</td>
<td>0.13</td>
<td>18.00</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>2 ATP + F6P ( \rightarrow ) ADP + FDP</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>0.13</td>
<td>18.00</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>FDP ( \leftrightarrow ) T3P1 ( \equiv ) T3P2</td>
<td>Acetaldehyde dehydrogenase</td>
<td>0.13</td>
<td>18.00</td>
<td>0.19</td>
</tr>
<tr>
<td>C</td>
<td>1 ACTP + ADP ( \leftrightarrow ) AC + ATP</td>
<td>Acetate kinase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>2 ACCOA + Pi ( \leftrightarrow ) ACTP + COA</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>3 ATP + F6P ( \rightarrow ) ADP + FDP</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>FDP ( \leftrightarrow ) T3P1 ( \equiv ) T3P2</td>
<td>Fructose-1,6-bisphosphatase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>2 ACCOA + 2 NADH ( \rightarrow ) COA + ETH + 2 NAD</td>
<td>Acetaldehyde dehydrogenase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>4 GLC + ATP ( \rightarrow ) G6P + PEP</td>
<td>Glucose kinase</td>
<td>0.12</td>
<td>18.13</td>
<td>10.53</td>
</tr>
</tbody>
</table>

### 1,3-Propanediol

<table>
<thead>
<tr>
<th>ID</th>
<th>Knockouts</th>
<th>Enzyme</th>
<th>( \text{max} v_{\text{biomass}} ) (1/hr)</th>
<th>( \text{max} v_{\text{1,3-PD}} ) (mmol/hr)</th>
<th>( \text{min} \sqrt{\sum (v - y)^2} ) 1,3-PD (mmol/hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild</td>
<td>'Complete'</td>
<td></td>
<td>1.06</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>1 FDP ( \rightarrow ) F6P + Pi</td>
<td>Fructose-1,6-bisphosphate</td>
<td>0.16</td>
<td>9.66</td>
<td>8.66</td>
</tr>
<tr>
<td></td>
<td>2 13PDG + ADP ( \leftrightarrow ) 3PG + ATP</td>
<td>Fructose-1,6-bisphosphate</td>
<td>0.21</td>
<td>9.66</td>
<td>8.66</td>
</tr>
<tr>
<td></td>
<td>NAD + Pi ( \leftrightarrow ) T3P1 ( \equiv ) 13PDG + NADH</td>
<td>Fructose-1,6-bisphosphate</td>
<td>0.21</td>
<td>9.66</td>
<td>8.66</td>
</tr>
<tr>
<td></td>
<td>3 GL + NAD ( \leftrightarrow ) GL + NADH</td>
<td>Fructose-1,6-bisphosphate</td>
<td>0.21</td>
<td>9.66</td>
<td>8.66</td>
</tr>
<tr>
<td>B</td>
<td>1 T3P1 ( \leftrightarrow ) T3P2</td>
<td>Triose phosphate isomerase</td>
<td>0.13</td>
<td>18.00</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>2 G6P + NADP ( \rightarrow ) DMG6P + NADPH</td>
<td>Glucose-6-phosphate-1-kinase</td>
<td>0.13</td>
<td>18.00</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>DMG6P ( \rightarrow ) T3P1</td>
<td>Glucose-6-phosphate-1-kinase</td>
<td>0.13</td>
<td>18.00</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>3 D3P ( \rightarrow ) ACAL + T3P1</td>
<td>Glucose-6-phosphate-1-kinase</td>
<td>0.13</td>
<td>18.00</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td>4 GL + NAD ( \leftrightarrow ) GL + NADH</td>
<td>Glucose-6-phosphate-1-kinase</td>
<td>0.13</td>
<td>18.00</td>
<td>10.53</td>
</tr>
</tbody>
</table>
Figure 5.1: The bilevel optimization structure of OptKnock. The inner problem performs the flux allocation based on the optimization of a particular cellular objective (e.g., maximization of biomass yield, MOMA, etc.). The outer problem then maximizes the bioengineering objective (e.g., chemical production) by restricting access to key reactions available to the optimization of the inner problem.

\[
\begin{align*}
\text{maximize} & \quad \text{bioengineering objective} \\
& \quad \text{(through gene knockouts)} \\
\text{subject to} & \quad \text{maximize} \quad \text{cellular objective} \\
& \quad \text{(over fluxes)} \\
\text{subject to} & \quad \begin{array}{l}
\circ \quad \text{fixed substrate uptake} \\
\circ \quad \text{network stoichiometry} \\
\circ \quad \text{blocked reactions identified by outer problem} \\
\end{array} \\
& \quad \text{number of knockouts} \leq \text{limit}
\end{align*}
\]
Figure 5.2: The flux distributions of the (A) wild-type E. coli, (B) succinate mutant B, (C) succinate mutant C, and (D) lactate mutant C networks that maximize biomass yield under anaerobic conditions.
Figure 5.3: (A) Succinate or (B) lactate production limits under anaerobic conditions for mutant A, mutant B, mutant C and the wild-type E. coli network. The production limits are obtained by separately maximizing and minimizing succinate or lactate production for the biomass yields available to each network. The points depict the solution identified by OptKnock (i.e., maximum chemical production at the maximum biomass yield).
Figure 5.4: The aerobic flux distributions of the (A) wild-type *E. coli*, (B) mutant A, and (C) mutant B networks that maximize biomass yield. Results for mutants A and B assume the reactions responsible for 1,3-propanediol production are available.
Figure 5.5: 1,3-propanediol (PDO) production limits under aerobic conditions for mutant A, mutant B, and the wild-type *E. coli* network. The points depict the solution identified by OptKnock (i.e., maximum chemical production at the maximum biomass yield).
Figure 5.6: Projection of the multidimensional flux space onto two dimensions. The pink region represents flux ranges potentially reachable by both the mutant and complete networks, while the blue region corresponds to flux distributions rendered unreachable by the gene deletion(s). Point A represents the maximum biomass yield solution. Point B is the solution assuming the minimization of metabolic adjustment hypothesis for the mutant network, while point C is the solution assuming the mutant network will maximize its biomass yield.
Chapter 6

Flux Coupling Analysis of Genome-Scale Metabolic Reconstructions

6.1 Background

An overarching attribute of metabolic networks is their inherent robustness and ability to cope with ever changing environmental conditions. Despite this flexibility, network stoichiometry and connectivity do establish limits/barriers to the coordination and accessibility of reactions. The recent abundance of complete genome sequences has enabled the generation of genome-scale metabolic reconstructions for various microorganisms (Covert et al. 2001a; Price et al. 2003; Reed and Palsson 2003). These models provide a largely complete skeleton of the metabolic reactions present in an organism. Examination of the structural and topological properties of metabolic networks is important at both the conceptual level to reveal the organizational principles of metabolic interactions within cellular networks and at the practical level for more effectively focusing engineering interventions and ensuring the consistency of the underlying reconstructions.

To this end, the identification of blocked reactions (i.e., reactions incapable of carrying flux due to the stoichiometry of the metabolic network under steady-state conditions) and enzyme subsets (i.e., groups of reactions that operate together in fixed flux proportions under steady-state conditions) in metabolic models has attracted considerable interest in recent years (Kholodenko et al. 1995; Rohwer et al. 1996; Pfeiffer et al. 1999; Klamt et al. 2003). The output of these analyses provides significant
biological insight as to which reactions are potentially missing from metabolic models as well as which reactions may be under coordinated regulation alluding to a mechanism for the continuous refinement of metabolic reconstructions through an iterative model building process. Specifically, stoichiometric models of *Escherichia coli* metabolism utilized within the flux balance analysis (FBA) framework have been used for (i) qualitatively predicting the outcomes of gene knockout experiments (Edwards and Palsson 2000a; Badarinarayana et al. 2001), (ii) identifying the correct sequence of byproduct secretion under increasingly anaerobic conditions (Varma et al. 1993b), (iii) quantitatively predicting cellular growth rates (Edwards et al. 2001; Ibarra et al. 2002), (iv) assessing the performance limits of metabolic networks in response to gene additions or deletions (Burgard and Maranas 2001), and (v) suggesting gene knockout strategies for enhancing biochemical production (Burgard et al. 2003; Pharkya et al. 2003).

In the post-genomic era, each cellular function, biological entity, or physiological event is seen in the context of a complex network of interactions. Following this spirit, several frameworks for examining structural and topological network properties based on convex analysis have been developed and demonstrated for small-scale metabolic representations (~100 reactions) to identify extreme pathways (Schilling et al. 2000) or elementary modes (Schuster and Hilgetag 1994; Schuster et al. 2000). An elementary mode refers to a minimal set of enzymes that could operate under steady-state conditions. Any feasible flux distribution can be represented by a non-negative linear combination of elementary modes. Elementary mode analysis is often used as a quantitative measure of network robustness as more elementary modes for a given environmental condition and organism hints at a more flexible metabolism (Stelling et al. 2002). This concept has
proven effective in the rational strain design for poly-\(\beta\)-hydroxybutyrate production in *Saccharomyces cerevisiae* by quantifying the additional flexibility gained by the addition of a non-native transhydrogenase reaction (Carlson et al. 2002). Correspondingly, the set of extreme pathways refers to the minimum set of flux vectors capable of describing all steady-state flux distributions and are consequently a subset of elementary modes (Schilling et al. 2000). As with elementary modes, the number of extreme pathways provides a measure of pathway redundancy. The application of extreme pathway analysis has revealed that the *Haemophilus influenzae* network has an order of magnitude larger degree of pathway redundancy for amino acid production than *Helicobacter pylori* (Papin et al. 2002; Price et al. 2002). For more details, reviews of the similarities, differences, and applicability of elementary mode analysis and extreme pathway analysis have been recently published (Klamt and Stelling 2003; Palsson et al. 2003).

However elegant, all existing algorithms for the exhaustive identification of elementary modes and extreme pathways do not scale well for genome-scale models of complex microorganisms due to the combinatorial explosion of the identified pathways (Klamt and Stelling 2002). For example, while the central metabolic network utilized by (Stelling et al. 2002) contained only 110 reactions, it gave rise to 43,279 elementary modes. Similarly, while a small example involving 20 reactions contained only 80 extreme pathways (Covert and Palsson 2003), the *Haemophilus influenzae* metabolic network contained over 1,000 extreme pathways even after its *a priori* breakup into six distinct metabolic subsystems (Schilling and Palsson 2000).

In this chapter, we introduce the Flux Coupling Finder (FCF) procedure for finding coupled reaction sets and blocked reactions in genome-scale metabolic systems. Whereas
previous algorithms require the computation of null-space matrices (Heinrich and Schuster 1996; Pfeiffer et al. 1999), a computationally prohibitive task for large networks (Golub and Van Loan 1996), the approach proposed here circumvents this challenge by requiring instead the solution of a sequence of linear programs (LPs). The set of blocked reactions for a given network is identified by maximizing each particular flux subject to the network stoichiometry. If the maximum possible value of a particular flux is zero, then the reaction is said to be unusable or blocked because it cannot carry any flux. Similarly, linear fractional programming is employed to identify the maximum and minimum flux ratios (i.e., max $v_1/v_2$, min $v_1/v_2$) for every pair of metabolic fluxes. Comparison of flux ratios allows one to determine if any two fluxes, $v_1$ and $v_2$, share any of the following types of coupling: (i) Directional coupling ($v_1 \rightarrow v_2$), if a non-zero flux for $v_1$ implies a non-zero flux for $v_2$ but not necessarily the reverse; (ii) Partial coupling ($v_1 \leftrightarrow v_2$), if a non-zero flux for $v_1$ implies a non-zero, though variable, flux for $v_2$ and vice-versa; or (iii) Full coupling ($v_1 \Leftrightarrow v_2$), if a non-zero flux for $v_1$ implies not only a non-zero but also a fixed flux for $v_2$ and vice-versa. (see Figure 6.1). Reaction pairs not falling into one of these categories are classified as uncoupled. Reactions which are mutually partially and/or fully coupled to one another are grouped into coupled reaction sets. The identification of directionally and partially coupled reactions in addition to fully coupled reactions implies that the FCF method is not only more tractable but also more encompassing than previously described algorithms for enzyme subset identification. The proposed procedure is versatile enough to allow the incorporation of additional constraints (e.g., oxygen uptake limitations and/or substrate restrictions) or performance requirements (e.g., minimum levels of ATP and/or biomass production) during the
calculation of blocked and/or coupled reactions. The mathematical frameworks for identifying the maximum and minimum flux ratios and blocked reactions are described next followed by their application to the following genome-scale stoichiometric models of increasing size and complexity: (i) *H. pylori* (Schilling et al. 2002) (389 reactions), (ii) *E. coli* (Edwards and Palsson 2000a) (740 reactions), and (iii) *S. cerevisiae* (Forster et al. 2003) (1173 reactions).

### 6.2 Exhaustive Identification of Blocked and Coupled Reactions

**Blocked Reactions**

Blocked reactions are defined as reactions incapable of carrying flux under steady-state conditions. Here we identify blocked reactions by identifying fluxes whose maximum and minimum values are zero for a particular uptake scenario. The maximization of a particular flux $v_j$ for a steady-state metabolic network comprised of a set $N = \{1, \ldots, N\}$ of metabolites and a set $M = \{1, \ldots, M\}$ of reactions is expressed mathematically as the following linear program (LP),

\[
\begin{align*}
\text{maximize} & \quad v_j \\
\text{subject to} & \quad \sum_{j=1}^{M} S_{ij} v_j = 0, \quad \forall i \in N \\
& \quad v_{j_{\text{uptake}}} \leq v_{j_{\text{uptake max}}}, \quad \forall j \in M_{\text{transport}} \\
& \quad v_j \geq 0, \quad \forall j \in M
\end{align*}
\]

where $S_{ij}$ is the stoichiometric coefficient of metabolite $i$ in reaction $j$. Reversible reactions are expressed as two irreversible reactions in opposite directions (i.e., $v_j = v_j^f - v_j^b$) thus constraining all fluxes to positive values. The optimization problem can be easily modified to examine not only the effects of changing which metabolites are internal or external, but more specifically changing which metabolites can be taken up, secreted, or
both. Constraint (3) limits the uptake of resources (i.e., carbon, oxygen, etc.) to the network and the maximum uptake of any metabolite absent from the external medium is set to zero. Transport mechanisms for metabolites out of the cell can be blocked by changing the inequality in constraint (4) to an equality. All isozymes catalyzing a given reaction are lumped into a single flux eliminating duplicate reactions. The set of blocked reactions is identified by solving the above linear programming problem once for every flux. If the maximum value of the flux is zero, then the reaction is said to be unusable or blocked. Examples of blocked reactions are provided in Figure 6.2. The identified blocked reactions may have either biological meaning such as the reaction being a part of an incomplete pathway at an intermediate stage of evolution or they could signify errors/omissions in the metabolic reconstruction. It is important to note that the results depend upon the (i) steady-state assumption, (ii) imposed uptake/secretion scenarios, (iii) growth requirements, and (iv) energy production requirements.

**Coupled Reactions**

The identification of all coupled reactions and corresponding coupled reaction sets hinges upon the calculation of the upper and lower limits of all flux ratios (i.e., $R_{\text{max}} = \max v_1/v_2$, $R_{\text{min}} = \min v_1/v_2$). Note that the calculation of these ratios originally gives rise to nonlinear optimization problems. However, by performing the variable transformation ($\tilde{v} = v \cdot t$) inspired by fractional programming, a completely equivalent (see Appendix A for proof) linear programming formulation is obtained whose solution time is on the order of milliseconds.
maximize \( R_{\text{max}} = \tilde{v}_1 \) or (minimize \( R_{\text{min}} = \tilde{v}_1 \))

subject to \( \sum_{j=1}^{M} S_{ij} \tilde{v}_j = 0, \quad \forall \ i \in N \)

\( \tilde{v}_2 = 1 \)

\( \tilde{v}_{\text{uptake}}^j \leq v_{\text{uptake max}}^j \cdot t, \quad \forall \ j \in M_{\text{transport}} \)

\( \tilde{v}_j \geq 0, \quad \forall \ j \in M \)

\( t \geq 0 \)

Here the variables \( \tilde{v} \) are the metabolic fluxes normalized by \( v_2 \). The above linear program has a clear biological interpretation in terms of responses of metabolic networks to the perturbation of particular fluxes. Constraint \( \tilde{v}_2 = 1 \) sets a reference flux to a unit value while the optimization criteria is used to probe flux variability for each tested reaction. Uncoupled fluxes do not “feel” flux perturbations while fluxes through coupled reactions decrease or increase in accordance with the encountered type of coupling. This formulation is guaranteed to obtain globally optimal values for the flux ratios since it relies on linear programming.

The various outcomes for the maximum and minimum flux ratios are depicted in Figure 6.1. The first case occurs whenever \( R_{\text{min}} \) is equal to zero and \( R_{\text{max}} \) is equal to some finite value \( c \). The fluxes are directionally coupled \( (v_1 \rightarrow v_2) \) because the activity of \( v_1 \) implies \( v_2 \) (i.e., \( v_2 \geq v_1 / c \)). Similarly, if \( R_{\text{min}} \) is equal to a finite constant \( c \) and \( R_{\text{max}} \) is unbounded, then the fluxes are directionally coupled in the opposite direction \( (v_2 \rightarrow v_1) \) as \( v_2 \) implies \( v_1 \) (i.e., \( v_1 \geq v_2 \cdot c \)). Two fluxes can also be partially coupled \( (v_1 \leftrightarrow v_2) \) if \( R_{\text{max}} \) and \( R_{\text{min}} \) are both finite and unequal, or fully coupled \( (v_1 \Leftrightarrow v_2) \) if \( R_{\text{max}} \) is finite and equal.
to $R_{\min}$. The final case occurs whenever the two fluxes are completely uncoupled and is encountered if their ratio can vary freely from zero to infinity.

Figure 6.2 shows an example of a fully coupled reaction set. Note that because the partial and/or full coupling of reactions is a transitive property (i.e., $v_1 \leftrightarrow v_2$ and $v_2 \leftrightarrow v_3$ imply that $v_1 \leftrightarrow v_3$), complete coupled reaction sets can be subsequently inferred from the maximum and minimum flux ratios. Directional coupling, unlike partial and full coupling, can capture the one-way type of connectivity between metabolic reactions. This information enables the global identification of *equivalent knockouts* defined as the set of all possible reactions whose deletion forces the flux through a particular reaction to zero and sets of *affected reactions* defined as all reactions whose fluxes are forced to zero if a particular reaction is deleted. These concepts are illustrated in Figure 6.3 where the reactions $v_1$, $v_2$, and $v_3$ all imply reaction $v^*$. This means that if any of these fluxes assumes a non-zero value, then $v^*$ must also attain a non-zero value. Therefore, knocking out reaction $v^*$ from the network forces the fluxes through reactions $v_1$, $v_2$, and $v_3$ to zero. Thus, we refer to reactions $v_1$, $v_2$, and $v_3$ as the set of reactions *affected* by the removal of $v^*$. Similarly, a non-zero flux through $v^*$ implies that the fluxes through $v_4$, $v_5$, and $v_6$ are also non-zero. This means that removing any of $v_4$, $v_5$, or $v_6$ from the network forces the flux through $v^*$ to zero. Reactions $v_4$, $v_5$, and $v_6$ are thus referred to as *equivalent knockouts* for $v^*$. Note that while the sets of reactions affected by each of the equivalent knockouts may differ, the directionality of coupling for partially/fully coupled reactions remains the same.
6.3 Flux Coupling Analysis Algorithm

Although the identification of all blocked and coupled reactions by inspection is possible for small networks such as the one shown in Figure 6.2, exhaustively identifying blocked and coupled reactions in genome-scale metabolic models requires a rigorous computational procedure. The developed FCF procedure for identifying both blocked and coupled reactions is summarized in pseudo-code as follows:

**Step 1.** Aggregate all isozymes (i.e., duplicate reactions) from the stoichiometric matrix $S_{ij}$.

**Step 2.** For ($j = 1$ to $M$),

- Solve for the maximum value of each flux $v_j$.
  - If $v_j^{\text{max}} = 0$, then the reaction is blocked. Column $j$ is removed from $S_{ij}$.

**Step 3.** Set $\text{AlreadyCoupled}(j) = 0$.

**Step 4.** For ($j = 1$ to $M - 1$) and $\text{AlreadyCoupled}(j) = 0$,

- For ($j' = j + 1$ to $M$),
  - Solve problem (2) for $R_{\text{min}}$ and $R_{\text{max}}$ (i.e., the minimum and maximum ratios).
    - **A.** If $R_{\text{min}} = 0$ and $R_{\text{max}}$ is unbounded, then the reactions are uncoupled.
    - **B.** If $R_{\text{min}} = 0$ and $R_{\text{max}} = c > 0$, then ($v_j \rightarrow v_{j'}$).
    - **C.** If $R_{\text{min}} = c_1 > 0$ and $R_{\text{max}} = c_2 > 0$, then
      - a. If $(c_2 - c_1) > 0$, then ($v_j \leftrightarrow v_{j'}$).
      - b. If $(c_2 - c_1) = 0$, then ($v_j \leftrightarrow v_{j'}$).
  - Set $\text{AlreadyCoupled}(j') = 1$.
  - Reactions $j$ and $j'$ belong to the same coupled reaction set.
  - All subsequent partially/fully coupled reactions $j'$ are added to the coupled reaction set.
  - **D.** If $R_{\text{min}} = c > 0$ and $R_{\text{max}}$ is unbounded, then ($v_{j'} \rightarrow v_j$).

End

The array $\text{AlreadyCoupled}(j)$ enables the outer loop of Step 4 to skip reactions that have already been found to be a part of a coupled reaction set. All reactions in the same coupled reaction set have identical coupling properties. Computational requirements are in the order of minutes for genome-scale models involving as many as 1,173 reactions.
upon implementing the FCF procedure utilizing LINDO (Lindo Systems, Inc.) accessed via C++ on an Intel Pentium IV, 2.4 GHz, 512 MB RAM computer. Note that the FCF procedure substantially reduces the number of reaction ratios to be calculated by employing a number of key tests as described above. Typically, only 10-45% of potential reaction pairs needs to be examined.

6.4 Blocked Reaction Identification

Here we examine the percentage of blocked reactions in the genome-scale models of *H. pylori* (Schilling et al. 2002) (389 reactions), *E. coli* (Edwards and Palsson 2000a) (740 reactions), and *S. cerevisiae* (Forster et al. 2003) (1173 reactions). For each one of them, five separate scenarios are explored to examine the effect of varied external/internal conditions on the number of potentially active fluxes in the models. First, any metabolite is allowed to enter or leave the metabolic network provided that there exists a corresponding transport mechanism into or out of the cell for that metabolite (i.e., Complex Media/Aerobic). Reactions blocked under this assumption are unconditionally blocked and cannot be active under any set of conditions. Second, aerobic growth on a glucose minimal media is explored (i.e., Glucose/Aerobic). For this condition, the uptake of any carbon source other than what is required for growth on glucose is not allowed. Next, we further constrain the second scenario by setting the oxygen uptake to zero (i.e., Glucose/Anaerobic). Likewise, the fourth and fifth scenarios are identical to the second and third except that we enforce that the network apportions its fluxes to maximize the biomass yield (i.e., Optimal Glucose/Aerobic and Optimal Glucose/Anaerobic). Reactions that are blocked under any of the last four conditions but can carry flux under the first condition are referred to as conditionally blocked.
The total numbers and percentages of blocked reactions in each model are summarized in Figure 6.4. As expected, the percentage of blocked reactions increases as more constraints/conditions are imposed on the three network models. For the *E. coli* metabolic network (Edwards and Palsson 2000a), we find that 14% of the 740 reactions are blocked under any condition (i.e., all metabolites with a transport mechanism into or out of the cell can be consumed or secreted, respectively) while 28% are blocked for aerobic growth on a glucose minimal media. Restricting the oxygen uptake to zero blocks only a few additional reactions (i.e., the oxygen uptake reaction and cytochrome oxidases) for growth on glucose. For both aerobic and anaerobic growth on glucose, about 55% of the *E. coli* reactions are found incapable of carrying any flux if the network is to attain the maximum biomass yield on a glucose minimal media. Notably, enforcing biomass maximization reveals distinct differences between aerobic and anaerobic *E. coli* metabolism. Specifically, optimal anaerobic growth requires that the 2-ketoglutarate dehydrogenase reaction is blocked preventing the cyclic operation of the TCA cycle. In addition, optimal aerobic growth prevents the activity of reactions involved in fermentation product formation such as pyruvate formate lyase (i.e., formate production) and acetaldehyde dehydrogenase (i.e., ethanol production) in contrast to optimal anaerobic growth.

For the yeast model (Forster et al. 2003), a much larger percentage of reactions are blocked under all examined scenarios. Many more reactions are blocked under anaerobic conditions than aerobic conditions for both growth on glucose and optimal growth on glucose. This is due to the large number of yeast reactions utilizing oxygen as a substrate, particularly in ergosterol and zymosterol synthesis, which are biomass constituents for
yeast but not \textit{E. coli}. Finally, for \textit{H. pylori}, far fewer reactions are blocked as compared to the \textit{E. coli} and \textit{S. cerevisiae} networks implying a much more compact and largely essential network. The complete lists of reactions blocked under the examined conditions are available upon request.

6.5 Coupled Reaction Sets

In this section, we identify all coupled sets of reactions for the three metabolic networks. First, coupled reaction sets are identified assuming a biomass reaction exists for draining the set of compounds necessary for cell growth in their pre-specified biological ratios. This aggregate biomass reaction description imposes a constant biomass composition. We then relax the constant biomass composition assumption by removing the centralized biomass drain from each metabolic network while allowing all biomass components to be drained independently of one another. The biomass components are slightly different for the three stoichiometric models (Edwards and Palsson 2000a; Schilling et al. 2002; Forster et al. 2003) although they all include similar lists of amino acids, cofactors, currency metabolites, etc. True cellular behavior is bound to reside between the two extremes of a fixed biomass composition and uncontrolled biomass component draining. For each case, we investigate two distinct uptake scenarios: (i) only glucose can be taken up in addition to all essential growth resources (i.e., glucose minimal media) and (ii) any metabolites with a transport mechanism into the cell can be consumed (i.e., complex media). Reactions that are coupled for the most general case (i.e., complex media uptake environment, no biomass reaction) are coupled under any set of conditions. For \textit{H. pylori}, the glucose minimal media consists of glucose, alanine, arginine, histidine, isoleucine, methionine, phenylalanine, valine, thiamine, phosphate,
oxygen, and sulfate as determined by Schilling et al. (2002). For *E. coli* and *S. cerevisiae*, the glucose minimal media is comprised of glucose, sulfate, nitrate, phosphate, and oxygen. For *S. cerevisiae*, the biomass components, ergosterol and zymosterol, are also included in the glucose minimal media under anaerobic conditions as the model lacks the pathways necessary to synthesize these compounds in the absence of oxygen.

**Genome-scale Identification of Coupled Reaction Sets**

The numbers of coupled reaction sets for each organism under the different conditions are provided in Table 6.1. These entries denote the total numbers of both partially or fully coupled reactions in each set. The complete lists of coupled reaction sets are available upon request. For all cases, we find that a much higher percentage of reactions are members of coupled sets in *H. pylori* than for the larger and more complex *E. coli* and *S. cerevisiae* networks (see Figure 6.5) alluding to a much more flexible metabolism in the larger networks. Additionally, we find that a constant biomass composition leads to the generation of one large coupled reaction set. Thus, if the biomass composition is fixed, the fluxes through tens of reactions are “locked” due to stoichiometry. In fact, the biomass-coupled reaction set of *H. pylori* comprises 38% and 46% of the entire network for the complex and glucose minimal media, respectively. Also, the size of this biomass-coupled reaction set is much larger for *H. pylori* than for *E. coli* and *S. cerevisiae* as increasing network redundancy leads to the decoupling of reactions from the large biomass coupled reaction set. Note that although the coupled reaction sets include both partially and fully coupled reactions, the reactions comprising the biomass reaction sets are almost exclusively fully coupled. In addition, the handful of partially coupled reactions can vary only within tight ranges.
**Coupled Reaction Set Example: Purine Biosynthesis in E. coli**

A representative example of the information gained from flux coupling analysis is shown in Figure 6.6 with a coupled reaction set for *E. coli* purine biosynthesis. This reaction set is identified for aerobic growth on a glucose minimal media assuming a constant biomass composition. Here the concepts of partially coupled, fully coupled, and uncoupled reactions can be more clearly discerned. The numbers indicate the relative values or range of values for each flux in any particular flux distribution for the examined conditions. The FCF framework identifies ten coupled reactions, eight fully coupled and two partially coupled. This coupled reaction set encompasses two multi-gene operons: purDH (three reactions with EC#’s 6.3.4.13, 2.1.2.3 and 3.5.4.10) and purEK (two reactions with EC# 4.1.1.21). The two reactions converting AICAR to IMP are partially coupled to the rest of the reaction set because AICAR is also formed during histidine biosynthesis. Therefore the partially coupled reactions can assume values slightly greater than the eight fully coupled reactions. If we relax the constant biomass composition assumption, the coupled reaction set is “broken up” into two fully coupled reaction sets of two and eight reactions, respectively. Interestingly, the two reactions capable of converting GAR to FGAR are not a part of this coupled reaction set because they decouple one another by offering alternative conversion routes. However, the sum of their fluxes is coupled to the rest of the reaction set. Also, both fluxes are directionally coupled to the other reactions because a non-zero flux through either one implies that the coupled reaction set carries flux.
Biomass Formation Coupled Reaction Sets

In all cases, the biomass coupled reaction sets are broken up into smaller sets whenever the biomass reaction is replaced by independent drains of biomass precursors. This decomposition is expected given that allowing biomass constituents to be drained (i.e., allowing them to become external metabolites) adds degrees of freedom to the network leading to the uncoupling of reaction sets. Note that considering certain highly-connected metabolites (e.g., ATP, NADH, etc.) as external helps to decompose large metabolic networks into smaller subsystems for elementary mode analysis (Schuster et al. 2002).

The decomposition of the yeast biomass coupled reaction set is shown in Figure 6.7. It is comprised of 34 reactions (including the biomass reaction) for growth on glucose. When the biomass reaction is removed and replaced with biomass component drains which could operate independently of one another, the 34 reaction set is decomposed into one 5, 17, and 2 member reaction sets as well as two 3 member reaction sets. The biomass set for E. coli and H. pylori undergo similar fragmentations as the biomass reaction is removed from the model. However, even though the coupled reaction sets are decoupled under uncontrolled biomass component draining, they are indeed linked to one another based on the cell’s requirement to provide itself with biomass constituents for growth.

Genome-scale versus Subsystem-based Coupling Analysis

A key advantage of the FCF framework over previous methods is that it does not require the a priori decoupling of the metabolic network into subsystems for analysis. For example, enzyme subset identification for the H. pylori model performed by Schilling et
al. (2002) using extreme pathway analysis required breaking the network into six metabolic subsystems: (1) amino acid biosynthesis and degradation, (2) central metabolism, (3) lipid and cell envelope biosynthesis, (4) nucleotide biosynthesis and degradation, (5) transport and energy-redox metabolism, and (6) vitamin and cofactor biosynthesis. Forty-nine fully coupled enzyme subsets were identified spanning the six subsystems. This study corresponds exactly to our *H. pylori* complex media case without the presence of a biomass reaction. A complete comparison of the coupled reaction sets identified using FCF with the enzyme subsets identified by Schilling et al. (2002) is shown in Figure 6.8. In addition to reproducing the enzyme subsets of Schilling et al. (2002), flux coupling analysis reveals additional information about the coupling of enzymes across the putative functional classifications. For example, the (GLCD, GLLDHR, KATA) transport subset is fully coupled with the (FOLE, DNTPH, DHPPH, FOLB, FOLK, PABB, PABC, FOLP, FOLC) vitamin and cofactor subset. Also, fully coupled with this reaction set is the ACEB reaction from central metabolism bringing the total number of coupled reactions in this set to thirteen. Thus breaking the metabolic network into subsystems often leads to reactions being missed during enzyme subset identification. Entire enzyme subsets can also be missed if they are comprised of no more than one reaction from each subsystem. One such example is the enzyme subset of OOR_ and FRDO which is comprised of one reaction from central metabolism and one reaction from transport and energy-redox metabolism, respectively. The results demonstrate the importance of investigating metabolism at the genome-scale as many connections between seemingly unrelated subsystems are uncovered.
6.6 Directional Coupling of Metabolic Reactions

In the previous section, we focused on partially and fully coupled flux pairs. Here we highlight results for directional coupling where a non-zero flux through certain reactions implies non-zero fluxes through others but not necessarily the reverse. First, we utilize the FCF procedure to detect coupled reaction sets, affected reactions, and equivalent knockouts in *E. coli* central metabolism for aerobic growth on glucose. We then extend our analysis to the genome-scale models of *H. pylori*, *E. coli*, and *S. cerevisiae* to identify the essential core of reactions required for biomass formation on a glucose minimal media and to examine the topological features of the identified reaction flux connected networks linked through directional coupling.

*E. coli Central Metabolism*

The coupling interactions for *E. coli* central metabolism are depicted pictorially in Figure 6.9. Due to the significant amount of redundant connectivity in central metabolism, no large fully coupled reaction sets are found. Instead, we identify seven fully coupled sets of two reactions (i.e., EDD/EDA, ZWF/PGL, ACEA/ACEB, GAP/PGK, GPM/ENO, PTA/ACK, GLT/ACN) corresponding to consecutive reactions in the network. We also find that the forward and backward directions of glycolysis, the pentose phosphate pathway, and the TCA cycle are completely disconnected from one another although Figure 6.9 reveals a significant amount of internal coupling between the various reactions within each pathway. In addition, while Entner-Doudoroff glycolysis (i.e., EDD and EDA) is dependant on the activity of the forward direction of the pentose phosphate pathway, the anaplerotic and respiration reactions are not coupled with any reactions of the three major central metabolic pathways.
We next examine how FCF-derived directionality data along with knowledge of partially and fully coupled reactions enables the identification of missing elements of a metabolic reconstruction. Specifically, consider the set of reaction fluxes in Figure 6.9 which must be forced to zero under the steady-state assumption if the ZWF reaction is removed from the network. As explained previously, PGL is forced to zero if ZWF is knocked out because the two reaction fluxes are fully coupled for aerobic growth on glucose. The FCF procedure also identifies five additional functionalities (i.e., EDD, EDA, GND, RPE_F, and TKT2_F) comprising the Entner Doudoroff pathway and part of the pentose phosphate pathway which are eliminated upon the removal of ZWF. Interestingly, however, a recent study demonstrated that an *E. coli* mutant lacking glucose-6-phosphate dehydrogenase activity (ZWF) had residual activity through the Entner Doudoroff and/or pentose phosphate pathway accounting for 7% of glucose metabolized (Fischer and Sauer 2003). Here, flux coupling analysis is useful by pinpointing where the metabolic reconstruction may be incomplete. Specifically, the “bypass” of the ZWF reaction could be explained by a glucose dehydrogenase reaction which enables the Entner Doudoroff and pentose phosphate pathways to operate even without the ZWF functionality (Fischer and Sauer 2003). Accordingly, the most recent *E. coli* stoichiometric model (Reed et al. 2003) includes this reaction, which is absent from the *E. coli* model used in this work (Edwards and Palsson 2000a).

Flux coupling analysis can also be used to identify all equivalent knockouts or multiple targets for the removal of a particular reaction. For example, suppose one wants to prevent the pentose phosphate pathway reaction TKT2 from carrying flux in the forward direction. The FCF procedure identifies six functionalities (i.e., ZWF_F, PGL,
GND, RPE_F, TKT1_F, and TALB_F) which could be alternatively eliminated to prevent TKT2_F from carrying flux under steady-state conditions. In the TCA cycle, the FCF procedure finds that the removal of GLTA or ACN_F ensures that the ICD_F reaction carries no flux. This is interesting because the elimination of isocitrate dehydrogenase (ICD_F) prevents *E. coli* growth on a glucose minimal media (Helling and Kukora 1971). Thus flux coupling analysis correctly points that the citrate synthase (GLTA) (Lakshmi and Helling 1976) or aconitase (ACN_F) (Gruer et al. 1997) mutations are also lethal for *E. coli* growth on glucose because they prevent the activity of isocitrate dehydrogenase.

**Essential Reaction Core for Biomass Formation**

The FCF procedure can be used to identify the essential core of metabolic reactions necessary for biomass formation for a given environmental condition. In an earlier work, it was shown that the minimal set of *E. coli* reactions needed to support various levels of growth is a strong function of the uptake environment (Burgard et al. 2001). While the minimum number of reactions required for a given biomass yield is constant, there exist myriads of alternate minimal reaction sets having the same number of reactions due to network redundancy. Using FCF, we can identify the conserved core of reactions present in all these minimal reaction sets. Specifically, all reactions which are either partially (*v_{biomass} \leftrightarrow v_j*), fully (*v_{biomass} \Leftrightarrow v_j*), or directionally coupled (*v_{biomass} \rightarrow v_j*) to biomass production are essential for cellular growth. Overall, FCF determined the percentage of reactions in the essential core for aerobic growth on a glucose minimal media to be 59% (229 reactions), 28% (206 reactions), and 14% (166 reactions) for *H. pylori*, *E. coli*, and *S. cerevisiae*, respectively. These data are available upon request. Note that the size of the
essential core of reactions is smaller than the minimal reaction sets (e.g., the minimal reaction set for *E. coli* growth on glucose contains 224 reactions (Burgard et al. 2001)). This is because non-unique, though necessary, functions are essential for growth in each network. This set of non-unique required reactions for growth is larger in the more complex *S. cerevisiae* and *E. coli* networks than in *H. pylori* due to their inherent flexibility.

*Scaling Properties of Directional Coupling*

The connectivity of the directional couplings of the three metabolic networks is examined by constructing reaction maps where nodes correspond to metabolic functionalities and arcs denote the presence and directionality of the coupling between reactions. It is important to note that unlike the study of Barabasi and Albert (1999), here nodes denote metabolic functionalities not metabolites. While metabolites cannot be “deleted” from a network, metabolic functionalities can be eliminated by deleting the appropriate gene or genes. Therefore, the vulnerability of the network to gene deletions can now be directly assessed. Reactions associated with coupled reaction sets are lumped together into super-nodes, one per partially/fully coupled set, because their directional coupling relationships are equivalent (see Figures 6.3 and 6.9). Specifically, we examine whether the directional coupling between metabolic reactions is scale-free, characterized by a relatively small number of well-connected nodes, or random, where the number of arcs associated with each node follows a Poisson distribution.

The number $N(k)$ of nodes/reactions implying a certain number of $k$ reactions is plotted in Figure 6.10 for *H. pylori*, *E. coli*, and *S. cerevisiae* growth on a glucose minimal media. We find that the connectivity of the three reaction maps is consistent
with that of scale-free networks as the number of nodes implying k other nodes decreases exponentially with k (Barabasi and Albert 1999). In all cases, the correlation exponents are less than two meaning that if the networks continue to expand through evolution, the total number of directional flux couplings will grow faster than the total number of reactions and no finite value can characterize the average coupling degree. These characteristics were encountered previously for many other types of investigated networks (Dorogovtsev and Mendes 2003). Thus, not only do the static features of metabolic networks (i.e., the connectivities of their metabolites) exhibit a scale-free topology (Edwards and Palsson 1999; Jeong et al. 2000; Wagner and Fell 2001), but also the stoichiometry driven couplings linking the individual reactions conform to a scale-free architecture. Interestingly, while the distribution of vertex degrees in the reaction-centered graph investigated by Wagner and Fell (2001) does not follow a power law, we find that the distribution of vertex degrees in the reaction flux-centered graphs does.

The genome-wide coupling between metabolic reactions for *E. coli* growth on a glucose minimal media is shown in Figure 6.11 with and without the presence of a centralized biomass drain. Visual inspection of Figure 6.11 reveals many fundamental organizational principles of mature scale-free networks: the existence of a giant component consisting of nodes interconnected with short paths, hubs dominating the topology, and inhomogeneity and clustering features. Note that the presence of the biomass drain reaction, shown in the bottom left-hand corner of Figure 6.11A, is responsible for connecting a large percentage of the metabolic network through directional coupling.
6.7 Conclusions

In this chapter, we introduced the Flux Coupling Finder (FCF) procedure for identifying blocked and coupled reactions in genome-scale metabolic models. This identification can be made under different environmental conditions and growth demands. The approach is based on the successive solution of linear programming problems which allows it to remain tractable for large metabolic networks involving many hundreds or even thousands of reactions. Consequently, the FCF method does not require breaking large metabolic networks into smaller subnetworks to identify coupled reaction sets as this *a priori* clustering does indeed miss various couplings. It is important to note that flux coupling analysis identifies not only fully coupled reactions but also directionally and partially coupled reactions unlike other approaches that focused on pinpointing enzyme subsets composed of only fully coupled reactions. As was shown earlier, partially as well as directionally coupled reactions are equally important to track.

The FCF procedure was applied to the three stoichiometric models of *H. pylori*, *E. coli*, and *S. cerevisiae* to provide a detailed analysis of their topological features. It was determined that 10%, 14%, and 29% of their respective reactions are blocked unconditionally. Furthermore, we found that the optimal growth of the larger networks involves a much higher percentage of blocked reactions. The percentage of reactions in coupled sets decreases substantially with model size alluding to the greater flexibility and redundancy inherent in the larger models of *E. coli* and *S. cerevisiae*. Unlike blocked reactions, the partial and full coupling of reactions was found to be rather condition independent as the uptake conditions barely affect the percentage of reactions in coupled
sets. Anaerobic conditions also had little impact on the coupling of reactions (results not shown).

Flux coupling analysis also revealed that postulating a constant biomass composition leads to the coupling of large sets of reactions in all three organisms. Interestingly, the size of the biomass-coupled subset is much larger in *H. pylori* than in *E. coli* or *S. cerevisiae* as the more complex networks have the inherent flexibility to decouple the production of various biomass precursors from one another. For all three networks, allowing the biomass components to be drained independently of one another breaks these large sets of biomass coupled reactions into smaller sets associated with a particular biomass precursor. The FCF procedure also led to the identification of the essential core of reactions whose activity is required for cellular growth for a given condition. This allows the lethality of any knockout to be quickly evaluated by examining the essential core of metabolic reactions required for cellular growth under the condition of interest. Reactions contained in this set are essential for biomass production for the examined condition and thus their deletion is predicted to be fatal.

Concurrently, directional coupling data obtained from FCF were used to detect sets of affected reactions and equivalent knockouts in *E. coli* central metabolism for growth on glucose. Specifically it was revealed that the forward and backward directions of glycolysis, the pentose phosphate pathway, and the TCA cycle are not stoichiometrically coupled to one another for growth on glucose. Although FCF does not make quantitative phenotypic predictions for a particular knockout, it allows the identification of all reactions forced to zero following the removal of a certain network function. In addition, FCF suggests multiple targets for removing a particular metabolic reaction by locating
equivalent knockouts and allows an immediate assessment of the consequences (i.e., affected reactions) of implementing any one of the candidate deletions. Finally, it was shown that not only the static but also the systemic features of metabolic networks captured with directional coupling exhibit a scale-free topology. Furthermore, the exponential correlation between \( k \) and \( N(k) \) improved substantially with increasing network size suggesting that as network complexity/size increases, they are driven towards a scale-free architecture which is more resistant to random attacks (i.e., mutations). In this respect, it is important to mention that while general graph-theoretic approaches predict the functional vulnerability of free-scale networks to the removal of hubs (the most highly-connected vertices), FCF allows for the classification between essential and nonessential hubs as essential hubs are implied by many other reactions.

The FCF procedure can be used in both aiding metabolic reconstructions and guiding genetic manipulations. For example, blocked reactions may signify model omissions or incomplete pathways. Coupling information can be used to suggest multiple avenues for achieving a particular reaction inactivation and to enable the quick assessment of the reactions forced to be inactivated upon a given deletion. The complete coupling characteristics of mutant or regulated networks can be established by reapplying the FCF procedure with the modified stoichiometric relations. Clearly the applications and implications of flux coupling analysis are not limited to those discussed in this thesis. For example, it will be interesting to examine whether the expression levels of genes associated with partially or fully coupled reactions are correlated. Thus, the FCF output may also be used to supplement operon prediction tools as coupled reactions could be under coordinated regulation. Preliminary comparisons of coupled reaction sets with
operons available from the RegulonDB (Salgado et al. 2001) database has revealed that about 30% of coupled reaction sets identified for *E. coli* growth on a complex media include two or more genes from common operons. Almost half of such coupled reaction sets correspond to operons exactly. Due to its wide range of features and applicability to genome-scale networks, the Flux Coupling Finder procedure provides a useful framework for both modelers and experimentalists seeking to extract biologically meaningful information from metabolic reconstructions.

### 6.8 APPENDIX: Flux Ratio Maximization/Minimization Transformation Proof

In this appendix, we show that the nonlinear programming problem of flux ratio maximization or minimization can be recast as the linear programming problem presented previously in the Coupled Reactions section. For any two fluxes, $v_1$ and $v_2$, the maximization or minimization of their respective ratios is described mathematically as

$$\begin{align*}
\text{maximize (or minimize)} \quad & \frac{v_1}{v_2} \\
\text{subject to} \quad & \sum_{j=1}^{M} S_{ij} v_j = 0, \quad \forall \; i \in N \\
& v_j^{\text{uptake}} \leq v_j^{\text{uptake max}}, \quad \forall \; j \in M_{\text{transport}} \\
& v_j \geq 0, \quad \forall \; j \in M
\end{align*}$$

By multiplying the numerator and denominator of the objective function as well as all constraints by a positive variable $t$, an equivalent problem $(\mathcal{P})$ is obtained.

$$\begin{align*}
\text{maximize (or minimize)} \quad & \frac{v_1 \cdot t}{v_2 \cdot t} = \frac{v_1}{v_2} \\
\text{subject to} \quad & \sum_{j=1}^{M} S_{ij} (v_j \cdot t) = 0, \quad \forall \; i \in N \\
& v_j^{\text{uptake}} \cdot t \leq v_j^{\text{uptake max}} \cdot t, \quad \forall \; j \in M_{\text{transport}} \\
& v_j \cdot t \geq 0, \quad \forall \; j \in M \\
& t \geq 0
\end{align*}$$
We next show that the following linear formulation \((P')\) is completely equivalent to problem \((P)\).

\[
\begin{align*}
\text{maximize (or minimize)} & \quad \hat{v}_1 \\
\text{subject to} & \quad \hat{v}_2 = 1 \\
& \quad \sum_{j=1}^{M} S_{ij} \hat{v}_j = 0, \quad \forall \; i \in N \\
& \quad \hat{v}_{\text{uptake}} \leq v_{\text{uptake, max}} \cdot t, \quad \forall \; j \in M_{\text{transport}} \\
& \quad \hat{v}_j \geq 0, \quad \forall \; j \in M \\
& \quad t \geq 0
\end{align*}
\]

Specifically, the goal is to solve \((P')\) to obtain \((\hat{V}, t)\) and have \(v = \hat{V} / t\) solve \((P)\). Note the proof provided below is for the maximization case of the above formulations as the proof for the minimization case can be obtained with the appropriate modifications.

**Theorem:** If there exists an optimal solution to \((P)\) with \(v_2 > 0\) and also there is an optimal solution \((\hat{v}^*, t^*)\) to \((P')\), then \(v^* = \hat{v}^* / t^*\) solves \((P)\).

**Proof:** Because \((\hat{v}^*, t^*)\) is optimal to \((P')\) it follows that \(\hat{v}^* \geq \hat{v}\). By contradiction, suppose that \(v^* = \hat{v}^* / t^*\) does not solve \((P)\) but there exists an optimal solution \(v^*\) to \((P)\).

Therefore, the following three conditions must be satisfied:

1. \(v^*\) must be feasible to \((P)\)
2. \(v_2^* > 0\)
3. \(v_1^*/v_2^* > v_1^*/v_2^*\) because \(v^*\) is suboptimal while \(v^*\) is optimal.

Now let \(\hat{t} = 1 / v_2^*\). Because \(v_2^* > 0\), we can deduce that

(i) \(\hat{t} > 0\).

Also let \(\hat{v} = \hat{t} \cdot v^*\) implying
(ii) $\hat{v}_2 = 1$.

Following from condition (1) above,

(iii) $\sum_{j=1}^{M} S_g \hat{v}_j = 0, \quad \forall i \in N$

$\hat{v}_j^{\text{uptake}} \leq v_j^{\text{uptake max}} t', \quad \forall j \in M_{\text{transport}}$

$\hat{v}_j \geq 0, \quad \forall j \in M$

From (i-iii), we see that $t'$ and $\hat{v}$ are feasible to ($P'$). Finally, from condition (3) we determine that $v_1^* \cdot t' > v_1 \cdot t^*$ because $v_2^* \cdot t = v_2^* \cdot t^* = 1$. Therefore, we recover $\hat{v} > \hat{v}^*$ which contradicts the original assumption.

Therefore,

$v_1^* / v_2^* = (\hat{v}_1 / t^*) / (\hat{v}_2 / t^*) = \hat{v}_1^* \text{ because } \hat{v}_2^* = 1$

confirming that the optimal objective function value to ($P'$) is equivalent to that of ($P$).
Table 6.1: Reaction coupling statistics for the *H. pylori*, *E. coli*, and *S. cerevisiae* metabolic networks. The first number in each entry denotes the number of coupled sets while the number in parenthesis denotes the size of the coupled set. For example, 19(2) implies that there are 19 coupled sets composed of 2 reactions.

<table>
<thead>
<tr>
<th></th>
<th><em>H. pylori</em></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>biomass reaction</td>
<td>glucose minimal</td>
<td>no biomass reaction</td>
<td>glucose minimal</td>
</tr>
<tr>
<td>complex media</td>
<td>19 (2)</td>
<td>15 (2)</td>
<td>25 (2)</td>
<td>23 (2)</td>
</tr>
<tr>
<td>glucose minimal</td>
<td>45 (2)</td>
<td>33 (2)</td>
<td>48 (2)</td>
<td>37 (2)</td>
</tr>
<tr>
<td></td>
<td>8 (3)</td>
<td>7 (3)</td>
<td>10 (3)</td>
<td>9 (3)</td>
</tr>
<tr>
<td></td>
<td>9 (3)</td>
<td>6 (3)</td>
<td>9 (3)</td>
<td>9 (3)</td>
</tr>
<tr>
<td></td>
<td>1 (6)</td>
<td>1 (7)</td>
<td>1 (5)</td>
<td>2 (5)</td>
</tr>
<tr>
<td></td>
<td>3 (5)</td>
<td>1 (5)</td>
<td>5 (5)</td>
<td>5 (5)</td>
</tr>
<tr>
<td></td>
<td>2 (6)</td>
<td>1 (10)</td>
<td>2 (6)</td>
<td>3 (6)</td>
</tr>
<tr>
<td></td>
<td>1 (10)</td>
<td>1 (174)</td>
<td>3 (7)</td>
<td>2 (7)</td>
</tr>
<tr>
<td></td>
<td>1 (8)</td>
<td>1 (8)</td>
<td>1 (8)</td>
<td>1 (8)</td>
</tr>
<tr>
<td></td>
<td>1 (9)</td>
<td>1 (9)</td>
<td>2 (9)</td>
<td>3 (9)</td>
</tr>
<tr>
<td></td>
<td>1 (10)</td>
<td>1 (13)</td>
<td>4 (10)</td>
<td>1 (12)</td>
</tr>
<tr>
<td></td>
<td>1 (13)</td>
<td>1 (13)</td>
<td>1 (66)</td>
<td>1 (17)</td>
</tr>
<tr>
<td></td>
<td>1 (20)</td>
<td>1 (20)</td>
<td>1 (20)</td>
<td></td>
</tr>
<tr>
<td>total reactions</td>
<td>248</td>
<td>247</td>
<td>220</td>
<td>213</td>
</tr>
<tr>
<td>in subsets:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>total subsets:</td>
<td>34</td>
<td>26</td>
<td>52</td>
<td>49</td>
</tr>
</tbody>
</table>

|                  | *E. coli* |                  |                  |                  |
|                  | biomass reaction | glucose minimal | no biomass reaction | glucose minimal |
| complex media    | 8 (3)      | 7 (3)            | 10 (3)           | 9 (3)            |
| glucose minimal  | 9 (3)      | 6 (3)            | 9 (3)            | 9 (3)            |
|                  | 1 (6)      | 1 (7)            | 1 (5)            | 2 (5)            |
|                  | 5 (5)      | 2 (5)            | 3 (5)            | 3 (5)            |
|                  | 2 (6)      | 1 (10)           | 2 (6)            | 3 (6)            |
|                  | 1 (10)     | 1 (174)          | 3 (7)            | 2 (7)            |
|                  | 1 (7)      | 1 (10)           | 2 (7)            | 2 (7)            |
|                  | 1 (7)      | 1 (7)            | 1 (7)            | 1 (7)            |
|                  | 1 (7)      | 1 (7)            | 1 (7)            | 1 (7)            |
|                  | 1 (8)      | 1 (8)            | 1 (8)            | 1 (8)            |
|                  | 1 (9)      | 1 (9)            | 2 (9)            | 3 (9)            |
|                  | 1 (10)     | 1 (13)           | 4 (10)           | 1 (12)           |
|                  | 1 (13)     | 1 (13)           | 1 (66)           | 1 (17)           |
|                  | 1 (20)     | 1 (20)           | 1 (20)           |                  |
| total reactions  | 259         | 236              | 252              | 226              |
| in subsets:      |             |                  |                  |                  |
| total subsets:   | 68          | 46               | 68               | 46               |

|                  | *S. cerevisiae* |                  |                  |                  |
|                  | biomass reaction | glucose minimal | no biomass reaction | glucose minimal |
| complex media    | 13 (3)        | 11 (3)           | 14 (3)           | 14 (3)           |
| glucose minimal  | 6 (4)         | 5 (4)            | 4 (5)            | 4 (5)            |
|                  | 2 (5)         | 3 (5)            | 2 (5)            | 3 (5)            |
|                  | 2 (6)         | 2 (6)            | 2 (6)            | 2 (6)            |
|                  | 1 (7)         | 1 (7)            | 1 (7)            | 1 (7)            |
|                  | 1 (7)         | 1 (7)            | 1 (7)            | 1 (7)            |
|                  | 1 (7)         | 1 (7)            | 1 (7)            | 1 (7)            |
|                  | 1 (8)         | 1 (8)            | 1 (8)            | 1 (8)            |
|                  | 1 (9)         | 1 (9)            | 1 (9)            | 1 (9)            |
|                  | 1 (9)         | 1 (9)            | 1 (9)            | 1 (9)            |
|                  | 1 (10)        | 1 (13)           | 1 (10)           | 1 (17)           |
|                  | 1 (12)        | 1 (12)           | 1 (12)           | 1 (12)           |
|                  | 1 (12)        | 1 (12)           | 1 (12)           | 1 (12)           |
|                  | 1 (13)        | 1 (13)           |                  |                  |
|                  | 1 (17)        | 1 (17)           | 1 (17)           | 1 (17)           |
| total reactions  | 261         | 248              | 255              | 242              |
| in subsets:      |             |                  |                  |                  |
| total subsets:   | 80          | 72               | 82               | 76               |
\[ R_{\text{min}} = \min v_1/v_2 \quad R_{\text{max}} = \max v_1/v_2 \quad \boxed{R_{\text{min}} \leq v_1/v_2 \leq R_{\text{max}}} \]

**Directionally Coupled:** \[ v_1 \rightarrow v_2 \]
- \( R_{\text{min}} = 0 \)
- \( R_{\text{max}} = c \)

**Partially Coupled:** \[ v_1 \leftrightarrow v_2 \]
- \( R_{\text{min}} = c_1 \)
- \( R_{\text{max}} = c_2 \)

**Fully Coupled:** \[ v_1 \leftrightarrow v_2 \]
- \( R_{\text{min}} = R_{\text{max}} = c \)

**Directionally Coupled:** \[ v_2 \rightarrow v_1 \]
- \( R_{\text{min}} = c \)
- \( R_{\text{max}} = \infty \)

**Uncoupled:**
- \( R_{\text{min}} = 0 \)
- \( R_{\text{max}} = \infty \)

**Figure 6.1:** Two reaction fluxes are (1) directionally coupled if the activity of one flux implies the activity of the other without the converse necessarily holding true, (2) partially coupled if the activity of one flux implies the activity of the other and vice versa, or (3) fully coupled if activity of one flux fixes the activity of the other. Reactions in enzyme subsets as defined by (Pfeiffer et al. 1999) are exclusively fully coupled. Various types of coupling are related to the flux ratio limits \( R_{\text{min}} \) and \( R_{\text{max}} \) as shown.
Figure 6.2: Examples of blocked reactions (dashed lines) and a fully coupled enzyme subset (heavy lines). Flux $v_4$ is blocked due to the absence of a reaction consuming metabolite H, while $v_9$ and $v_{10}$ are blocked because there are no reactions forming I or consuming K. Note, however, that $v_4$ can carry flux if metabolite H is allowed to accumulate (i.e., unsteady-state). Assuming that the biomass composition is pre-specified, knowledge of any flux in the enzyme subset confers the values of all other fluxes in that subset. For example, if $v_3$ is fixed, then $v_6$ and $v_{bio}$ are also fixed as they are the only outlets for the flux towards metabolites C and F, respectively. Similarly, fluxes $v_7$ and $v_E$ are fixed as a consequence of fixing $v_{bio}$ and $v_6$. 
Corresponding flux ratio outcomes,

\[ 0 \leq \frac{v_{1,2, or 3}}{v^*} \leq c \quad \quad \quad 0 \leq \frac{v^*/v_{4,5, or 6}}{c} \leq \infty \]

or \[ \frac{c}{\infty} \leq \frac{v^*/v_{1,2, or 3}}{\infty} \leq c \quad \text{or} \quad \frac{c}{\infty} \leq \frac{v^*/v_{4,5, or 6}}{\infty} \leq \infty \]

Figure 6.3: Examples of affected reaction sets and equivalent knockouts for reaction \( v^* \). Removing \( v^* \) from the network results in reaction fluxes \( v_1, v_2, \) and \( v_3 \) being forced equal to zero at steady-state and thus they are referred to as affected by \( v^* \). Removing any of \( v_4, v_5, \) or \( v_6 \) ensures that \( v^* \) cannot carry flux at steady-state so they are said to be equivalent knockouts for \( v^* \).
<table>
<thead>
<tr>
<th></th>
<th>H. pylori 389 runs</th>
<th>E. coli 740 runs</th>
<th>S. cerevisiae 1173 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex Media (Aerobic)</td>
<td>35</td>
<td>103</td>
<td>338</td>
</tr>
<tr>
<td>Glucose (Aerobic)</td>
<td>66</td>
<td>207</td>
<td>460</td>
</tr>
<tr>
<td>Glucose (Anaerobic)</td>
<td>210</td>
<td>515</td>
<td></td>
</tr>
<tr>
<td>Optimal Glucose (Aerobic)</td>
<td>77</td>
<td>408</td>
<td>774</td>
</tr>
<tr>
<td>Optimal Glucose (Anaerobic)</td>
<td>407</td>
<td></td>
<td>791</td>
</tr>
</tbody>
</table>

**Figure 6.4:** Total numbers and percentages of blocked reactions for the three networks under different growth conditions.
Figure 6.5: Percentage of reactions contained in coupled sets in the *H. pylori*, *E. coli*, and *S. cerevisiae* metabolic networks for growth on either a complex or glucose minimal media (with and without a biomass reaction).
**Figure 6.6:** Coupled reaction set identified for purine biosynthesis in *E. coli* on a glucose minimal media assuming a constant biomass composition. The numbers indicate the relative values or range of values for each flux in any particular flux distribution for given growth condition. Secondary metabolites and cofactors are omitted for simplicity.
Figure 6.7: Reactions coupled to biomass formation for aerobic *S. cerevisiae* growth on a glucose minimal media. Secondary metabolites and cofactors are omitted for simplicity. All reactions are fully coupled meaning that knowledge of one reaction flux is sufficient to specify the flux through all reactions at steady-state. Note that PAP is converted to AMP which is a precursor to biomass. This enzyme subset is decomposed into numerous subsystems, indicated by different colored arrows, if the biomass reaction is replaced with drains on the various biomass precursors.
Figure 6.8: Comparison of the FCF-identified coupled reaction sets for *H. pylori* with the enzyme subsets identified by Schilling et al. (2002). The later approach subdivides the network into six smaller sub-networks based on functional classification and finds subsets for each one of them The FCF procedure considers the network in its entirety. The reaction names in each row correspond to different coupled reaction sets. Underlined reactions highlight coupling relationships only identified using the FCF method and the two arrows indicate the coupling of enzyme subsets across functional classifications. Reaction abbreviations can be found in the supplemental material of Schilling et al. (2002).
Figure 6.9: The complete reaction coupling relationships in *E. coli* central metabolism for aerobic growth on glucose. Reversible reactions are listed by the reaction name followed by _F_ and _B_ to denote the forward and backward directions, respectively. The reaction names and stoichiometry corresponding to the reaction abbreviations are found in the Supplementary Material.
Figure 6.10: The number of reactions N(k) implying k other reactions are plotted as a function of k for *Helicobacter pylori*, *Escherichia coli*, and *Saccharomyces cerevisiae* growth on a glucose minimal media.
Figure 6.11: Genome-wide metabolic coupling for *E. coli* growth on a glucose minimal medium (A) with or (B) without the presence of a biomass reaction. The biomass reaction is located in the bottom left corner of (A).
Synopsis

Recent developments in molecular biology and recombinant DNA technology have ushered a new era in the ability to shape the gene content and expression levels for microbial production strains in a direct and targeted fashion. The advent of genome-scale models of metabolism laid the foundation for the development of computational procedures for suggesting genetic manipulations that lead to overproduction. To this end, the primary focus of this thesis work was the development of computational tools to be used in conjunction with genome-scale metabolic models for selecting the optimal sets of genetic modifications for strain optimization projects. In this pursuit, numerous optimization-based analysis methods were also developed and directed towards addressing various questions of substantial scientific importance.

In Chapter 1, an optimization-based procedure for studying the capabilities of metabolic networks after the addition of foreign functionalities was introduced and applied to a linear flux balance analysis *Escherichia coli* model. The developed modeling and optimization framework was tested by investigating the effect of gene additions on the maximum theoretical production of all twenty amino acids for aerobic growth on glucose and acetate substrates. This study revealed that the maximum theoretical production of six amino acids could be improved by the addition of *only one or two* genes to the native amino acid production pathway of *E. coli*, even though the model could have chosen from 3,400 foreign reaction candidates. The mechanism of all suggested
enhancements was either by: (i) improving the energy efficiency and/or (ii) increasing the carbon conversion efficiency of the production route.

Next, a computational procedure for identifying the minimal set of metabolic reactions capable of supporting various growth rates on different substrates was introduced and applied to the *Escherichia coli* metabolic network. The minimal reaction sets capable of supporting specified growth rates were determined for two different uptake conditions (i) limiting the uptake of organic material to a single organic component (*e.g.*, glucose or acetate) and (ii) allowing the importation of any metabolite with available cellular transport reactions. The results revealed the minimal reaction sets to be highly dependent on the imposed uptake environment and the growth requirements.

The ObjFind framework was introduced in Chapter 4 for testing whether experimental flux data are consistent with different hypothesized objective functions. Specifically, it was examined whether the maximization of a weighted combination of fluxes can explain a set of observed experimental data. Coefficients of importance (Col’s) were identified which quantify the fraction of the additive contribution of a given flux to a fitness (objective) function whose optimization explains the experimental flux data. ObjFind was applied to both an aerobic and anaerobic set of *Escherichia coli* flux data derived from isotopomer analysis. Results revealed the Col’s for both growth conditions to be strikingly similar even though the flux distributions for the two cases are quite different. This is consistent with the presence of a single metabolic objective driving the flux distributions in both cases. Interestingly, the Col associated with a biomass production flux, complete with energy and reducing power requirements,
assumed a value nine and fifteen times higher than the next largest coefficient for the aerobic and anaerobic cases, respectively.

In Chapter 5, the computational OptKnock framework was introduced for suggesting gene deletions strategies leading to the overproduction of biochemicals in *E. coli*. This was accomplished by ensuring that a drain towards growth resources (i.e., carbon, redox potential, and energy) must be accompanied, due to stoichiometry, by the production of a desired product. Computational results for gene deletions for succinate, lactate, and 1,3-propanediol (PDO) production were in good agreement with mutant strains published in the literature. While some of the suggested deletion strategies were straightforward and involved eliminating competing reaction pathways, many others suggested complex and non-intuitive mechanisms of compensating for the removed functionalities. More importantly, by directly coupling biochemical production with the basal biomass requirements, the OptKnock procedure hints at a growth selection/adaptation system for indirectly evolving genetically stable overproducing mutants.

The Flux Coupling Finder (FCF) algorithm was introduced in Chapter 6 for elucidating the topological and flux connectivity features of genome-scale metabolic networks. The framework was demonstrated on genome-scale metabolic reconstructions of *Helicobacter pylori, Escherichia coli*, and *Saccharomyces cerevisiae*. This analysis allowed one to determine if any two metabolic fluxes, $v_1$ and $v_2$, are (i) directionally coupled, if a non-zero flux for $v_1$ implies a non-zero flux for $v_2$ but not necessarily the reverse; (ii) partially coupled, if a non-zero flux for $v_1$ implies a non-zero, though variable, flux for $v_2$ and vice-versa; or (iii) fully coupled, if a non-zero flux for $v_1$ implies not only a non-zero but also a fixed flux for $v_2$ and vice-versa. Flux coupling analysis also
enabled the global identification of blocked reactions, which are all reactions incapable of carrying flux under a certain condition, equivalent knockouts, defined as the set of all possible reactions whose deletion forces the flux through a particular reaction to zero, and sets of affected reactions denoting all reactions whose fluxes are forced to zero if a particular reaction is deleted. The FCF approach thus provides a novel and versatile tool for aiding metabolic reconstructions and guiding genetic manipulations.

Finally, two properties of biological networks became progressively more evident during the development of the computational tools presented in this thesis. First, metabolic networks are amazingly complex and robust motivating a systems-level approach for their analysis. Second, clever engineering strategies are necessary to shift microbial systems from satisfying their own intrinsic objectives into meeting our biotechnological targets.
Bibliography


Nakamura, C. E. (2002). "Production of 1,3-Propanediol by E. coli." presented at Metab Eng IV Conf: Tuscany, Italy.


Anthony Paul Burgard was born on July 27th, 1977 to the loving parents George and Anita Burgard. He grew up in a small East Pittsburgh town, Chalfant Boro, where many life-long lessons such as “giving everything you have” and “never giving up” were learned while playing in the local baseball organization where his father was and still is a coach. Anthony graduated from Woodland Hills High School in 1995 where he was asked out by his future wife, Alysa Mori, for the first time before his AP biology class senior year. Half asleep, he obliged and found himself happily married seven years later. After high school, Anthony attended The Pennsylvania State University where he majored in Chemical Engineering and minored in Environmental Engineering. He decided to continue the research he pursued as an undergraduate under the exceptional guidance of Dr. Costas Maranas at Penn State and he obtained his doctoral degree in Chemical Engineering in the fall of 2004. This research, in the area of metabolic network analysis and redesign, has produced the following peer-reviewed publications:


