UNIVERSITY OF CALIFORNIA, SAN DIEGO

Model-Driven Metabolic Engineering of *Escherichia coli:* A Systems Biology Approach

A Dissertation submitted in partial satisfaction of the Requirements for the degree
Doctor of Philosophy

in

Bioengineering

by

Adam Michael Feist

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Steven P. Briggs
Professor Jeff Hasty
Professor Milton H. Saier
Professor Shankar Subramaniam

2008

The Dissertation of Adam Michael Feist is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2008

To Ashley, My Parents, and Grandparents

Opportunity is rare, and a wise man will never let it go by him.

Bayard Taylor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The best scientists are those who ask the right questions. This was told to me by my advisor, Bernhard Palsson, and I wrote it in the back of my lab notebook for motivation along with the statement “keep in mind the BIG picture”, with a few lines under the word “BIG” to emphasize its importance. During my years in graduate school, it was no longer a question of being able to accomplish a task. As accomplishing tasks such as passing a test with good marks, or finishing a homework assignment on time was a proven recipe for success in an undergraduate education. Research at the graduate level was and still is about determining which research questions are worth answering, what is to gain from the effort in answering them, and then working hard to stay focused on doing so clearly and conclusively. Of course, the path from question to conclusion was not always a clear and smooth process, but I have tried to keep these two thoughts in mind for the work presented herein.

As the work presented here is also a reflection of my experience in graduate school, I want to take a few sentences to point out key concepts and experiences that were significant for me in my graduate career that do not fall into the outline of this technically rich dissertation. The first point I want to make is that I have benefitted greatly personally and professionally by taking advantage of opportunities outside of my research graduate curriculum. Graduate school, especially here at the University of California, San Diego in the Bioengineering Department, is full of opportunities beyond academic and research goals and I would encourage anyone in a similar situation to actively seek them out. I found my benefit in becoming active in the Bioengineering Graduate Students Group working with fellow students, members of the Department, and also with members of the Jacobs School of Engineering. I feel

this opportunity provided invaluable insight into career choices after graduate school, allowed me to build fruitful relationships, and educated me in how to work successfully in a team environment. The second point I want to make is to again keep in mind the ultimate goal of your work. Often times, I would be buried into the details of research and in these instances, one is sometimes moving away from project success and essentially away from gaining new knowledge and experience. Getting into the details is necessary in many instances, but perspective is extremely important to keep in mind. With this, I recommend to the reader to refer to the Abstract to place each chapter in context of the entire dissertation and to keep in mind the two thoughts outlined in the previous paragraph as they review the content of this dissertation.

ACKNOWLEDGEMENTS

We would like to thank Andrew Joyce, Jennifer Reed, Daniel Segre, Nathan Price, Markus Herrgard and Christian Barrett for their invaluable insight.

Chapter 1, in full, is adapted from a review that originally appeared in Nature Biotechnology, volume 26, number 6, pages 659-667, published June, 2008. The dissertation author was the primary author of this paper, which was co-authored by Dr. Bernhard Ø. Palsson.

Chapter 2, in full, is adapted from *Genome-scale reconstruction, modeling, and simulation of E. coli's metabolic network* in Systems Biology and Biotechnology of *E. coli*. Sang Yup Lee, Ed., Springer, that is scheduled to appear. The dissertation author was the primary author of this paper, which was co-authored by Dr. Bernhard Ø. Palsson and Ines Thiele.

We would like to thank A. Osterman and N. Jamshidi for their insights.

Chapter 3, in full, is adapted from Reconstruction of biochemical networks in microbial organisms that is scheduled to appear in Nature Reviews Microbiology. The dissertation author was the primary author of this paper, which was co-authored by Dr. Markus J. Herrgård, Ines Thiele, Dr. Jennie L. Reed, and Dr. Bernhard Ø. Palsson.

We would like to thank Kenyon Applebee, Edward Chuong, Ingrid Keseler, Sean Nihalani, Alan Ruttenberg, Milton Saier, Jan Schellenberger and Jeremy Zucker for their help in the generation and analysis of the reconstruction.

Chapter 4, in full, is adapted from an article that appeared in Nature Molecular Systems Biology, volume 3, number 121, pages 1-18, published June, 2007. The dissertation author was the primary author of this paper, which was co-authored by Dr. Christopher S Henry, Dr. Jennifer L Reed, Markus Krummenacker, Dr. Andrew R Joyce, Dr. Peter D Karp, Dr. Linda J Broadbelt, Dr. Vassily Hatzimanikatis, and Dr. Bernhard Ø. Palsson.

We would like to thank Jennifer Reed, Thuy Vo, Natalie Duarte, Sharon Wiback, Iman Famili, Radhakrishnan Mahadevan and Chris Workman for their invaluable insight in preparation of this chapter.

Chapter 5, in full, is adapted from an article that appeared in Nature Molecular Systems Biology, volume 2, number 2006.0004, pages 1-14, published January, 2006. The dissertation author was the primary author of this paper, which was co-authored by Dr. Johannes C. M. Scholten, Dr. Bernhard Ø. Palsson, Dr. Fred J. Brockman and Dr. Trey Ideker.

Chapter 6, in full, is adapted from Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli* that is in preparation. The dissertation author was the primary author of this paper, which was co-authored by Daniel C. Zielinski, Jeff D. Orth, Jan Schellenberger, Dr. Markus J. Herrgård, and Dr. Bernhard Ø. Palsson.

We would like to thank Alex Azuma for help with HPLC analysis and medium preparation, Karsten Zengler for his help with anoxic cultures, and additionally Vasiliy Portnoy, Kenyon Applebee, and Dae-hee Lee for their invaluable insight in various project aspects.

Chapter 7, in full, is adapted from construction and evolution of *E. coli* production strains designed through model-driven metabolic engineering that is in preparation. The dissertation author was the primary author of this paper, which was co-authored by Jeff D. Orth, Daniel C. Zielinski, and Dr. Bernhard Ø. Palsson.

During the work towards my graduate degree, I have gained knowledge, insight, and motivation from nearly everyone that I have had a chance to interact with over the course of my graduate career. Although there are too many to name, I would like to specifically thank my advisor, Dr. Bernhard Palsson for his guidance, teaching, and support. I truly have no reservations about knowing that I found an ideal environment and advisor in the Systems Biology Research Group at University of California, San Diego and for this I am grateful. Next, I would like to thank the members on my thesis committee for their insight and guidance on the work presented in this dissertation. Each has provided me with both general and specific details that aided in the generation of the work presented. I would like to thank all of the members of the Systems Biology Research Group, past and present, as I can specifically name at least one area where each of them has helped me along the way. These are too numerous to list here, but if you ask me, I will gladly elaborate. I would also like to thank all of the undergraduate researchers who have helped along the way on the work contained in this dissertation, each of whom is acknowledged above. Lastly, I would like to thank my friends and family for their unyielding and undoubting support along the way, who I share this thesis with.

VITA

2003	Bachelor of Science, Chemical Engineering, Honors Program Graduate, University of Nebraska, Lincoln
2005	Master of Science, Bioengineering, University of California, San Diego
2008	Doctor of Philosophy, Bioengineering, University of California, San Diego

PUBLICATIONS

Feist AM, Scholten JCM, Palsson BØ, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2:2006.0004 (2006)

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

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2 (2007)

Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 26:6 (2008)

Lee J, Yun H, Feist AM, Palsson BØ, Lee SY. Genome-scale reconstruction and in silico analysis of the *Clostridium acetobutylicum* ATCC 824 metabolic network. *Appl Microbiol Biotechnol* 80:5 (2008)

Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microbial organisms. *Nat Rev Microbiol*, *In Press* (2008)

Feist AM, Thiele, I, Palsson, BØ, Genome-scale reconstruction, modeling, and simulation of E. coli's metabolic network in *Systems biology and biotechnology of E. coli* Eds: Lee, S.Y., Springer, *In Press* (2008)

Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgard MJ, Palsson BØ. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *In preparation* (2008)

Feist AM, Orth JD, Zielinski DC, Palsson BØ. Construction and evolution of *E. coli* production strains designed through a model-driven analysis. *In preparation* (2008)

ABSTRACT OF THE DISSERTATION

Model-Driven Metabolic Engineering of *Escherichia coli:* A Systems Biology Approach

by

Adam Michael Feist

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2008

Professor Bernhard Ø. Palsson, Chair

Metabolic engineering of microorganisms will be necessary to advance mankind over the coming centuries. Systems biology has the potential to significantly aide in this effort through design, interpretation, and expansion of experimental implementation. This dissertation outlines work towards advancing the field of

systems biology, in general, and specifically focuses on applying this technology to metabolically engineer the bacterium *Escherichia coli*.

The first part of this thesis dissertation focuses on the impact of systems biology in science and engineering through an introduction of the topic and demonstration of systems biology case studies centered on the reconstruction of *E. coli* metabolism. The history of reconstruction of *E. coli* metabolism prior to and since the genomic era is presented and provides the scope of the fundamental biological platform, the metabolic reconstruction, for which later computations are based. The process and product of network reconstruction and the developed methods necessary for validation and use are outlined.

The second part of the thesis dissertation describes the generation, properties, and biological characterization of two organism-specific genome-scale metabolic reconstructions. These reconstructions are for an environmentally important archaea, *Methanosarcina barkeri*, and the aforementioned bacteria and model organism, *E. coli*. The transformation of these reconstructions to computational models is presented along with validation of modeling results through comparison with experimental data. Demonstrations of the utility of metabolic reconstructions as platforms for systems analyses to answer biological questions are presented in application specific examples.

The third part of this thesis dissertation describes how the generated metabolic reconstruction of *E. coli* was used for model-driven metabolic engineering. A computation evaluation of the production potential for native products of *E. coli* from different feedstocks is presented. This study characterizes the range and number of products that can be coupled to growth in *E. coli*. Lastly, the *in vivo* construction,

evolution, and characterization of strains computationally designed from this analysis are presented for validation of approach. The generated strains possess production capabilities suitable for further development at a larger scale.

Taken in whole, this thesis dissertation describes the process developed, outcomes, and future potential of performing systems metabolic engineering of microorganisms.

Chapter 1

Introduction: How Systems Biology is Impacting Science and Engineering – Case Studies on Analyses of *E. coli* Metabolism

1.1 Abstract

The number and scope of methods developed to interrogate and use metabolic network reconstructions has significantly expanded since the first review of the use of constraint-based analysis in *Nature Biotechnology* some 14 years ago. In particular, the *Escherichia coli* metabolic network reconstruction has reached the genome-scale and has been broadly adapted. Specifically, it has been used to address a broad spectrum of basic and practical applications, falling into five main categories: 1) metabolic engineering, 2) model-directed discovery, 3) interpretations of phenotypic screens, 4) analysis of network properties, and 5) studies of evolutionary processes. With these accomplishments in hand, the field is expected to move forward and seek to further, i) broaden the scope and content of network reconstructions, ii) develop new and novel *in silico* analysis tools, and iii) expand in adaptation to uses of proximal and distal causation in biology. Taken together, these efforts will solidify a mechanistic genotype-phenotype relationship for microbial metabolism.

1.2 Introduction

The availability of reconstructed metabolic networks for microorganisms has increased rapidly in recent years, and a growing number of research groups are reconstructing metabolic networks for organisms of interest[1]. A network reconstruction represents a highly curated set of primary biological information for a particular organism and thus can be considered a biochemically, genetically and genomically structured (BiGG) data base[1,2]. A curated BiGG data base (*de facto* a knowledge base) can be converted into a mathematical format (i.e., an *in silico* model), and used to computationally assess phenotypic properties using a variety of computational methods[2,3]. Genome-scale reconstructions are thus, a key step in quantifying the genotype-phenotype relationship and can be used to 'bring genomes to life'[4]. The purpose of this review is to summarize and classify applications utilizing the *E. coli* reconstruction to answer a broad spectrum of biological questions. These studies provide both an up to date review of the applications of constraint-based analysis and a guide to similar applications for the growing number of organisms for which genome-scale reconstructions are becoming available.

1.3 The key steps in the formulation of genome-scale metabolic network models

The four key steps in the formulation and use of genome-scale models are illustrated in **Figure 1.1**. Foundational to the process is the generation of global, or genome-scale, omics data. Omics data, along with legacy information (i.e., the 'bibliome') and small-scale detailed experiments, can be used to define the interactions amongst the biological components that are used to reconstruct organism-specific networks[1]. Network reconstruction is also an iterative, on-going process that continually integrates data in a formal fashion as it becomes available[5].

As a result, a current and well curated genome-scale network reconstruction is a common denominator for those studying systems biology of an organism. An in depth review on the bottom-up reconstruction process can be found in[2] and will not be described here.

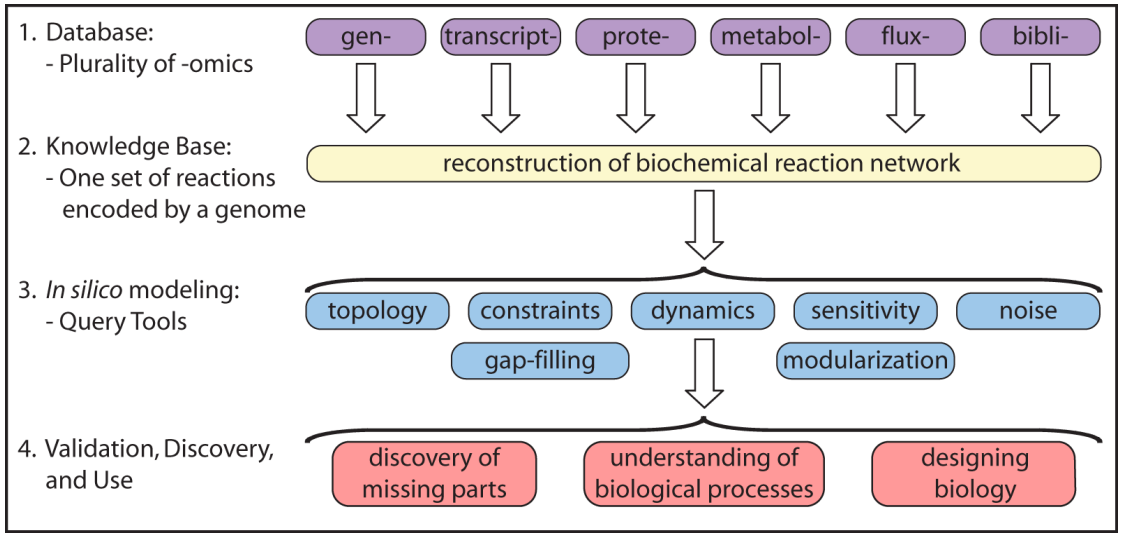


Figure 1.1: Formulation and use of GEMs as a four-step process. Step 1, the process is based on a variety of high-throughput data sets (i.e., omics data) and a comprehensive assessment of the literature (i.e., bibliomic data). Step 2, all of the data types are used to reconstruct the list of biochemical transformations that make up a network as well as their genetic basis[1]. In principal, the network is unique. Step 3, the data contained in the reconstruction can be formally represented (i.e., in the form of matrices and logical statements) that can be mathematically characterized by a variety of methods. Step 4, the computational model enables a broad spectrum of applications, as reviewed in this article. Figure adapted from[2].

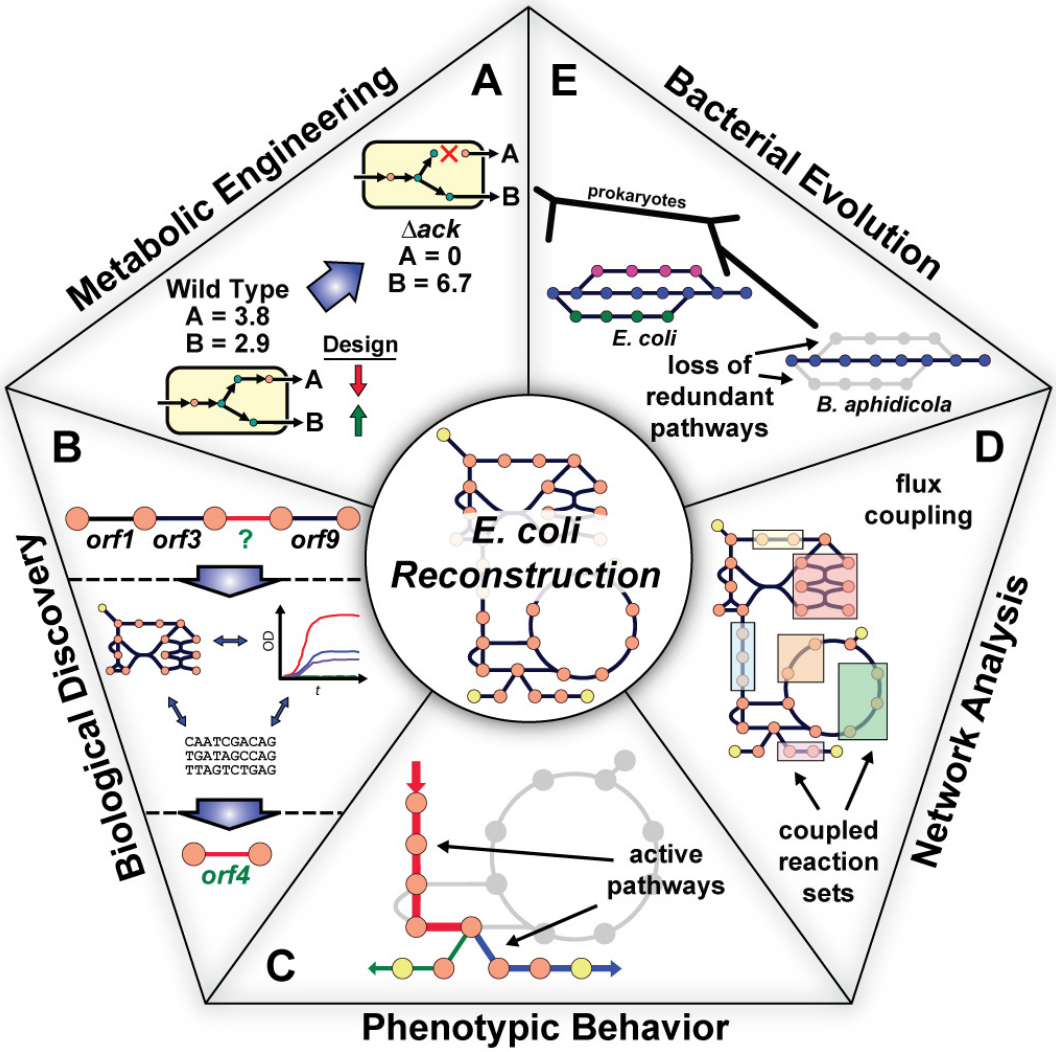
The arrow from step 2 to step 3 in **Figure 1.1** involves a somewhat subtle, but critical, transition. With the definition of systems boundaries and other details, a network reconstruction can be converted into a mathematical format that can be computationally interrogated and subsequently used for experimental design[2]. Thus, a network reconstruction is converted into a Genome-scale Model (GEM)[3]. This arrow represents a bridge between the realms of high-throughput data/bioinformatics on

one hand and systems science on the other. A network reconstruction (or BiGG knowledge base) is accessible to all and significant strides have been made to make computation with GEMs more readily accessible and free of use[6-11]. This availability of both genome-scale reconstruction and GEMs has unleashed creativity in research groups around the world and resulted in the series of studies reviewed below.

1.4 The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli

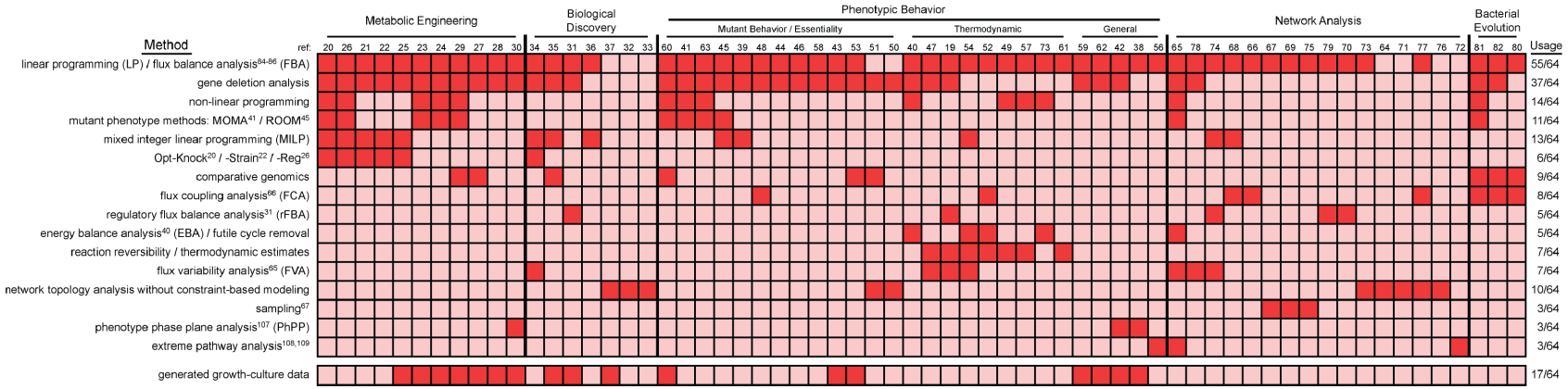
A growing number of research groups utilize the *E. coli* GEM for predicting, interpreting and understanding *E. coli* phenotypic states and function, in addition, the reconstruction itself has been used as a context for the interpretation of large amounts of experimental data. Applications of the *E. coli* GEM range from pragmatic to theoretical studies, and can be classified into five general categories (**Figure 1.2**): 1) metabolic engineering[12-22]; 2) biological discovery[23-29]; 3) assessment of phenotypic behavior[30-56]; 4) biological network analysis[57-72]; and 5) studies of bacterial evolution[73-75]. The *in silico* methods used to probe the *E. coli* GEM in each study are summarized in **Figure 1.3**. It should be noted that these methods perform an assessment of the solution spaces associated with the mathematical representation of a reconstruction[2]; these methods are categorized as unbiased and biased methods[3]. The latter category relies on an observer bias that is stated through an objective function (that is now beginning to be experimentally examined[76]) and is utilized in most of the studies reviewed here use the general application of flux balance analysis (FBA)[77-79]. Each category of application is now detailed, with emphasis on the first three that have the greatest practical utility.

Figure 1.2: Uses of the *E. coli* reconstructions divided into five categories. (facing page) Top: (**A**) A drawing of a predicted effect from a loss of function mutation in a simple system is shown. Metabolic engineering studies have investigated *in silico* strain design using *E. coli* metabolic reconstructions to overproduce desired products[12-22]. (**B**) Recent studies utilizing the reconstruction in a prospective manner have aimed to use the current biochemical and genetic information included in the metabolic network along with additional data types to drive biological discovery, such as predicting genes encoding for orphan reactions[24,25,27-29]. (**C**) Utilizing the reconstruction in phenotypic studies, computational analyses have examined gene[38,43,45,50,56], metabolite[36,53] and reaction[31,39,40,51] essentiality along with considering thermodynamics[32,39,41,44,46,47,49,50,54] to make better predictions about the physiological state (i.e., the active pathways) of the cell for a given environmental condition. (**D**) The *E. coli* reconstructions have been used to analyze and interpret the intrinsic properties of biological networks. One example being finding coupled reaction activities[59] (as shown in the drawing) across different growth conditions. (**E**) Using the network reconstruction, evolutionary studies have examined the cellular network in the context of adaptive evolution events[74], horizontal gene transfer[73,74] and minimal metabolic network evolution (as shown in the drawing)[75]. Bottom: Spectrum of uses of the of the genome-scale *E. coli* metabolic network reconstruction.



Type of Analysis:	Metabolic Engineering	Biological Discovery	Phenotypic Behavior	Network Analysis	Bacterial Evolution
Application:	Practical				Basic

Figure 1.3: Summary of the *in silico* methods utilized in published *E. coli* GEM studies. (facing page) This heatmap characterizes the incorporation of different computational methods into studies utilizing genome-scale models of *E. coli*. A dark box indicates that a particular method (one method per row) was utilized in a corresponding study (one citation per column, for citation numbering[80]); the frequency of usage of a particular method is given on the right. Studies were grouped into one of five general categories and studies examining phenotypic behavior were further divided into three subgroups. Studies that contributed new experimental growth data are also marked along the bottom offset row.


Metabolic Engineering
Biological Discovery
Phenotypic Behavior
Mutant Behavior / Essentiality
Thermodynamic
General
Network Analysis
Bacterial Evolution
Method
ref:
20 26 21 22 25 23 24 29 27 28 30
34 35 31 36 37 32 33
60 41 63 45 39 48 44 46 58 43 53 51 50
40 47 19 54 52 49 57 73 61
59 62 42 38 56
65 78 74 68 66 67 69 75 79 70 73 64 71 77 76 72
81 82 80
Usage
linear programming (LP) / flux balance analysis[84-86] (FBA) 55/64
gene deletion analysis 37/64
non-linear programming 14/64
mutant phenotype methods: MOMA[41] / ROOM[45] 11/64
mixed integer linear programming (MILP) 13/64
Opt-Knock[20] / -Strain[22] / -Reg[26] 6/64
comparative genomics 9/64
flux coupling analysis[66] (FCA) 8/64
regulatory flux balance analysis[31] (rFBA) 5/64
energy balance analysis[40] (EBA) / futile cycle removal 5/64
reaction reversibility / thermodynamic estimates 7/64
flux variability analysis[65] (FVA) 7/64
network topology analysis without constraint-based modeling 10/64
sampling[67] 3/64
phenotype phase plane analysis[107] (PhPP) 3/64
extreme pathway analysis[108,109] 3/64
generated growth-culture data 17/64

1.5 Applications of GEMs to metabolic engineering of *E. coli*

Through the application of computational methods that incorporate linear, mixed integer linear, and non-linear programming, it has been demonstrated that model-directed strain design can lead to increased metabolite production[12-22]. In these studies, the *E. coli* GEM is principally used to analyze the metabolite production potential of *E. coli* and identify metabolic interventions needed to enable the production of the product of interest. Thus, *E. coli* strains have been systematically designed through *in silico* analysis to overproduce target metabolites such as lycopene[15,16], lactic acid[17], ethanol[18], succinic acid[19,20], L-valine[21], L-threonine[22], additional amino acids[13], as well as diverse products from hydrogen to vanillin[14]. Select exemplary metabolic engineering applications will be described in more detail.

To increase the production of an already high producing strain, a systematic computational search was developed[16] to explore the *E. coli* metabolic network and report gene deletions that diverted metabolic flux towards the desired product. This process resulted a knock-out strain, that when constructed, showed a two-fold increase in the production of lycopene over the parental strain. In this analysis, the computational algorithm MOMA[33] and the *i*JE660[81] *E. coli* GEM were utilized to sequentially examine additive genetic deletions that would improve lycopene production while maintaining cell viability. Strain designs were constructed through genetic manipulations using the predicted modifications and it was found that this computational approach yielded the twofold increase in production rate over a previously engineered overproducing strain and an 8.5 fold increase over wild-type production harboring only a lycopene biosynthesis plasmid[16]. Strain performance was evaluated by monitoring lycopene production through enzymatic assays and mutant

growth rates. In addition, the strain designs identified computationally were compared to mixed combinatorial transposon mutagenesis and it was found that the maximum production observed could be designed solely using the systematic GEM aided computational method[15,16]. Furthermore, a deleterious effect was observed when targets identified in individual computational designs were combined in an attempt to achieve an overall more desirable phenotype. Thus, the overall systematic effects from individual designs were not additive and needed to be interpreted in the context of the entire network.

Two studies producing the amino acids L-valine[21] and L-threonine[22] have demonstrated the broad usage of GEM aided computation for strain design. In the first study, GEM aided modeling was employed in three different areas to increase the production of L-threonine to industrial titers[22]. In one instance, *in silico* modeling was used to identify the optimal activity of a key enzymatic reaction towards maximum L-threonine production using a parametric sensitivity analysis that compared reaction activity to L-threonine production rate. The optimal activity prediction was subsequently used to tune the overexpression of the gene which encodes for this enzymatic reaction through comparison to base-line activity and the result was a production increase. This method proved to be vital to the success of this strain, as a previous transcription profiling guided attempt at overexpression resulted in an undesirable surplus of activity and was detrimental to L-threonine production. For the same strain, a GEM aided flux analysis in conjunction with mRNA expression data levels also guided the elimination of negative regulation on a gene which encoded for a reaction that channeled flux towards the final product. The third use of the GEM for the design of this strain occurred when an unwanted byproduct was observed in the culture medium and computation was utilized to divert the flux from

this byproduct to L-threonine[22] through overexpression of another key gene encoded activity. The second analysis applied the systematic computational search algorithm previously described[16] to the updated *E. coli* GEM MBEL979[7] (similar to the *i*JR904 GEM[82]) to improve L-valine production. The *in silico* analysis of beneficial knock-outs to divert flux towards the desired product once again resulted in a significant increase in the production of the desired metabolite over an existing overproducing strain; more than a two-fold increase in this case[21]. Furthermore, in this same study, a number of additional metabolic engineering approaches to increase overproduction were performed (i.e., relieving feedback inhibition and regulation through attenuation, removing competing pathways, up-regulation of primary biosynthetic pathways, and overexpression of exporting machinery). When compared to each of the other individual strain modifications, the *in silico* GEM aided interventions resulted in the greatest increase in L-valine production[21]. Taken together, these two studies demonstrate the broad applications for which GEMs can be utilized to design strains not only in a *de novo* fashion, but to make further improvements on strains through integrating and interpreting experimental data.

Several other strain designs utilizing *E. coli* GEMs have been reported. In a combined computational and experimental study, the bi-level optimization algorithm OptKnock[12] and *i*JR904[82] were utilized to overproduce lactate in *E. coli*[17]. The algorithm OptKnock optimizes two objective functions, biomass formation and product secretion, to produce strains that will couple the excretion of a desirable product to the growth rate. Using adaptive evolution with growth rate selection pressure, the lactate producing strains designed using OptKnock were found to possess this growth-coupling property. Growth rate, uptake and secretion rate profiles were the measures by which this property was examined and thus this study demonstrated the

utility of adaptive evolution as a design tool[83]. Additional noteworthy examples of GEM aided design are two studies which demonstrated[19,20] that GEM modeling using *i*JR904[82] was beneficial to screen genes that were deemed to be important for succinate production. Combinatorial knock-outs that were predicted to be overproducers *in silico* were experimentally verified to display the same overproducing phenotype *in vivo*. Furthermore, this method had an advantage over using comparative genomics for strain design, which was also performed in one of the studies[19].

Taken together, a growing number of metabolic engineering studies demonstrate the use of GEMs to generate strain designs that are often non-intuitive and non-obvious. An excellent example of a non-intuitive strain improvement outlined in this section was when modeling was used to not only study the effect of a gene removal, but to tune the expression of a gene to an optimally predicted level, that when expressed too highly, was detrimental to product formation. Genome-scale reconstructions thus allow the examination and simulation of metabolism as an integrated network, circumventing the possible shortcomings of methods that rely on manual assessment of a limited number of interactions and fail to detect non-intuitive causal interactions. With the growing availability of organism and strain specific GEMs, applications for designing microbial strains for industrial production are expected to continue to grow. This growth expectation is in part based on the on-going reconstruction of additional cellular processes, such as transcriptional regulation and protein production. Computations based on genome-scale models are also beginning to influence other areas of industrial microbiology such as generation of renewable energy[84-86] and bioremediation[85].

1.6 Directing discovery: GEM-driven discovery in *E. coli*

GEMs can provide a guide to biological discovery. This capability is based on comparison of computed and actual experimental outcomes. Given the fact that BiGG knowledge bases are incomplete and that they contain gaps[87], they provide a context for systematic discovery of missing information. The comparison between computation and experiments are summarized in **Figure 1.4** highlighting how agreements and disagreements are analyzed.

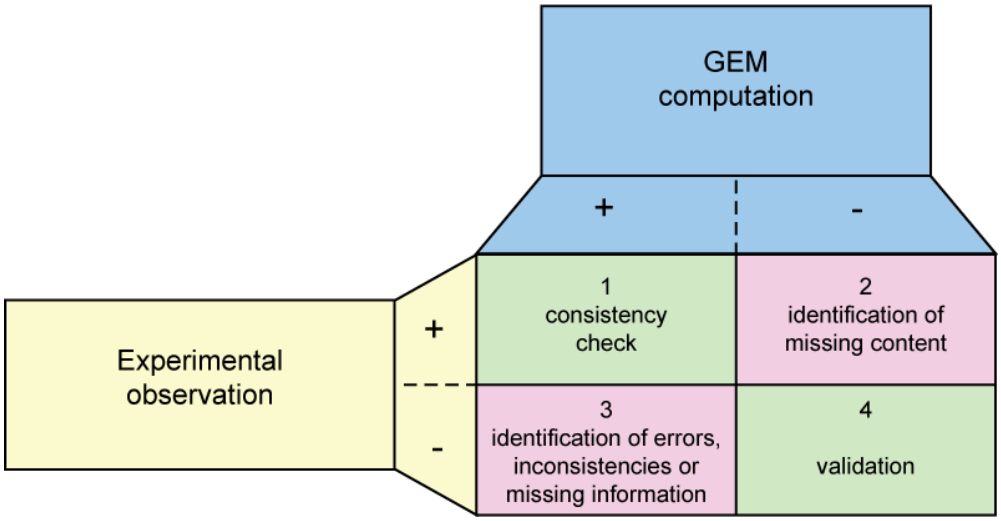


Figure 1.4: Comparison of computation and experimental data: identification of agreements and disagreements. The comparison of GEM computation and organism-specific experimental measurements identifies agreements and disagreements. The phenotypic outcomes are tabulated for genetic perturbations examined in a given environment (e.g., growth or no growth). A '+' indicates that a given phenotype is not affected by the perturbation, and '-' indicates it does. Each outcome of comparison has a different implication; 1: consistency check - a perturbation has no affect on the property being measured and modeling predicts the same; 4: validation - the perturbation affects the experimental outcome and modeling with the GEM predicts this outcome; 2: identification of missing content - when GEM modeling fails to predict the positive confirmation of the property being measured, this outcome indicates that there is missing content in the GEM and can lead to the identification of specific areas for biological discovery; 3: identification of errors, inconsistencies or missing context-specific information – a positive prediction for the measured property and an opposite experimental observation indicates a possible error in the current organism-specific knowledge or that additional context-specific information is lacking from the GEM or modeling method (e.g., transcriptional regulation).

The current area of most significant interest is to direct discovery efforts towards characterizing unknown ORFs in the *E. coli* genome. Ten years after the first release of the complete genome-sequence[88], many unknown ORFs still exist in the *E. coli* genome (see Supplementary Data[80]), with many of these likely to encode metabolic functions. ORF discovery utilizing GEMs also has significant potential to impact not only how new and less studied genomes are annotated, but to fill out the missing pieces in *E. coli* metabolism.

To address this challenge, algorithms have been developed to determine the probable gene candidates that fill knowledge gaps in the *E. coli* and other network reconstructions. These algorithms utilize global network topology and genomic correlations, such as genome context and protein fusion events[24], as well as local network topology and/or phylogenetic profiles[24,25]. Similar tools has been developed which utilize mRNA coexpression[89] and which can evaluate more general metabolic pathway databases[90]. In addition to these network topology-based methods, an optimization based procedure has also been developed to fill network gaps and evaluate reaction reversibility along with adding additional transport and intracellular reactions from databases of known metabolic reactions[28]. These studies produce specific targets for drill-down experiments needed for confirmation of these computationally generated hypotheses.

Two recent studies have integrated a combined computational and experimental approach to aid the ORF discovery process in *E. coli* through utilizing the GEM and high-throughput phenotype data[27,29]. The first study utilized an iterative process[27] in which, 1) differences in modeling predictions and high-throughput growth phenotype data were identified, 2) potential missing reactions that remedy these

disagreements were algorithmically determined, 3) bioinformatics was utilized to identify likely encoding ORFs, and 4) resulting targeted ORFs were cloned and experimentally characterized. Application of this process led to the functional characterization of eight ORFs that are involved in transport, regulatory and metabolic functions in *E. coli*[27]. The discovery process was aided by a high-throughput growth phenotyping analysis and the genome-wide single-gene mutant collection[91], along with other characterization analyses such as targeted expression profiling. The second GEM-based analysis which resulted in ORF discovery utilized network topology to examine orphan reactions in the *E. coli* network (i.e., reactions known to exist in *E. coli* that have not been linked to an encoding gene) identified by the previously mentioned network topology-based gap-filling algorithms[24,25,89]. The basic premise behind these algorithms is the utilization of an orphan reaction's network neighbors as constraints to assign metabolic function. With the resulting tentative ORF assignment, biochemical characterization studies utilizing genetic mutants[91], analysis of growth under different substrate conditions, and expression data were all utilized to characterize and assign function to an orphan ORF that is responsible for a metabolic conversion that has been known for 25 years[29].

Further studies in this category of biological discovery applications (not focused on ORF identification) have utilized GEMs of *E. coli* to identify potential bottleneck reactions in the metabolic network[26] and as of yet uncharacterized transcription factor target interactions in *E. coli*[23]. The aforementioned study targeting the elucidation of regulatory and metabolic interactions in *E. coli* developed an iterative procedure focused on reconciling computational and experimental discrepancies stemming from high-throughput growth phenotype and gene expression data where selected expression changes were validated using RT-PCR[23].

With the advancement of high-throughput technologies to test the hypotheses generated from computational studies, these and similar algorithmic approaches are likely to continue to aid in the quest to achieve full functional annotation of the *E. coli* genome and its context-specific uses.

1.7 Phenotypic functions: GEM aided assessment

The area where the *E. coli* GEMs has been most extensively utilized is for the examination and quantitative interpretation of metabolic physiology for wild-type, genetically perturbed and adaptively evolved strains of *E. coli*[30-56]. These efforts have implications in both the quantitative and qualitative understanding of physiological states of the cell. Furthermore, these efforts have examined *E. coli* physiology for a vast number of given genetic and environmental conditions and incorporation of the developed methods will have an impact on future design of biological systems and modeling approaches. A large subset of these studies of phenotypic behavior aim to utilize thermodynamic laws and information to refine phenotype predictions of GEMs and to incorporate metabolomic and fluxomic data into modeling[32,39,41,44,46,47,49,50,54].

A set of distinct computational methods using GEMs have been developed to determine the physiological state of *E. coli* after genetic perturbations[33,37,42]. These studies have utilized ^{13}C flux measurements and growth rate phenotype data to evaluate the predictability of the developed algorithms when compared to experimental observations. Whereas comparisons to flux data from wild-type and *E. coli* mutants reveals that the computational algorithm MOMA[33] provides better predictions for transient growth rates (early post perturbation state), the algorithm ROOM[37] (and basic FBA) was found to be more successful in predicting final steady-

state growth rates and overall lethality[37]. These algorithms have been utilized, in addition to basic FBA, for genome-wide essentiality screens, as now outlined.

A range of computational studies have sought to understand phenotypes through determining the essential genes[38,43,45,50,56], metabolites[36,53] and reactions[31,39,40,51] in the *E. coli* metabolic network. A common benchmark for examining GEM predictive ability is to determine the agreement with growth phenotype data from knock-out collections of *E. coli*. Such studies will be further enabled by the recent availability of a comprehensive single-gene knock-out library for *E. coli*[91] (for example[45,50]). Implications for examining network essentiality in *E. coli* include determining network essentiality in similar organisms[31,40,45,51], deciphering network makeup and enzyme dispensability (i.e., measures of robustness)[38,51,53], aiding in metabolic network annotation, validation and refinement[36], and even rescuing knock-out strains through additional gene deletions[56], to name a few. The predictive capability of the *E. coli* GEM, as demonstrated by these studies, has been instrumental in the adaptation of its use. One particular study examining knock-out phenotypes has demonstrated that the *E. coli* GEM was able to predict the outcomes of adaptively evolved strains to a high degree (78%) when knock-out *E. coli* strains were grown in a number of different substrate environments by examining growth rates at the beginning and end of adaptive evolution[35]. This study represents a demonstration of a GEM's ability to look at adaptive behavior (or 'distal' causation[92]), in addition to immediate behavior (or 'proximal' causation[92]). Predictive capability is expected to improve through examining growth behavior across a greater number of environments (additional phenotyping screens will be necessary) and with an increase of integration of additional cellular processes. Genetic perturbations have played a key role in the study of the genotype-phenotype relationship in biology and

GEMs can be used to mechanistically interpret the results and predict the outcomes of such perturbations.

Incorporating thermodynamic information into *E. coli* GEMs has shown promise in narrowing predictions of allowable physiological states in a given environment[32,39,41,44,46,47,49,50,54] and in identifying reactions likely to be subject to active allosteric or genetic regulation[41,46]. This field is progressing rapidly and should prove to increase the predictive capabilities of genome-scale modeling through the addition of governing thermodynamic physiochemical constrains. One particular analysis incorporating compound formation and reaction energies for the content of the GEM based on *i*JR904[82] identified reactions that are likely to be effectively irreversible for any realistic metabolite concentration[46]. The hypothesis was advanced that these reactions are candidates for cellular regulation in their respective pathways since enzyme regulation will likely be the dominant mechanism for control of flux through these reactions[46].

The addition of thermodynamics enables the analysis of metabolomic data in the context of a reconstruction. A study utilizing high-throughput metabolomic data and GEMs proposed likely regulatory interactions by deciphering the metabolite concentrations in the context of overall network functionality[41]. Not only did the metabolomic data benefit computations by constraining the system using physiological measurements, but the computational predictions were also able to validate quantitative metabolomic data sets for consistency through providing a functional context to relate metabolite concentrations. This application is one example of how metabolomic data will directly influence modeling and metabolite concentration data is likely to greatly influence future metabolic modeling due to its

intimate connection with GEM content. Similar work incorporating other quantitative values with FBA, such as metabolite concentrations[49] and flux ratios at branch points in metabolism[48] is also appearing.

Applying a different physiochemical constraint, molecular crowding, a framework has also been developed to incorporate spatial constraints into FBA[52]. The functional states predicted with this method (i.e., FBA with molecular crowding, FBAwMC) and the *E. coli* GEM were validated against generated growth, substrate, and production rate data along with gene expression profiles and enzyme activity measures to demonstrate predictive accuracy, including substrate preferentiality, when examining growth in complex substrate environments[52,55]. Overall, these studies which incorporate reaction thermodynamics and additional cellular constraints should further narrow the range of allowable functional network states that can made based on stoichiometry alone and thus improve the utility of GEMs.

In addition to analyses on the genomic scale, a number of studies modeling the metabolism of *E. coli* on a smaller-scale have been performed. These analyses typically utilize models containing approximately 100 reactions or less and most often, focus on incorporating non-linear analysis to understand quantitative experimental data (e.g., isotopomer modeling). With the advancement of computational power and developed platforms, the networks that can be analyzed will grow in size[93]. Given that the results produced from analyses such as isotopomer modeling have been shown to be highly dependent on the content of a reduced model, the logical starting point for building such models is the *E. coli* GEM[93]. A number of noteworthy studies have been conducted with reduced models, but not detailed here as they are outside the scope of this review.

1.8 Systems biology: analysis of network properties

E. coli is generally viewed as having the most complete characterization of any model organism[94,95]. Due to the incorporation of thousands of metabolic interactions with relatively high reliability (e.g., 92% of the genes included in the latest reconstruction of *E. coli*[50] have experimentally determined annotated functions[95], (see **Table 4.1**, **§4.3.1**), validated genome-scale reconstructions of *E. coli* have become popular resources for the analysis of various network properties[57-72]. The methods designed to analyze the underlying network structure of *E. coli* metabolism, some characterizing its interplay with regulation, have been developed to determine a number of physiological features. These features include the most probable active pathways and utilized metabolites under all possible growth conditions[60,62,66,68], the existence of alternate optimal solutions and their physiological significance[58], conserved intracellular pools of metabolites[61], coupled reaction activities[59] and their relationship to gene co-expression[70], metabolite coupling[64], metabolite utilization[65], the organization of metabolic networks[57,69], strategies for *E. coli* to incorporate metabolic redundancy[71], and the dominant functional states of the network across various environments[63,67,72]. These findings are both driven by biased approaches utilizing FBA and biomass objective function optimization and by unbiased approaches such as graph-based analyses (see **Figure 1.3**). One noteworthy study utilizing the GEM outlined network examined thousands of different potential growth conditions and observed a 'high-flux backbone' in *E. coli* that both carried high levels of flux across the different environmental conditions and was composed of a relatively small set of enzymatic reactions[60]. This result can be of practical importance for synthetic biology efforts aimed towards manipulating flux within biological systems. Furthermore, this finding was hypothesized to be a universal feature of metabolic

activity in all cells and was consistent with flux measurements from ^{13}C labeling experiments[60].

The studies in this category have a common systems biology theme; namely the development and subsequent demonstration of methods that identify sets of reactions or metabolites with correlated or coordinated functions and systematic relationships. The systems biology that these methods enable and demonstrate has potential implications for, i) antimicrobial drug-target discovery[61,62], ii) aiding the development of additional metabolic reconstructions[59,61], iii) guiding genetic manipulations[59], iv) improving metabolic engineering applications[60,61], and v) increasing the general understanding of biological network behavior[58,67,70] and resilience[71]. The role that the *E. coli* GEM has taken is a comprehensive and curated set of up to date metabolic knowledge; thus providing a scaffold for these large-scale computations.

1.9 Bacterial evolution: GEM aided studies of distal causation

The GEMs of *E. coli* have been used to examine the process of bacterial evolution[73-75]. Specifically, the network reconstructions have been used to interpret adaptive evolution events[74], horizontal gene transfer[73,74] and evolution to minimal metabolic networks[75]. These studies, which utilize the *E. coli* reconstruction as an organism-specific genetic and metabolic content database, and the corresponding GEM, have been able to provide insight into evolutionary events through combining known physiological data (e.g., in various environmental conditions) with hypotheses and *in silico* computation. Examining the evolution of minimal metabolic networks through simulation demonstrated that it was possible to predict the gene content of close relatives of *E. coli* by examining the necessity of genes and reactions in the

overall context of the system functionality for a specific lifestyle[75]. Similarly, by re-examining network functionality in a number of different environments and through the utilization of comparative genomics, it was shown that recent evolutionary events (i.e., horizontal gene transfer) likely resulted from a response to a change in environment[74]. Furthermore, computational analysis led to the additional conclusion that these horizontal gene transfer events are more likely if the host organism contains an enzyme that catalyzes a coupled metabolic flux related to the transferred enzyme's function[73,74]. Taken together, these studies demonstrate the importance of having high-quality curated reconstructions to enable studies on an organism's response to environmental changes and for understanding the fundamental forces driving bacterial evolution.

1.10 Closing

The myriad of studies described in this review highlights the rapid development and use of genome-scale reconstruction and derived computational models to address a growing spectrum of basic research and applied problems. The experience with genome-scale reconstructions has demonstrated that they are a common denominator in the systems analysis of metabolic functions. With the recognition of its basic paradigms and a growing spectrum of practical uses enabled, there are several exciting challenges that this field now faces. Accordingly, further development is necessary, and three major areas where it will be influential are now discussed; i) network reconstructions and the reconstruction process, ii) computational BiGG query tools (i.e., modeling), and iii) application to proximal and distal causation in biology.

The scope of reconstructions is bound to grow, representing more and more BiGG knowledge in the structured format of a GEM[87]. Growth in scope in the near-term will on one front, involve the transcriptional and translational machinery of bacterial cells[96-98]. Such an extension will enable a range of studies including the direct inclusion of proteomic data, fine graining of growth requirements and the explicit consideration of secreted protein products. Another expansion in scope in the near-term is the reconstruction of the genome-scale transcriptional regulatory network (TRN). Such reconstruction at the genome-scale is now enabled by new experimental technologies, such as ChIP-chip[99]. Experimental interrogation of the currently available TRN suggests that we know about one-fourth to one-third of its content[23], indicating that there is much to be discovered. Once reconstructed, the TRN will allow computational predictions of the context-specific uses of the *E. coli* genome and the responses of two-component signaling systems. Taken together, these near-term expansions in content will encompass the activity of apparently 2000 ORFs in the *E. coli* genome.

Mid-term expansions in scope will include the growth cycle, shock responses and additional cellular functions. Such a reconstruction should eventually be a comprehensive representation of the chemical reactions and transactions enabled by *E. coli*'s gene products. Longer-term reconstruction may begin to address the 3-dimensional organization of the bacterial cell. In particular, high-resolution ChIP-chip data on the DNA binding protein could enable the estimation of the topological arrangement of the genome, and potentially elucidate the structure of the cell wall and other cellular structures that will allow us a full 3-dimensional reconstruction of *E. coli*.

We now know how to represent BiGG data in either a stoichiometric format or in the form of causal relationships[100] and how to use them to perform several lines of computational inquiries. Computational query tools of GEMs will continue to be developed. New advances will likely include modularization methods, use of fluxomic data and eventually kinetics. As the scope and content of the reconstruction grows, the need to modularize its content becomes more pressing. Fine or course grained views of cellular processes are needed for different applications. For instance, as previously mentioned, current computational limitations force the reduction in a network for the analysis of isotopomer data, and a rational way to carry out such reduction is needed. Given the systemic nature of fluxomic data and its phenotypic relevance, there is a pressing need to increase the size of the networks that can be analyzed for experimental measurement and estimation of flux states. Finally, although detailed kinetic models of microbial functions may currently be mostly of academic interest, we will most likely be able to construct them in the mid-term based on advances with metabolomic and fluxomic data, in addition to the developments that are occurring with the incorporation of thermodynamic information. Such large-scale kinetic models are likely to differ from those resulting from traditional approaches for construction of kinetic models as, they come with different challenges.

As this review shows, the scope of applications of genome-scale reconstructions and GEMs is growing. Going forward, we wish to comment on three categories of applications: growth in coverage (i.e., gap-filling), engineering (i.e., synthetic biology), and the development of fundamental understanding. Growth in coverage will come through discovery of missing network components. For instance, the latest metabolic reconstruction, *i*AF1260, contains 14% blocked reactions[50]. This disconnected content means that we have knowledge gaps that have arisen due to

characterization of individual gene products outside the context of a given physiological function (i.e., outside a defined pathway). Metabolomic profiling is one measure that will provide us with the missing upstream or downstream routes to such dead ends in the network. Also, an expansion of scope in modeling will allow for further investigation of network content, such as tRNA charging reactions that are currently in this blocked reaction set[50]. Furthermore, growing metabolomic data suggests that we are discovering the existence of several new metabolites. Pathways that include these metabolites need to be discovered. Methods exist to compute missing pathways between molecules[101] that can be applied to such data. Such pathways, in turn, will lead to experimental programs to discover novel gene functions and to validate or refute the existence of such pathways. Similarly, we expect that a number of the components of TRNs are missing, such as new sRNA molecules (see Supplementary Data[80]). Clearly, maintaining the quality control/quality assurance of such reconstructions will help in guiding us to a comprehensive genome-scale representation of all major cellular processes in bacteria at the BiGG data level of resolution that, in turn, enables GEMs of growing coverage and resolution.

Predictive models allow for design. In fact, in engineering, there is ‘nothing more useful that a good theory.’ As this review demonstrates, genomics and high-throughput technologies have enabled the construction of predictive computational models. The scope of such predictions is limited at the moment, but with the growing scope and coverage of genome-scale reconstructions and advancements in the development of computational tools, this scope will broaden. Not only will GEMs influence design in synthetic biology, but their influence in discovery of cellular content will provide a more complete picture of the environment (i.e., the parts list in the cell) in which future synthetically engineered constructs and circuits will be placed.

The impact of GEMs on synthetic biology is thus likely to be notable; ranging from the provision of the cellular-context of a small-scale gene circuit design to engineering of the entire genome-scale network towards fundamentally new and useful (i.e., production) phenotypes.

Finally, we can speculate about the deep scientific impact that comprehensive predictive GEMs will have on our understanding of the living process. A comprehensive view of cellular functions will allow us to study the fundamental properties of both the underlying energy and information flows in living organisms. Such a view is likely to deeply affect our understanding of both distal and proximal causation in biology.

Acknowledgements

We would like to thank Andrew Joyce, Jennifer Reed, Daniel Segre, Nathan Price, Markus Herrgard and Christian Barrett for their invaluable insight.

Chapter 1, in full, is adapted from a review that originally appeared in Nature Biotechnology, volume 26, number 6, pages 659-667, published June, 2008. The dissertation author was the primary author of this paper, which was co-authored by Dr. Bernhard Ø. Palsson.

References

1. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

2. Palsson BO. Systems biology: properties of reconstructed networks. New York: Cambridge University Press, 2006.

3. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

4. Frazier ME, Johnson GM, Thomassen DG, Oliver CE, Patrinos A. Realizing the potential of the Genome Revolution: The Genomes to life Program. *Science* 2003;300:290-3.

5. Reed JL, Palsson BO. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol* 2003;185:2692-9.

6. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

7. Lee SY, Woo HM, Lee D-Y, Choi HS, Kim TY, Yun H. Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioproc. Eng.* 2005;10:425-431.

8. Klamt S, Saez-Rodriguez J, Gilles ED. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol* 2007;1:2.

9. Raman K, Chandra N. Pathway Analyser: for FBA/MoMA analyses of metabolic pathways 11th SBML Forum, 2006.

10. Luo RY, Liao S, Zeng SQ, Li YX, Luo QM. FluxExplorer: A general platform for modeling and analyses of metabolic networks based on stoichiometry. *Chinese Science Bulletin* 2006;51:689-696.

11. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin, II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;19:524-31.

12. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003;84:647-57.

13. Pharkya P, Burgard AP, Maranas CD. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol Bioeng* 2003;84:887-99.

14. Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* 2004;14:2367-76.

15. Alper H, Jin YS, Moxley JF, Stephanopoulos G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* 2005;7:155-64.

16. Alper H, Miyaoku K, Stephanopoulos G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* 2005;23:612-6.

17. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91:643-8.

18. Pharkya P, Maranas CD. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab Eng* 2006;8:1-13.

19. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY. Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. *Appl Environ Microbiol* 2005;71:7880-7.

20. Wang Q, Chen X, Yang Y, Zhao X. Genome-scale in silico aided metabolic analysis and flux comparisons of *Escherichia coli* to improve succinate production. *Appl Microbiol Biotechnol* 2006;V73:887-894.

21. Park JH, Lee KH, Kim TY, Lee SY. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci U S A* 2007;104:7797-802.

22. Lee KH, Park JH, Kim TY, Kim HU, Lee SY. Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* 2007;3:149.

23. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429:92-6.

24. Chen L, Vitkup D. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol* 2006;7:R17.

25. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* 2006;7.

26. Herrgard MJ, Fong SS, Palsson BO. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* 2006;2:e72.

27. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 2006;103:17480-4.

28. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 2007;8:212.

29. Fuhrer T, Chen L, Sauer U, Vitkup D. Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in Escherichia coli. *J Bacteriol* 2007;189:8073-8078.

30. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001;19:125-130.

31. Burgard AP, Vaidyaraman S, Maranas CD. Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol Prog* 2001;17:791-7.

32. Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. *Biophys J* 2002;83:79-86.

33. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 2002;99:15112-7.

34. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186-9.

35. Fong SS, Palsson BO. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat Genet* 2004;36:1056-58.

36. Imielinski M, Belta C, Halasz A, Rubin H. Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities. *Bioinformatics* 2005;21:2008-16.

37. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* 2005;102:7695-700.

38. Ghim CM, Goh KI, Kahng B. Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. *J Theor Biol* 2005;237:401-11.

39. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys J* 2006;90:1453-61.

40. Samal A, Singh S, Giri V, Krishna S, Raghuram N, Jain S. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* 2006;7:118.

41. Kümmel A, Panke S, Heinemann M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* 2006;2:2006.0034.

42. Wunderlich Z, Mirny LA. Using the topology of metabolic networks to predict viability of mutant strains. *Biophys J* 2006;91:2304-11.

43. Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. Essential genes on metabolic maps. *Curr Opin Biotechnol* 2006;17:448-56.

44. Kümmel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 2006;7.

45. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. *J Bacteriol* 2006;188:8259-8271.

46. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* 2007;92:1792-1805.

47. Ederer M, Gilles ED. Thermodynamically feasible kinetic models of reaction networks. *Biophysical Journal* 2007;92:1846-1857.

48. Choi HS, Kim TY, Lee DY, Lee SY. Incorporating metabolic flux ratios into constraint-based flux analysis by using artificial metabolites and converging ratio determinants. *J Biotechnol* 2007;129:696-705.

49. Hoppe A, Hoffmann S, Holzhutter HG. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. *BMC Syst Biol* 2007;1.

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

51. Guimera R, Sales-Pardo M, Amaral LA. A network-based method for target selection in metabolic networks. *Bioinformatics* 2007;23:1616-22.

52. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabasi AL, Oltvai ZN. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proc Natl Acad Sci U S A* 2007;104:12663-8.

53. Kim PJ, Lee DY, Kim TY, Lee KH, Jeong H, Lee SY, Park S. Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proc Natl Acad Sci U S A* 2007;104:13638-42.

54. Warren PB, Jones JL. Duality, thermodynamics, and the linear programming problem in constraint-based models of metabolism. *Phys Rev Lett* 2007;99:108101.

55. Vazquez A, Beg QK, Demenezes MA, Ernst J, Bar-Joseph Z, Barabasi AL, Boros LG, Oltvai ZN. Impact of the solvent capacity constraint on *E. coli* metabolism. *BMC Syst Biol* 2008;2:7.

56. Motter AE, Gulbahce N, Almaas E, Barabasi AL. Predicting synthetic rescues in metabolic networks. *Mol Syst Biol* 2008;4:168.

57. Gagneur J, Jackson DB, Casari G. Hierarchical analysis of dependency in metabolic networks. *Bioinformatics* 2003;19:1027-34.

58. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 2003;5:264-76.

59. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions. *Genome. Res.* 2004;14:301-12.

60. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 2004;427:839-843.

61. Nikolaev EV, Burgard AP, Maranas CD. Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys J* 2005;88:37-49.

62. Almaas E, Oltvai ZN, Barabasi AL. The Activity Reaction Core and Plasticity of Metabolic Networks. *PLoS Comput Biol* 2005;1:e68.

63. Barrett CL, Herring CD, Reed JL, Palsson BO. The global transcriptional regulatory network for metabolism in *Escherichia coli* attains few dominant functional states. *Proc Natl Acad Sci U S A* 2005;102:19103-19108.

64. Becker SA, Price ND, Palsson BO. Metabolite coupling in genome-scale metabolic networks. *BMC Bioinformatics* 2006;7.

65. Imielinski M, Belta C, Rubin H, Halasz A. Systematic Analysis of Conservation Relations in *Escherichia coli* Genome-Scale Metabolic Network Reveals Novel Growth Media. *Biophys J* 2006;90:2659-72.

66. Beasley JE, Planes FJ. Recovering metabolic pathways via optimization. *Bioinformatics* 2007;23:92-8.

67. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* 2007;3:101.

68. Almaas E. Optimal flux patterns in cellular metabolic networks. *Chaos* 2007;17:026107.

69. Sales-Pardo M, Guimera R, Moreira AA, Amaral LA. Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci U S A* 2007;104:15224-9.

70. Notebaart RA, Teusink B, Siezen RJ, Papp B. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol* 2008;4:e26.

71. Mahadevan R, Lovley DR. The degree of redundancy in metabolic genes is linked to mode of metabolism. *Biophys J* 2008;94:1216-20.

72. Samal A, Jain S. The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst Biol* 2008;2:21.

73. Pal C, Papp B, Lercher MJ. Horizontal gene transfer depends on gene content of the host. *Bioinformatics* 2005;21 Suppl 2:ii222-ii223.

74. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 2005;37:1372-5.

75. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. *Nature* 2006;440:667-70.

76. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 2007;3.

77. Varma A, Palsson BO. Metabolic Flux Balancing: Basic concepts, Scientific and Practical Use. *Nat Biotechnol* 1994;12:994-998.

78. Edwards JS, Ramakrishna R, Schilling CH, Palsson BO. Metabolic Flux Balance Analysis. In: Lee SY, Papoutsakis ET, eds. Metabolic Engineering: Marcel Deker, 1999.

79. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;14:491-6.

80. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 2008;26:659-667.

81. Edwards JS, Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A.* 2000;97:5528-5533.

82. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology* 2003;4:R54.1-R54.12.

83. Fraser-Liggett CM. Insights on biology and evolution from microbial genome sequencing. *Genome Res.* 2005;15:1603-1610.

84. Bro C, Regenberg B, Forster J, Nielsen J. In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. *Metab Eng* 2006;8:102-11.

85. Mahadevan R, Bond DR, Butler JE, Esteve-Nunez A, Coppi MV, Palsson BO, Schilling CH, Lovley DR. Characterization of Metabolism in the Fe(III)-Reducing Organism Geobacter sulfurreducens by Constraint-Based Modeling. *Appl. Environ. Microbiol.* 2006;72:1558-1568.

86. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2006;2:1-14.

87. Breitling R, Vitkup D, Barrett MP. New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol* 2008;6:156-61.

88. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;277:1453-74.

89. Kharchenko P, Vitkup D, Church GM. Filling gaps in a metabolic network using expression information. *Bioinformatics* 2004;20 Suppl 1:I178-I185.

90. Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 2004;5:76.

91. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;2:2006.0008.

92. Mayr E. This is biology : the science of the living world. Cambridge, Mass.: Belknap Press of Harvard University Press, 1997:xv, 327.

93. Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien S, Keasling JD, Maranas CD. Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metab Eng* 2007;9:387-405.

94. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;6:397-9.

95. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, 3rd, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. *Nucleic Acids Res* 2006;34:1-9.

96. Allen TE, Palsson BO. Sequenced-Based Analysis of Metabolic Demands for Protein Synthesis in Prokaryotes. *Journal of Theoretical Biology* 2003;220:1-18.

97. Mehra A, Hatzimanikatis V. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys J* 2006;90:1136-46.

98. Thomas R, Paredes CJ, Mehrotra S, Hatzimanikatis V, Papoutsakis ET. A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data. *BMC Bioinformatics* 2007;8:228.

99. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002;298:799-804.

100. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO. Matrix Formalism to Describe Functional States of Transcriptional Regulatory Systems. *PLoS Comput Biol.* 2006;2:e101.

101. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Science* 2004;59:5051-5060.

Chapter 2

The History of Network Reconstruction of *Escherichia coli* metabolism: A platform for systems analysis

2.1 Introduction

Since the release of the first genome-scale metabolic reconstruction of the *E. coli* metabolic network in 2000[1], there has been a growing number of researchers around the world adapting it for a broad range of studies[2]. The uses range from practical to obtaining basic biological understanding of cellular behavior. This range of uses is further expected to expand as the reconstruction broadens in scope and as new *in silico* methods are developed, implemented, and put to use.

In this chapter, we will describe foundational concepts central to the reconstruction process and model formulation, the history of reconstruction of the *E. coli* metabolic network, the development of reconstruction technology, and insights into the future of the field. As such, this chapter should serve as a guide to those interested in either expanding the application of the *E. coli* reconstruction or adapting established applications to other organisms.

2.2 Foundational concepts

The reconstruction of the *E. coli* metabolic network has led the development of 'bottom-up' reconstruction technology, genome-scale modeling methods, and basic

and practical uses. A number of foundational concepts have developed during this period that we introduce here to provide background and a conceptual framework for the reader (see[3,4]).

2.2.1 Forming a BiGG knowledge base

A network reconstruction is based on a highly curated set of primary biological information for a particular organism; or a biochemically, genetically and genomically structured (BiGG) knowledge base[5]. Such a knowledge base represents a large body of experimental data that is meticulously assembled and curated through the systems biology and reconstruction approaches detailed herein.

2.2.2 Genome-scale network reconstruction (GENRE)

An organism-specific BiGG knowledge base is the basis for a GENRE. The term GENRE applies to a particular organism, for example, GENRE of *Escherichia coli* (below we will see four of these, specifically called *i*JE660, *i*JR904, *i*MBEL979, and *i*AF1260). A GENRE contains a list of all the known (and some predicted) chemical transformations that are believed to take place in the particular network (e.g. metabolic, transcriptional regulatory network, etc.).

2.2.3 The central role of network reconstruction in systems biology

Systems biology research generally can be conceptualized as a four-step process (see **§1.3**). Foundational to the field is the generation of global, or genome-scale, data. The growing number of available 'omics' data types has created the need for formal and structured multi-'omic' data integration[6]. Omics data, along with legacy information (i.e., the 'bibliome') and detailed small-scale experiments, can be used to define the interactions among biological components that are used to reconstruct

networks in particular organisms[5]. Network reconstruction is also an iterative, on-going process that continually integrates data in a formal fashion as it becomes available[7]. These characteristics render the network reconstruction as a common denominator for those studying systems biology. The reconstruction effectively represents a 2-D annotation of a genome detailing not only the parts for an organism, but the interactions between specific components[8]. Genome-scale reconstruction technologies for metabolic[5], transcriptional regulation[9-11] and signaling networks[12] have been established, and transcriptional/translational network reconstruction methods are currently under development[13]. An in depth review on the bottom-up reconstruction process[4] as well as a current review of biological network reconstruction[14] have been generated.

2.2.4 Constraint-based reconstruction and analysis (COBRA)

COBRA is the overall philosophy and approach of applying constraints to limit the range of achievable functional (phenotypic) states of GENREs (outlined below). A GENRE operates under defined constraints. These constraints fall into at least four categories[4]: physico-chemical, topological, regulatory, and environmental. Such constraints can be mathematically represented and imposed on the functional states that a GENRE can take on. Functional states can be assessed using a variety of computational methods[3,4] and have been disseminated in the form of a COBRA Toolbox[15] that is a MATLAB (The MathWorks Inc., Natick, MA) based software package.

2.2.5 Converting network reconstructions into a Genome-scale Model (GEM)

A GENRE can be converted into a mathematical form (i.e., an *in silico* model) and used to computationally assess phenotypic properties (reviewed in[3]). The COBRA approach is used to analyze the properties of GENREs by assessing allowable functional states. Genome-scale reconstructions are thus a key step in quantifying the genotype-phenotype relationship and can be used to 'bring genomes to life'[16]. The availability of reconstructed metabolic networks for micro-organisms has increased rapidly in recent years and a growing number of research groups are synthesizing GENREs for target organisms of interest (see **Figure 2.2**)[5,14].

The conversion of a reconstruction (GENRE) to an *in silico* model (GEM), represented by the arrow from step 2 to step 3 in **Figure 1.1**, involves a subtle, but critical, transition. The chemical transformations of which a GENRE is comprised can be represented stoichiometrically (as well as other formats, e.g., a directed graph). Stoichiometric representations form a matrix, the rows of which represent the compounds, the columns of which represent the chemical transformations, and the entries of which are the stoichiometric coefficients. With the definition of systems boundaries and other details, a network reconstruction can be converted into a mathematical format that can be computationally interrogated. The process that this arrow represents is the bridge between the realms of high-throughput data/bioinformatics and systems science.

2.3 History of the *E. coli* metabolic network reconstruction: an ongoing and iterative process

The 18-year history of metabolic reconstruction for *E. coli*[2,7] is outlined in **Figure 2.1**. *E. coli* served as a model organism in the era of discovery of metabolic

biochemistry, and thus, comprehensive metabolic reconstructions could be developed before its genome sequence was available[17,18].With the publication of the *E. coli* genomic sequence in 1997[19], the development and use of the metabolic reconstruction in *E. coli* grew rapidly in scope.

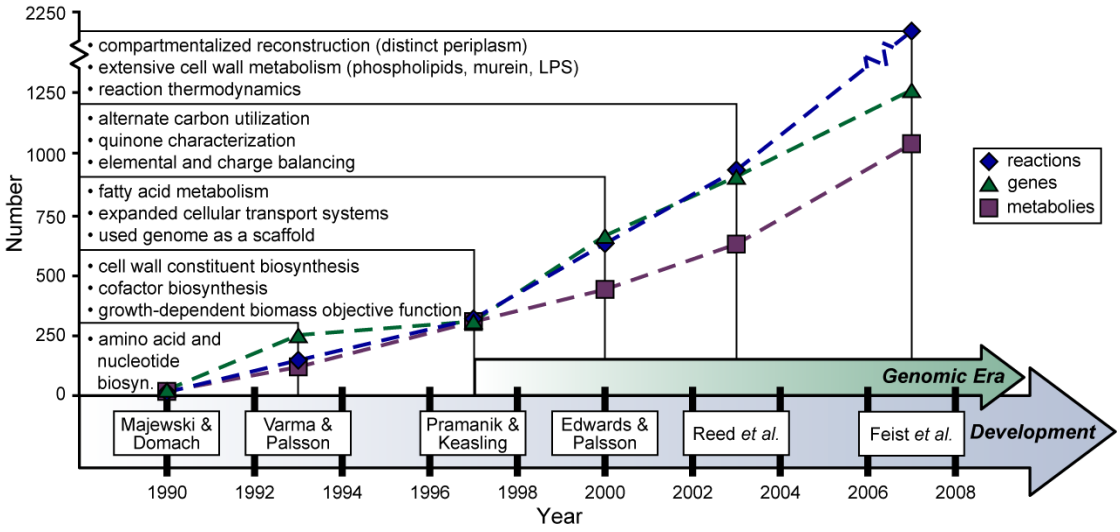


Figure 2.1: The ongoing reconstruction and history of the *E. coli* metabolic network. Shown are six milestone efforts contributing to the reconstruction of the *E. coli* metabolic network. For each of the six reconstructions [1,17,18,20-24], the number of included reactions (blue diamonds), genes (green triangles) and metabolites (purple squares) are displayed. Also listed are noteworthy properties that each successive reconstruction provided over previous efforts. For example, Varma & Palsson[17,18] included amino acid and nucleotide biosynthesis pathways in addition to the content that Majewski & Domach[20] characterized. The start of the genomic era[19] (1997) marked a significant increase in included reconstruction components for each successive iteration. The reaction, gene and metabolite values for pre-genomic era reconstructions were estimated from the content outlined in each publication and in some cases, encoding genes for reactions were unclear.

2.3.1 Pre-genome era

Beginning in 1990, a network reconstruction consisting of 14 reactions (characterizing primarily the TCA cycle and partially glycolysis) was generated to analyze the production and secretion of acetate during aerobic growth on glucose[20]. This example demonstrates the scope of initial uses of network reconstructions of *E.*

coli. Later, in 1993, a larger metabolic reconstruction consisting of 146 reactions was generated, representing key catabolic and anabolic metabolic pathways[17,18]. With its increased scope, this reconstruction was used for computing[18,25-27]:

i) Optimal production of cofactors and biosynthetic precursors,

ii) Maximum allowable generation of amino acids and nucleic acids, and

iii) Internal network flux distributions for optimal and sub-optimal growth.

The computational predictions based on the model were compared to experimental data and found to be consistent with measurements under both aerobic and anaerobic glucose minimal media conditions[26]. The comparison of computation and experimental findings in this work demonstrated the important concept of comparison to *in vivo* data as computational outcomes have to be considered as a hypothesis that need experimental confirmation.

Following these developments in the early 1990s, an expanded reconstruction consisting of 317 reactions was generated in 1997. It included cofactor and cell wall biosynthesis, and other additional metabolic pathways[21,22]. This expanded reconstruction was used for computations that incorporated measured metabolite uptake and secretion rates to predict central metabolic fluxes which were found to be consistent with enzymatic flux values determined from isotopomer based measurements[21,22]. These studies also incorporated a growth rate dependent biomass objective function that had not been considered in previous studies. It should be noted that isotopomer based measurements are also network dependent and studies are currently emerging looking specifically at this issue[28].

Note that these pre-genome era reconstructions of *E. coli* metabolism were based solely on biochemical information and provided an important foundation for subsequent work at the genomic scale.

2.3.2 Genome era

The complete genome sequence for *E. coli* K-12 MG1655 was published in 1997[19]. Its availability fueled a significant change in network reconstruction content and scope as the genome sequence directly provided a list of parts (components) present in *E. coli* (**Figure 1.1**). Utilizing the annotated sequence, a genome-scale metabolic reconstruction was generated for *E. coli* consisting of 627 unique reactions catalyzed by 660 gene products[1]. This reconstruction, later titled *i*JE660, was initially used to:

i) Predict the phenotypes for knock-out mutants of the central metabolic pathways[1],

ii) Design quantitative experiments[29], and

iii) Predict the outcome of adaptive evolution in the context of the metabolic machinery available to the cell[30].

These results demonstrated the utility of the reconstruction to understand growth characteristics of *E. coli*, the effects of gene deletions, and to point to areas of computational and experimental disagreement that identified targets for further biochemical characterization (see below).

An updated annotation of the *E. coli* K-12 MG1655 genome[31] and continual functional characterization of *E. coli* metabolic content enabled an expansion of the reconstruction in 2003 which consisted of 931 reactions catalyzed by 904 gene

products[23]. This reconstruction, titled *i*JR904, was an improvement over previous efforts in that it;

i) Contained both charge and elemental balancing of all reactions,

ii) Expanded the various carbon source utilization pathways,

iii) Contained a larger number of characterized transport systems and their encoding genes,

iv) Better accounted for quinone usage in the electron transport chain, and

v) Better detailed the relationship between given genes, proteins and reactions contained in the reconstruction (the GPR associations).

This reconstruction has been utilized for a broad number of applications reviewed later in this chapter. Utilizing the *i*JR904[23] reconstruction, an expanded reconstruction of *E. coli* was generated (containing 979 reactions and titled MBEL979) for the purpose of designing overproducing strains in the software framework, MetaFluxNet[32].

The most recent metabolic reconstruction for *E. coli*, titled *i*AF1260, incorporates data from the most recent *E. coli* K-12 MG1655 genome annotation[33] and consists of 2,077 reactions and 1,260 genes[24]. The advancements represented by *i*AF1260 over *i*JR904 lie in five main areas:

i) An increased scope with the inclusion of 357 additional ORFs,

ii) Compartmentalization into three distinct compartments (cytoplasmic, periplasmic and extra-cellular),

iii) The detailing of all grouped, or lumped, reactions (most often associated with lipid and lipopolysaccharide biosynthesis),

iv) The incorporation of reaction thermodynamics; calculated Gibbs free energy (ΔG°) values for 950 metabolites and 1935 reactions, and

v) Alignment with the EcoCyc database[34] which provided expanded coverage for the network and content mappings for further computational analyses.

This 18-year history of reconstruction of the *E. coli* metabolic network has culminated in a network containing a total number of 1,260 metabolic genes covering 28% of the 4,453 identified ORFs on the *E. coli* genome. More importantly, the 1260 ORFs represent 48% of the functionally annotated ORFs that have been confirmed by experimental data (**Table 4.1**). Thus, 92% of the 1,260 gene products included in *i*AF1260 have been experimentally verified[33] with the balance of 8% having a computationally predicted function and necessitate confirmation with focused experimentation. Model-aided gap-filling and discovery will aid in this process (see section 11.5.2). In addition, protein structures (computed or experimental) are available for a large fraction of the proteins in *i*AF1260[35]. Integration of protein structural data with the functional content of the reconstruction will lead to a better link between these two data types.

Reconstruction of the *E. coli* metabolic network is thus approaching exhaustion of known metabolic gene functions and is now being used in a prospective fashion to discover new metabolic capabilities in *E. coli* (see below). As a result of this 18-year history, the reconstruction of the *E. coli* metabolic network represents the best-developed genome-scale network to date.

2.4 Continuing development of reconstruction technology

2.4.1 Development of the reconstruction process for metabolic networks

The reconstruction process for metabolic networks is an iterative procedure, as illustrated in the previous section, that requires different types of experimental data and techniques at each phase of reconstruction. The experience with *E. coli* has led to the formulation of the workflows that underlie metabolic reconstruction. The four phases of the reconstruction process are depicted in **Figure 3.1** and the product at each phase can be used for different applications, with the number of applications increasing with network development. This procedure represents the current status of network reconstruction, and the most recent *E. coli* reconstruction, *i*AF1260, was built accordingly[24] with the advantage of starting from an already well-established reconstruction, *i*JR904. The end product of this reconstruction effort has culminated in a platform for design and discovery, and key examples of use are given later in this chapter. More extensive descriptions exist, which outline the conceptual basis[5] and the detailed process to generate genome-scale biological networks,[14] and will not be repeated here.

2.4.2 Development of the reconstruction process: beyond metabolism

The development and use of genome-scale reconstruction was rapid and many computational models were developed to address a growing spectrum of basic research and applied problems. Still, further development of reconstruction technology is necessary. The scope of reconstructions is bound to grow, representing more and more BiGG knowledge in the structured format of a GEM[36]. Growth in scope is likely to proceed in phases[2]:

i) Growth in scope in the near-term will involve the transcriptional and translational machinery[13,37-39]. Such an extension will enable a range of

studies including the direct inclusion of proteomic data, fine graining of growth requirements, and the explicit consideration of secreted protein products.

ii) Another expansion in scope in the near-term is the reconstruction of the genome-scale transcriptional regulatory network (TRN). Such reconstruction at the genome-scale is now enabled by new experimental technologies, such as ChIP-chip[40]. Experimental interrogation of the currently available TRN suggests that we know about one-fourth to one-third of its content[11], indicating that there is much to be discovered. This expectation is being confirmed with high-resolution ChIP-Chip data for *E. coli*[41]. Once reconstructed, the TRN will allow computational predictions of the context-specific uses of the *E. coli* genome and the responses of two-component signaling systems.

iii) Mid-term expansions in scope are likely to include the growth cycle, shock responses (e.g. heat and acid shock), and additional cellular functions (e.g. DNA replication and flagellar biosynthesis). Such a reconstruction should eventually be a comprehensive representation of the chemical reactions and transformation enabled by *E. coli*'s gene products.

iv) Longer-term reconstruction may begin to address the 3-dimensional organization of the bacterial cell. In particular, high-resolution ChIP-chip data on the DNA binding protein could enable the estimation of the topological arrangement of the genome, and potentially elucidate the structure of the cell wall and other cellular structures that will allow us a full 3-dimensional reconstruction of *E. coli*.

The two near-term expansions in content (i and ii in the list above) will encompass the activity of approximately 2000 ORFs in the *E. coli* genome. Clearly,

well quality-controlled reconstructions will help in guiding us to comprehensive genome-scale representation of all major cellular processes in bacteria at the BiGG data level of resolution that, in turn, enables GEMs of growing coverage and resolution.

2.4.3 Influence of the *E. coli* reconstruction on the *in silico* analysis of other micro-organisms:

The metabolic network reconstruction of *E. coli* has been influential in the generation of other organism-specific metabolic networks. The *E. coli* metabolic reconstruction has served:

i) As a content database where stoichiometrically and charge balanced reactions, and even pathways, have been incorporated into new reconstructions,

ii) As a database for defined metabolites, and

iii) As a source for a biomass objective function to query network content and functionality.

This influence has sparked an increase in the number of genome-scale network reconstructions that have been generated to formulate GEMs for a number of organisms. A detailed list of GEMs that have been developed, curated, and used for computation can be found online (http://systemsbiology.ucsd.edu/In_Silico_Organisms/Other_Organisms) Additionally, **Figure 2.2** shows the number of genome-scale reconstructions that have been developed over two year periods since 2000. The number of reconstructions generated for each period has increased since the release of the first genome-scale

reconstructions for *Haemophilus influenzae* in 1999[42] and *E. coli* in 2000[43]. Furthermore, the number of published studies utilizing the *E. coli* GEM has also increased significantly over time resulting in the applications outlined in the sections below[2].

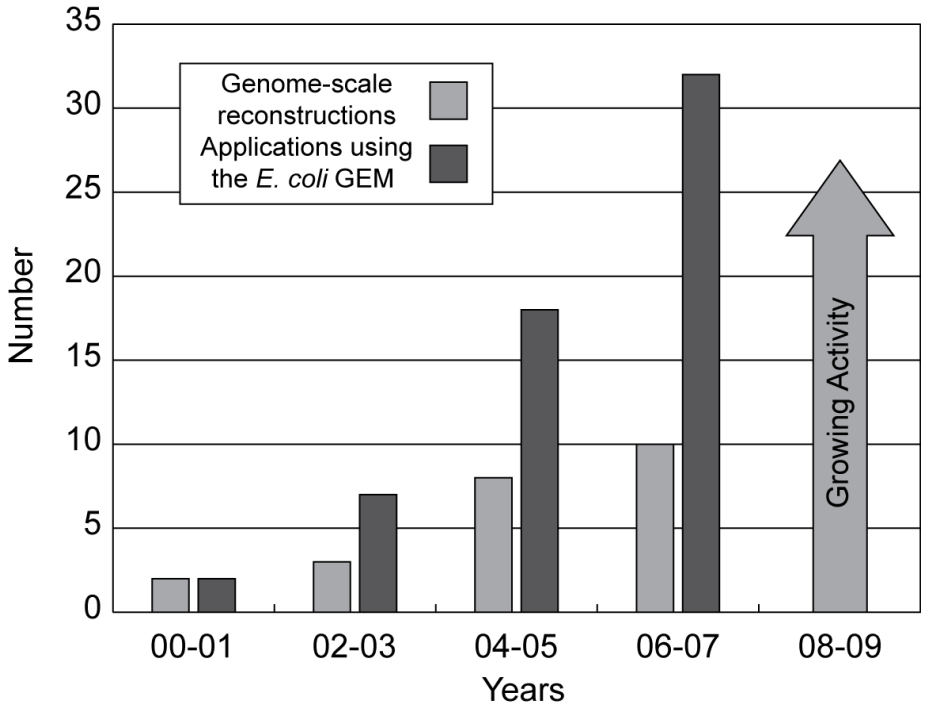


Figure 2.2: Appearance of organism-specific genome-scale reconstructions and applications of the *E. coli* metabolism reconstruction. The genome-scale reconstructions for metabolic networks that have appeared every two years since the release of the first *E. coli* GEM in 2000[1] and the number of published studies that have appeared utilizing the *E. coli* GEM[14]. Shown for the time period 2008-09, the dark bars are the numbers of each that have appeared in the first quarter of 2008, and the light bars are potential estimates for the total time period. Since the release of the first GEMs for *E. coli*[1] and that of *Haemophilus influenzae*[42], there has been a significant increase in both the number of genome-scale reconstructions and studies focused on the *E. coli* GEM for every time period.

2.5 Modeling strategy and philosophy

Models are a formal way of accounting for our knowledge about the phenomena being described. When describing biochemical reaction networks formally, we need to deal with the 'links' (i.e., the reactions) between 'nodes' (i.e., the

compounds). Our knowledge about links between biological molecules varies; from the abstract to the specific (**Figure 2.3**). Statistical models are built on correlations and a black box approach that is not mechanism based. Specific mechanisms based models are based on knowledge of chemistry, kinetics, and thermodynamics. Given the fact that kinetic and thermodynamic information is hard to obtain on a large-scale, stoichiometric models stop one step short of full specification (in the spectrum conveyed in **Figure 2.3**). The result is that we have chemistry (and its genetic basis) and network structure used as the foundation for building a mathematical description of network functions. Such models do not have a unique solution (e.g., see[4]). The lack of kinetic information can be dealt with by: 1) examining the properties of the entire set of solutions (i.e., the solutions space) or 2) by using constraint-based optimization to find specific solutions in the space[3]. The latter can be successful if we know the prevailing selection pressure on an organism and if the organism has been selected for under such conditions. The combination of network reconstruction that is based on a knowledge-base at the genome-scale and the inherent optimality properties of the selection process underlie the success of COBRA for a number of applications[2].

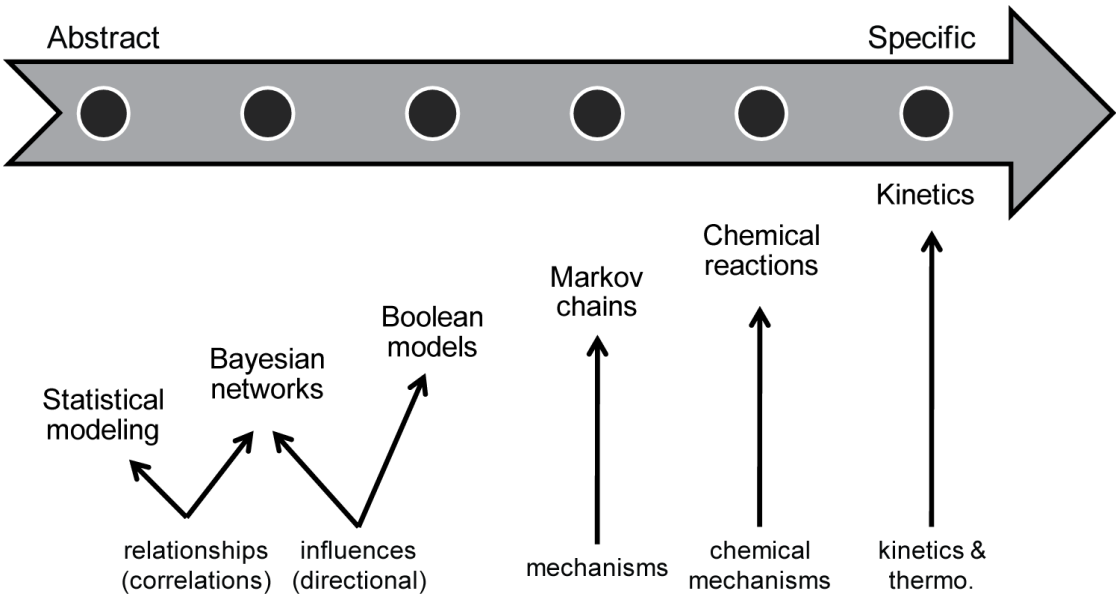


Figure 2.3: The different levels of knowledge used to generate biological models. Our knowledge about links in biochemical networks varies. At one extreme, the information is abstract and often takes the form of black-box correlations. At the other, we have detailed chemical mechanisms with kinetic and thermodynamic information. Stoichiometric models would be second from the right, accounting for mechanisms, but not incorporating kinetic and thermodynamic information.

2.6 Need for new *in silico* methods and applications

We now know how to represent BiGG data in either a stoichiometric format or in the form of causal relationships[10] and further how to use this data to perform several lines of computational inquiries. Computational query tools of GEMs will continue to be developed. New advances in these query tools will likely include, i) modularization methods, ii) use of fluxomic data, and iii) eventually kinetic information.

2.6.1 Modularization

As the scope and content of the reconstruction grows, the need to modularize its content becomes more pressing. Fine or coarse-grained views of cellular processes are needed for different applications.

2.6.2 Fluxomics

Currently, computational limitations force the reduction in network size for the analysis of isotopomer data. Given the systemic nature of fluxomic data and its phenotypic relevance, there is a pressing need to increase the size of the networks that can be utilized for experimental measurement and estimation of flux states. A network reconstruction will both guide the content that is needed for analyzing fluxomic data and offer a starting point for a rational reduction to generate relevant models in the meantime.

2.6.3 Kinetics/thermodynamics

Although detailed kinetic models of microbial functions may currently be mostly of academic interest, they will most likely be able to be constructed in the mid-term based on advances with metabolomic and fluxomic data, in addition to the developments that are occurring with the incorporation of thermodynamic information. Such large-scale kinetic models are likely to differ from those resulting from traditional approaches for construction of kinetic models as they come with different challenges.

2.7 Closing

The process underlying the *E. coli* metabolic reconstruction has pioneered many approaches, methods, and studies in the systems biology of microbial metabolism. This effort has effectively put a mechanistic basis into the genotype-phenotype relationship. In fact, this relationship is now broken down into four steps:

i) Components (a large knowledge base, BiGG), leading to

ii) Networks (the reconstruction process resulting GENRE), leading to

iii) *In silico* Models (GEMs), leading to

iv) Phenotypic States (estimated by COBRA methods)

GEMs will allow for gap-filling and systematic biological discovery[36] and for understanding of complex biological processes.

Predictive models also allow for experimental strain design. In fact, in engineering, there is '*nothing more practical than a good theory*.' As this chapter demonstrated, genomics and high-throughput technologies have enabled the construction of predictive computational models. The scope of such predictions is limited at the moment, but with the growing scope and coverage of genome-scale reconstructions and advancements in the development of computational tools, this scope will broaden. Not only will GEMs influence design in synthetic biology, but also their help with discovering cellular content will provide a more complete picture of the intra-cellular environment in which future synthetically engineered constructs and circuits will be placed. The impact of GEMs on synthetic biology is thus likely to be notable, ranging from the provision of the cellular-context of a small-scale gene circuit design to engineering of the entire genome-scale network towards fundamentally new and useful (i.e., production) phenotypes.

Finally, we can speculate about the deep scientific impact that comprehensive predictive GEMs will have on our understanding of the living process. A comprehensive view of cellular functions will allow us to study the fundamental properties of both the underlying energy and information flows in living organisms. Such a view is likely to deeply affect our understanding of both distal and proximal causation in biology.

Acknowledgements

Chapter 2, in full, is adapted from *Genome-scale reconstruction, modeling, and simulation of E. coli's metabolic network* in Systems Biology and Biotechnology of *E. coli*. Sang Yup Lee, Ed., Springer, that is scheduled to appear. The dissertation author was the primary author of this paper, which was co-authored by Dr. Bernhard Ø. Palsson and Ines Thiele.

References

1. Edwards JS, Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A.* 2000;97:5528-5533.

2. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 2008;26:659-667.

3. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

4. Palsson BO. Systems biology: properties of reconstructed networks. New York: Cambridge University Press, 2006.

5. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

6. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 2006;7:198-210.

7. Reed JL, Palsson BO. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol* 2003;185:2692-9.

8. Palsson BO. Two-dimensional annotation of genomes. *Nat Biotechnol* 2004;22:1218-9.

9. Herrgard MJ, Covert MW, Palsson BO. Reconstruction of Microbial Transcriptional Regulatory Networks. *Current Opinion in Biotechnology* 2004;15:70-77.

10. Gianchandani EP, Papin JA, Price ND, Joyce AR, Palsson BO. Matrix Formalism to Describe Functional States of Transcriptional Regulatory Systems. *PLoS Comput Biol.* 2006;2:e101.

11. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429:92-6.

12. Papin JA, Hunter T, Palsson BO, Subramaniam S. Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 2005;6:99-111.

13. Thiele I, Jamshidi N, Fleming RMT, Palsson BO. Genome-scale reconstruction of *E. coli*'s transcriptional and translational machinery: A knowledge-base and its mathematical formulation. *Under Review* 2008.

14. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microbial organisms. *Nat Rev Microbiol* 2008;Accepted.

15. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

16. Frazier ME, Johnson GM, Thomassen DG, Oliver CE, Patrinos A. Realizing the potential of the Genome Revolution: The Genomes to life Program. *Science* 2003;300:290-3.

17. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol* 1993;59:2465-73.

18. Varma A, Boesch BW, Palsson BO. Biochemical production capabilities of *Escherichia coli*. *Biotechnology and Bioengineering* 1993;42:59-73.

19. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;277:1453-74.

20. Majewski RA, Domach MM. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 1990;35:732-738.

21. Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering* 1997;56:398-421.

22. Pramanik J, Keasling JD. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and Bioengineering* 1998;60:230-238.

23. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology* 2003;4:R54.1-R54.12.

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

25. Varma A, Palsson BO. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. *Journal of Theoretical Biology* 1993;165:477-502.

26. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* 1994;60:3724-3731.

27. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. *Biotechnology and Bioengineering* 1995;45:69-79.

28. Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien S, Keasling JD, Maranas CD. Metabolic flux elucidation for large-scale models using 13C labeled isotopes. *Metab Eng* 2007;9:387-405.

29. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001;19:125-130.

30. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186-9.

31. Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M. A functional update of the Escherichia coli K-12 genome. *Genome Biol* 2001;2:RESEARCH0035.

32. Lee SY, Woo HM, Lee D-Y, Choi HS, Kim TY, Yun H. Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioproc. Eng.* 2005;10:425-431.

33. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, 3rd, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. *Escherichia coli* K-12: a cooperatively developed annotation snapshot-2005. *Nucleic Acids Res* 2006;34:1-9.

34. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* 2005;33:D334-7.

35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235-42.

36. Breitling R, Vitkup D, Barrett MP. New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol* 2008;6:156-61.

37. Allen TE, Palsson BO. Sequenced-Based Analysis of Metabolic Demands for Protein Synthesis in Prokaryotes. *Journal of Theoretical Biology* 2003;220:1-18.

38. Mehra A, Hatzimanikatis V. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophys J* 2006;90:1136-46.

39. Thomas R, Paredes CJ, Mehrotra S, Hatzimanikatis V, Papoutsakis ET. A model-based optimization framework for the inference of regulatory interactions using time-course DNA microarray expression data. *BMC Bioinformatics* 2007;8:228.

40. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 2002;298:799-804.

41. Cho BK, Knight EM, Barrett CL, Palsson BØ. Genome-wide Analysis of Fis Binding in Escherichia coli Indicates a Causative Role for A-/AT-tracts. *Genome Res.* 2008;18:900-910.

42. Edwards JS, Palsson BO. Systems properties of the Haemophilus influenzae Rd metabolic genotype. *Journal of Biological Chemistry* 1999;274:17410-6.

43. Edwards JS, and Palsson, B.O. Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. *BMC Bioinformatics* 2000;1.

Chapter 3

Metabolic Network Reconstruction of Microorganisms: The Process and Product

3.1 Abstract

Systems analysis of metabolic and growth functions in microbial organisms is rapidly developing and maturing. Such studies are enabled by the reconstruction, at the genomic-scale, of the biochemical reaction networks that underlie cellular processes. The network reconstruction process is organism-specific and is based on an annotated genome sequence, high-throughput network-wide data sets, and bibliomic data on the detailed properties of individual network components. This chapter describes the details of the process that is currently implemented to achieve comprehensive network reconstruction of a metabolic network and how it is curated and validated. The reconstruction process for genome-scale metabolic networks is well developed compared to other emerging networks, such as transcriptional regulation and for transcriptional / translational processes. This chapter should accelerate the progress of the growing number of researchers that are carrying out metabolic reconstruction for particular target organisms.

3.2 Introduction

Reconstructed networks of biochemical reactions are at the core of systems analysis of cellular processes. They form a common denominator for both

experimental data analysis and computational studies in systems biology. The conceptual basis for the reconstruction process has been outlined[1], and computational methods and tools used to characterize them have been reviewed[2,3]. Furthermore, the number of available well-curated organism-specific network reconstructions is growing (see Supplementary Data[4]) and the spectrum of their uses is broadening[5].

This chapter describes the detailed workflow that forms the basis of the metabolic reconstruction process and provides key procedural information needed for the growing number of researchers performing organism-specific reconstructions. We describe the procedures in which various experimental data types are integrated to reconstruct biochemical networks, the current status of metabolic network reconstruction, and how network reconstructions can be used in a prospective manner to discover new interactions and pathways.

3.3 Metabolic network reconstruction

Before annotated genomic sequences were available, primary literature and biochemical characterization of enzymes provided the major source of information for reconstructing metabolic networks in a select number of organisms. Accordingly, some of the earliest metabolic reconstructions that were subsequently used in modeling applications were for *Clostridium acetobutylicum*[6], *Bacillus subtilis*[7] and *Escherichia coli*[8-11].

Today, with the ability to sequence and annotate whole genomes, we can generate metabolic network reconstructions at a genomic scale, even for organisms for which little direct biochemical information is available in the published literature. To implement the metabolic reconstruction process, we need to answer the following

questions for each of the enzymes in a metabolic network: i) what substrates and products does an enzyme act on, ii) what are the stoichiometric coefficients for each metabolite participating in the reaction(s) catalyzed by an enzyme, iii) are the outlined reactions reversible, and iv) where does the reaction occur in the cell (e.g., cytoplasm, periplasm, etc.)? This data comes from a variety of sources. The establishment of a set of the chemical reactions that comprise a reaction network culminates in a database of proper chemical equations. Each reaction also has additional information associated, such as its cellular localization, thermodynamics, and genetic/genomic information. The genome-scale metabolic network reconstruction process is comprised of four fundamental steps (see **Figure 3.1**).

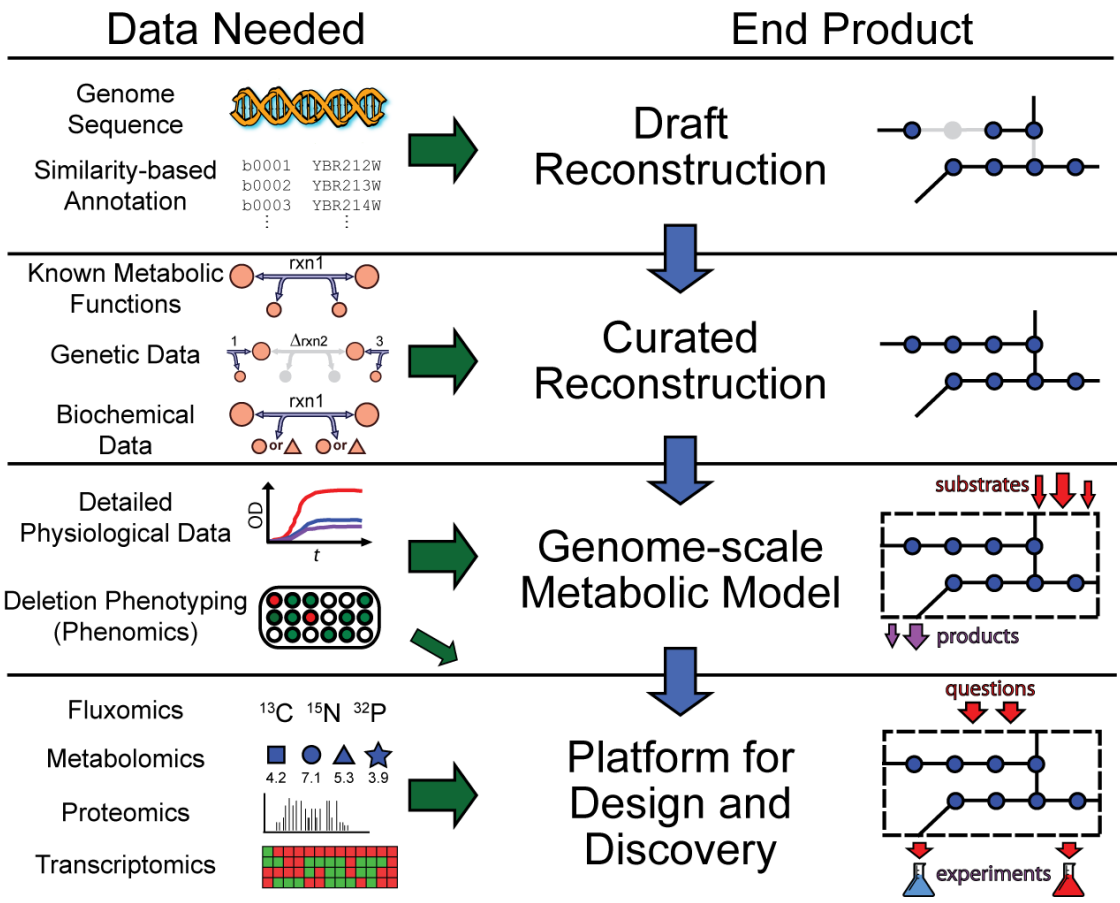


Figure 3.1: The phases and data utilized for generating a metabolic reconstruction. Genome-scale metabolic reconstruction can be summarized in four major phases, each of the latter phases building off the previous. Also characteristic of the reconstruction process is the iterative refinement of reconstruction content that is driven by experimental data and occurs in the three latter phases. For each phase, specific data types are necessary and these range from high-throughput data types (e.g., phenomics, metabolimics, etc.), to detailed studies characterizing individual components (e.g., biochemical data for a particular reaction). For example, the genome annotation can provide a parts list of a cell, whereas genetic data can provide information about the contribution of each gene product towards a phenotype (e.g., when removed or mutated). The product generated from each reconstruction phase can be utilized and applied to examine a growing number of questions with the final product having the broadest applications.

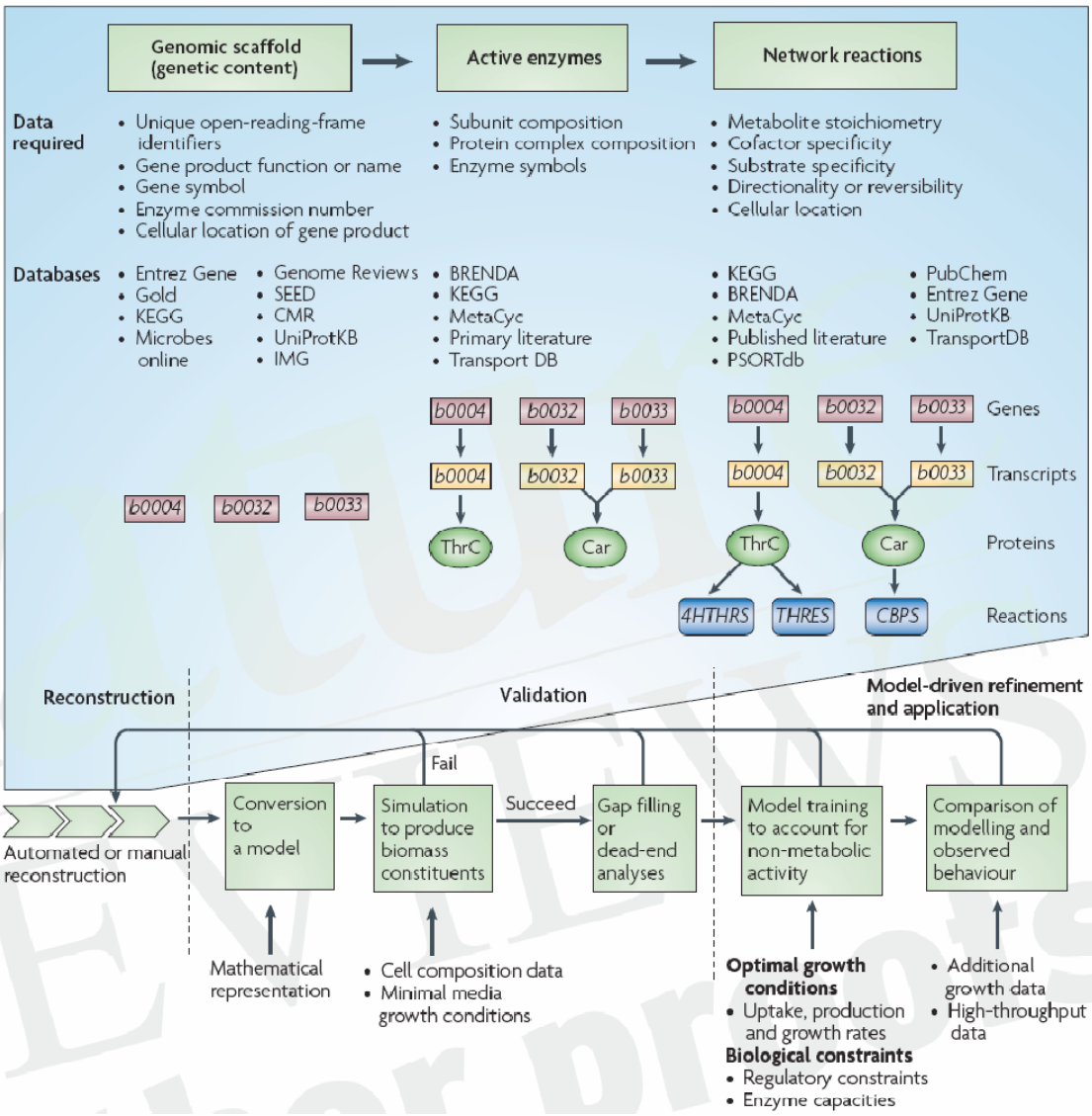
3.3.1 Step 1: Automated genome-based reconstruction.

The starting point for reconstructions is the annotated genome for a particular target organism and strain (**Figure 3.2**). Genome annotations can be found in organism-specific databases, such as EcoCyc[12] for *E. coli* and SGD[13] or CYGD[14] for *Saccharomyces cerevisiae*, or in databases with collections of genome annotations, such as EntrezGene[15], Comprehensive Microbial Resource (CMR)[16], Genome Reviews (through EBI)[17] or the Integrated Microbial Genomes (IMG)[18]. The genome annotation provides unique identifiers for the reconstruction content and a list of the metabolic enzymes thought to be present in the target organism and can indicate how the gene products interact (as subunits, protein complexes or isozymes) to form active enzymes that catalyze metabolic reactions. The next step in the reconstruction process is to determine which biochemical reactions these enzymes carry out, and this can be determined manually or by using automated tools.

Metabolic databases such as KEGG[19], BRENDA[20], MetaCyc[21], SEED[22] and Transport DB[23] contain collections of metabolic and transport reactions which have been shown to occur across a variety of organisms. Many of these databases link enzyme commission (EC) number(s) or transport commission (TC) numbers to individual or sets of reactions which have been observed biochemically in other organisms. However, substrate specificities and enzyme activities can vary between enzymes with the same EC/TC number, so the actual reactions that are catalyzed by the enzyme in the target organism may differ from that of the analogous enzyme in a reference organism. In addition, some information such as sub-cellular localization and reaction directionality might be missing but is needed for the metabolic reconstruction (see Supplementary Data[4]).

Information from metabolic databases can be extracted manually, where each active enzyme and reaction given for an organism is examined, or automated tools can be used to piece together reactions from the metabolic databases. A number of such automated tools to facilitate the reconstruction process have appeared. Some are used to map genes in the genome to reactions forming a draft metabolic network (PathwayTools[24], GEM System[25], metaShark[26], SEED[22], and others[27-29]) and others are used to refine the networks by filling in missing reactions (SMILEY algorithm[30], GapFind / GapFill[31], PathoLogic[32]) or by evaluating reaction directionality[33,34]. The later methods improve draft reconstructions built from gene to reaction mapping via databases, as they can correct incorrect or missing information from metabolic databases and/or genome annotations. Since automated methods rely heavily on metabolic and transport databases, along with genome annotations, errors will propagate into the reconstructed networks. A table of common issues encountered during automated network reconstruction has been generated[4] as a guide for use of such methods and should enable further advancement of such tools.

Figure 3.3: Reconstruction, validation and utilization of a metabolic reconstruction. (facing page) The process of metabolic reconstruction can be performed in a sequential manner. The process is initiated by obtaining the genetic content (that is, a parts list of the cell) from the genome annotation. Active enzymes on this scaffold are associated with the genetic content by using information from databases and published literature. The metabolic reactions that these enzymes catalyse are then delineated and a gene to protein to reaction association is ultimately generated. Automated reconstruction tools are available to aid in this process and several databases possess the necessary information for each data type (see below). Following the initial reconstruction process, a reconstruction is converted to a model in a mathematical format that can be used for computation. Further in the validation phase, the ability of the organism to produce biomass constituents and grow is examined using a biomass objective function (Figure 3.3). This analysis functionally tests the reconstruction for an experimentally observed phenomenon. A dead-end analysis should follow, for which computational algorithms are available (see the main text), to examine reactions on a pathway basis for their physiological role. For predictions of physiological behavior, a training data set is needed to examine non-metabolic energy needs and organism-specific components (for example, the electron transport system). In this phase, additional known key network properties can be applied in addition to the metabolic functions outlined in the reconstruction (for example, key regulatory interactions under a given condition) to improve predictive capabilities. For prospective use, high- and low-throughput data can also be compared with modeling simulations to validate the content and make predictions or find specific areas of disagreement between the functionality of the currently characterized content and experimental observations. BRENDA[19], CMR[15], Entrez Gene[14], Genome Reviews[16], GOLD[23], IMG[17], KEGG[18], MetaCyc[20], Microbes Online[24], PSORTdb[25], PubChem[26], SEED[27], Transport DB[22], UniProtKB[28].


Genomic scaffold (genetic content)
Active enzymes
Network reactions
Data required
• Unique open-reading-frame identifiers
• Gene product function or name
• Gene symbol
• Enzyme commission number
• Cellular location of gene product
• Subunit composition
• Protein complex composition
• Enzyme symbols
• Metabolite stoichiometry
• Cofactor specificity
• Substrate specificity
• Directionality or reversibility
• Cellular location
Databases
• Entrez Gene
• Gold
• KEGG
• Microbes online
• Genome Reviews
• SEED
• CMR
• UniProtKB
• IMG
• BRENDA
• KEGG
• MetaCyc
• Primary literature
• Transport DB
• KEGG
• BRENDA
• MetaCyc
• Published literature
• PSORTdb
• PubChem
• Entrez Gene
• UniProtKB
• TransportDB
b0004
b0032
b0033
Genes
Transcripts
ThrC
Car
Proteins
4HTHRS
THRES
CBPS
Reactions
Reconstruction
Validation
Model-driven refinement and application
Automated or manual reconstruction
Conversion to a model
Simulation to produce biomass constituents
Fail
Succeed
Gap filling or dead-end analyses
Model training to account for non-metabolic activity
Comparison of modelling and observed behaviour
Mathematical representation
• Cell composition data
• Minimal media growth conditions
Optimal growth conditions
• Uptake, production and growth rates
Biological constraints
• Regulatory constraints
• Enzyme capacities
• Additional growth data
• High-throughput data

3.3.2 Step 2: Curating the draft reconstruction.

While the automated extraction of metabolic reactions from databases gives an initial set of candidate biochemical reactions encoded on a genome, they cannot establish certain organism-specific features such as substrate or cofactor specificity and sub-cellular localization. Such information requires domain-specific knowledge of the organism. Therefore, the draft network reconstruction needs to be manually curated, ideally with input from organism-specific experts. An automatically reconstructed metabolic network will be incomplete, and it will have gaps and may also contain mistakenly included reactions that may actually not occur in the target organism. Manual curation is thus necessary to add and correct information that the automatic procedures misses or misplaces in the initial network reconstruction. While the automated reconstruction step is rapid, the manual curation process is labor intensive and at times tedious.

Organism-specific databases, textbooks[41-44], primary publications, review articles and experts familiar with the legacy data for an organism are the main sources of information for the manual curation step. These detailed sources contain information about properties such as reaction directionality and location that is not always found in more general databases. For example, protein localization studies[45] can be used to assign metabolic reactions to sub-cellular compartments. Similarly, biochemical studies of enzymes from the target organism (or a closely related organism) can provide information on reversibility and substrate specificity specific to that organism. These sources of information provide more direct evidence for the inclusion of specific reactions in the metabolic reconstruction. The availability of such sources for a given organism is highly variable[46]. The goal of manual reconstruction is to fill in gaps or holes in the network by inference or through direct evidence in the

available literature on the organism or its close relatives. Gap-filling is further discussed below and examples of gap-filling in metabolic networks has been generated (see Supplementary Data[4]).

A high-quality network reconstruction is thus, based on a combination of automated genome-based procedures coupled with detailed and laborious literature-based manual curation. This process effectively creates a biochemically, genomically and genetically (BiGG) structured knowledge base that is both organism-specific and available to all researchers working with the target organism. All the reactions placed in a BiGG knowledge base form a genome-scale network reconstruction (GENRE). GENREs are formed in an iterative fashion (for example, *E. coli*[47,48]) as the corresponding BiGG knowledge base grows for the target organism, based on new experimental data or new genome-annotation.

3.3.3 Step 3: Converting a genome-scale reconstruction to a computational model.

Before a reconstruction can be used for computations of network and/or physiological capabilities, there is a subtle, but critical step where a reconstruction is converted to a mathematical representation[27,28] (**Figure 3.2**). This conversion translates a GENRE into a mathematical format that becomes the basis for a genome-scale model (GEM). Subsequent computations serve as a way to interrogate data consistency and to compute which functions a reconstructed network can and cannot carry out.

Representation of a network in a mathematical format enables the deployment of a large range of computational tools to analyze network properties. These computational tools focus on the evaluation of network systemic properties and which

functions a network can perform under the physico-chemical constraints placed on the cell. This step competes the so called constraint-based reconstruction and analysis (COBRA) framework[1] for the target organism. Multiple computational platforms have been developed, which apply constraint-based methods to metabolic GEMs[3,49,50]. In addition to the stoichiometric representation, metabolic networks are commonly analyzed as graphs[51] or using a pathway or subsystem-based approach[52], but these essentially non-parametric approaches are not discussed further here.

With a mathematical representation and computational platform, the generation of a biomass objective function is necessary to compute a network's ability to support growth (**Figure 3.3**). Here, the macromolecular composition of the cell (and the building blocks which are used to generate them) is utilized to define a necessary functionality that the network must be able to execute. A useful consistency check performed on reconstructed networks is to use them to compute growth rates under a given condition. The set of experimental data necessary to perform such analysis includes, i) the composition of cellular biomass, ii) the composition of the minimal growth media necessary to support growth *in vivo*, and iii) a training data set including growth rate and substrate uptake rates. Phenotypic data (growth rates and uptake and secretion rates) can be obtained through growth experiments in minimal or complex media by monitoring media components. This data is typically available in published cell characterization studies, but may need to be generated for a specific organism-of-interest. Cellular biomass composition data is obtainable through experimental assays which determine overall cellular composition and further experimentation cataloging the breakdown of each macromolecule of the cell (this information has been cataloged extensively for *E. coli*[53]). With essentiality data (gene and/or cellular content), this equation can be refined[48]. Genome-scale gene

essentiality data sets are appearing for model organisms (listed in[54]), and these data sets are often times available through specific projects or organism-specific databases, such as the SGD yeast online database[55]. Overall, the analysis and testing of a network's ability to produce biomass components is often utilized to curate metabolic networks (see Supplementary Data[4]).

Aside from simulations to produce biomass constituents, i) additional gap-filling analyses can be performed to add missed pathways or to remove any pathways that have been incorrectly included from the automated reconstruction process, and ii) additional cellular objective functions can be evaluated computationally to understand cellular behavior[56,57]. The current state of gap-filling of metabolic networks has been recently reviewed[58].

Once gap-filling analyses are complete, additional steps are necessary to account for strain-specific parameters and non-metabolic activities in modeling simulations. In this phase, growth data is necessary to understand and quantify these key physiological parameters. Two major factors to consider during this phase are the stoichiometry for translocation (or energy-coupling) reactions and maintenance parameters[48,59]. Translocation reactions differ than other reactions in the network because the mass and energy balances around these ion pumping components are difficult to measure experimentally. Therefore characterizing reactions of this type is challenging, but can be accomplished given the proper experimental data (see Supplementary Data[4]). After this phase is complete, a model can be applied to study the specific growth condition from which the training data was based and can be used to explore additional environmental conditions.

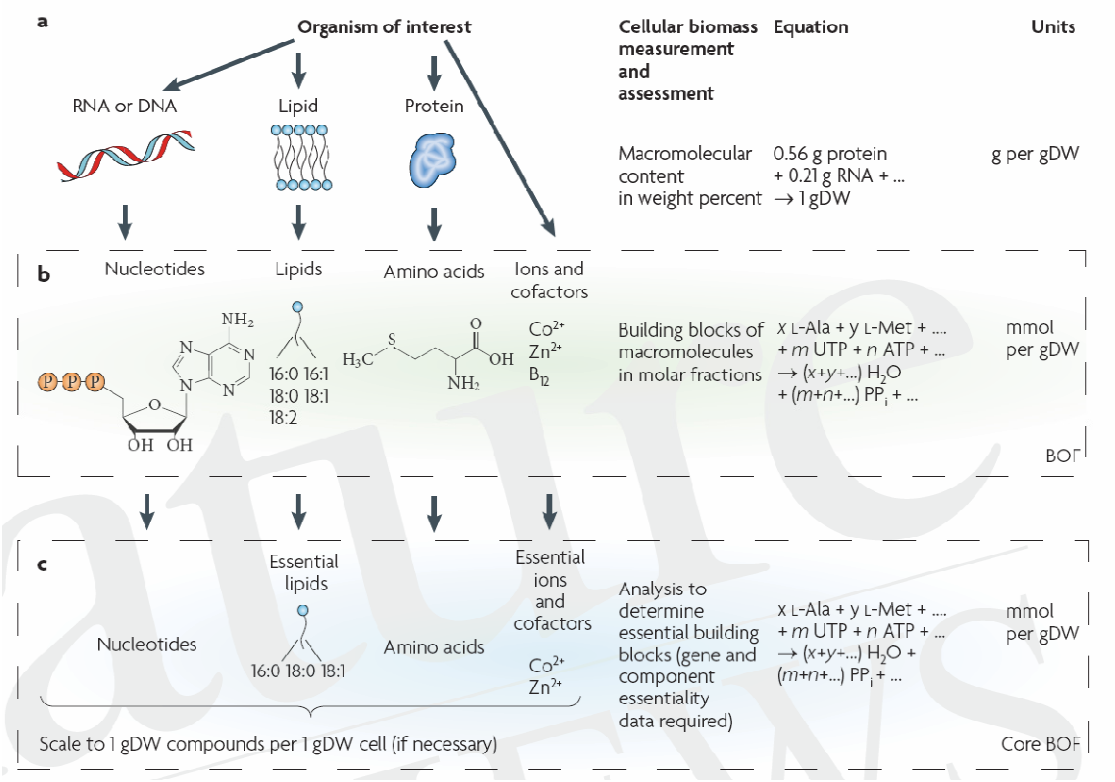


Figure 3.3: Procedure to generate a biomass objective function. An organism-specific biomass objective function (BOF) can be used to test the functionality of a network by examining the fundamental property of cellular growth and regeneration. The BOF, a known growth-supporting media condition and a reconstruction in a mathematical format are necessary for this test. Starting with the organism of interest, the macromolecular weight percent contribution of each component is determined (see the figure, part a). These data can be generated using readily available assay kits. Each macromolecule is then broken down into the cellular building blocks that constitute the macromolecule or those that are necessary to synthesize the macromolecule in terms of molar fractions (see the figure, part b). The building block will often be physiologically present in the network (for example, lipid molecules), but in some cases, the most appropriate metabolite in the network is used to generate the BOF (for example, protein is broken down into individual amino acids and the net product of protein synthesis, water). With the availability of gene and/or component essentiality data, a core BOF can be generated that possesses different metabolites compared with the wild-type BOF. In formulating the core BOF, gene essentiality data are used together with the pathway context to determine the most basic macromolecule that is necessary for cell viability (see the figure, part c). Alternatively, published data that determine minimally essential biomass components can be incorporated to generate the core BOF. A core BOF can be used in simulations to more accurately examine essential components or aspects of the network. This process ultimately results in a BOF (or BOFs) in mmol per gram of dry weight (gDW) that can be used to evaluate an organism-specific network.

3.3.4 Step 4: Reconstruction uses and integration of high-throughput data.

High-throughput data sets which evaluate a large number of interactions across different growth or genetic conditions can be utilized to refine and expand the metabolic content of a network. These types of comparisons and analyses have the potential to truly evaluate genome-scale omics data sets in an integrated manner by placing them in a functional and structured context. Several successful studies have been conducted for microbial species to uncover new metabolic knowledge using systematic data-driven discovery (**Table 3.1**). The necessary data types to support studies of discovery and expansion, as well as pilot studies for discovery have been recently reviewed[58]. Briefly, these studies fall into three categories, i) studies that have utilized a reconstruction to examine topological network properties, ii) those that have utilized a reconstruction in constraint-based modeling for quantitative or qualitative analyses, and iii) studies that are purely data driven.

One particular example of systematic data-driven discovery integrated a number of data types and GEM modeling to annotate unknown gene functions in *E. coli*[30]. In this analysis, an iterative process was utilized to, i) identify discrepancies between modeling predictions and high-throughput growth phenotyping data (Biolog data, http://www.biolog.com), ii) determine potential reactions which remedy disagreements (and the ORFs that might encode proteins to catalyze them) through a computational analysis, and iii) characterize targeted ORFs experimentally to confirm their function. To drive discovery, this approach analyzed a variety of data types (i.e., phenotyping, gene expression, and enzyme activity) to hypothesize and validate computational predictions. This one example demonstrates the promise of integrating modeling results and experimental, data and will likely become a key approach to

expanding current metabolic knowledge along with aiding discovery of new components and interactions in cellular processes.

Table 3.1: Approaches for systematic data-driven discovery of new pathways or enzymes.

Data type	Discovery type	References
Growth in diverse media conditions	New substrate utilization pathways	30
Deletion strain growth phenotyping	Alternative pathway discovery	60,61
Synthetic lethal interactions		
Systematic in vitro enzymatic assays	New metabolic reactions and pathways	62
Metabolomics	New metabolite utilization/production pathways	63
Proteomics	Candidate genes for filling network gaps	64-66
Transcriptomics		
Genomic neighborhood		

3.4 The effects of missing network content

An important issue in the conversion of a network reconstruction into a predictive computational model is the coverage and accuracy of available data from which the network was reconstructed. Therefore, it is important to understand the impact and influence network components can have on computational results. Intended use examples of *in silico* models are used to help understand this issue.

Qualitative predictions obtained using GEMs (e.g., will an organism grow given an environmental or genetic perturbation, or does expression of gene increase

or decrease) are likely to be less sensitive than quantitative predictions (e.g., what is the cellular growth rate or what level of gene expression is expected) to errors in the network content. This expectation is due to the fact that qualitative predictions are compared to binary outcomes (i.e. digital outcomes), rather than a range of numerical values (i.e., analog outcomes). If one is generating qualitative predictions regarding growth phenotypes, the effect of omitting an individual reaction from a network does not greatly affect your results. For example, removing approximately 87% of the 2077 reactions individually from an *E. coli* metabolic model (*i*AF1260[48]) did not affect the qualitative growth predictions for a given environmental condition.

In depth studies have been performed to assess the influence of individual network components, input parameters, and the querying methods used to probe GEMs on computational predictions. The results from these studies can be used to gauge the influence of the content of reconstructions. These analyses include examining input/output values[48,59,67,68], BOF composition[48,56,59,69,70], querying methods[71,72], and network components[61,68,73-77]. These initial studies demonstrate the necessity to identify the scope and intention of GENRE applications *a priori* and further show how computational analysis can help to identify missing components and errors when computational results are compared to biological functions. The latter model-driven gap-filling approach is expected to continue to develop and lead to GEMs with improved predictive capabilities.

3.5 Conclusions

The reconstruction process relies on workflows that organize and integrate various data types and other relevant information about the network of interest. Over the past ten years, such workflows have been developed for genome-scale metabolic

networks to the point where they represent BiGG knowledge bases and are in wide use. More recently, similar methods are being developed for other cellular processes such as transcriptional regulation, and for transcription and translation. The implementation of these workflows for a growing number of organisms should accelerate the systems analysis in a single organism, in communities of organisms, and through phyla. The workflow outlined herein have been implemented and enabled a wide variety of analyses[5]. To facilitate wider use and the development of additional analysis procedures, improvements in the distribution of GENREs in needed. Two areas that will aid distribution and usage are the standardization of a reconstruction format (e.g., SBML[78]) and available reconstruction database where they can be accessed.

It is expected that the reconstruction process will continue to grow in scope, depth and accuracy, and it should continue to enable a broadening spectrum of basic and applied studies. The availability of high-quality comprehensive reconstructions will accelerate the implementation of the systems biology paradigm (i.e., biological components *to* networks *to* computational models *to* phenotypic studies) and will thus help realize the broad transformative potential of this paradigm in the life sciences. Network reconstructions are a key factor in building a mechanistic genotype-phenotype relationship. Quantitative genotype-phenotype relationships have been best established for bacterial metabolism[5] to date and this review should aid new practitioners to build such relationships for their target organisms.

Acknowledgements

We would like to thank A. Osterman and N. Jamshidi for their insights.

Chapter 3, in full, is adapted from Reconstruction of biochemical networks in microbial organisms that is scheduled to appear in Nature Reviews Microbiology. The dissertation author was the primary author of this paper, which was co-authored by Dr. Markus J. Herrgård, Ines Thiele, Dr. Jennie L. Reed, and Dr. Bernhard Ø. Palsson.

References

1. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

2. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

3. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

4. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microbial organisms. *Nat Rev Microbiol* 2008;Accepted.

5. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 2008;26:659-667.

6. Papoutsakis ET. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and Bioengineering* 1984;26:174 - 187.

7. Papoutsakis E, Meyer C. Fermentation equations for propionic-acid bacteria and production of assorted oxychemicals from various sugars. *Biotechnology and Bioengineering* 1985;27:67-80.

8. Papoutsakis E, Meyer C. Equations and calculations of Product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnology and Bioengineering* 1985;27:50-66.

9. Majewski RA, Domach MM. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 1990;35:732-738.

10. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol* 1993;59:2465-73.

11. Varma A, Boesch BW, Palsson BO. Biochemical production capabilities of *Escherichia coli*. *Biotechnology and Bioengineering* 1993;42:59-73.

12. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, Bonavides-Martinez C, Ingraham J. Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res* 2007.

13. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, Hong EL, Issel-Tarver L, Nash R, Sethuraman A, Starr B, Theesfeld CL, Andrada R, Binkley G, Dong Q, Lane C, Schroeder M, Botstein D, Cherry JM. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* 2004;32 Database issue:D311-4.

14. Guldener U, Munsterkotter M, Kastenmuller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, Michael H, Kaps A, Talla E, Dujon B, Andre B, Souciet JL, De Montigny J, Bon E, Gaillardin C, Mewes HW. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* 2005;33:D364-8.

15. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2007;35:D26-31.

16. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O. The Comprehensive Microbial Resource. *Nucleic Acids Res* 2001;29:123-5.

17. Stoesser G, Tuli MA, Lopez R, Sterk P. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res* 1999;27:18-24.

18. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Lykidis A, Mavromatis K, Ivanova N, Kyrpides NC. The integrated microbial genomes (IMG) system. *Nucleic Acids Res* 2006;34:D344-8.

19. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30.

20. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;32:D431-3.

21. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes, 2004:D438-442.

22. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, Best A. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* 2007;8:139.

23. Ren Q, Chen K, Paulsen IT. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 2007;35:D274-9.

24. Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics* 2002;18 Suppl 1:S225-32.

25. Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* 2006;7.

26. Pinney JW, Shirley MW, McConkey GA, Westhead DR. metaSHARK: software for automated metabolic network prediction from DNA sequence and its application to the genomes of Plasmodium falciparum and Eimeria tenella. *Nucleic Acids Res* 2005;33:1399-409.

27. Borodina I, Nielsen J. From genomes to in silico cells via metabolic networks. *Curr Opin Biotechnol* 2005;16:350-5.

28. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B. Accelerating the reconstruction of genome-scale metabolic networks. *BMC Bioinformatics* 2006;7:296.

29. Goesmann A, Haubrock M, Meyer F, Kalinowski J, Giegerich R. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics* 2002;18:124-9.

30. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 2006;103:17480-4.

31. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 2007;8:212.

32. Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics* 2004;5:76.

33. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys J* 2006;90:1453-61.

34. Kümmel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 2006;7.

35. Bernal A, Ear U, Kyrpides N. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res* 2001;29:126-7.

36. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP. The MicrobesOnline Web site for comparative genomics. *Genome Res* 2005;15:1015-22.

37. Rey S, Acab M, Gardy JL, Laird MR, deFays K, Lambert C, Brinkman FS. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res* 2005;33:D164-8.

38. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008;36:D13-21.

39. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Ruckert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 2005;33:5691-702.

40. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004;32:D115-9.

41. Neidhardt FC. *Escherichia coli* and *Salmonella*: cellular and molecular biology. Washington, D.C.: ASM Press, 1996:2 v. (xx, 2822 , lxxvii).

42. Dickinson JR, Schweizer M. The metabolism and molecular physiology of Saccharomyces cerevisiae. London ; Philadelphia: Taylor & Francis Ltd, 2004:xii, 343.

43. Marre R. Legionella. Washington, D.C.: ASM Press, 2002.

44. Mobley HLT, Mendz GL, Hazell SL. Helicobacter Pylori. Washington, D.C.: ASM Press, 2001.

45. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. Global analysis of protein localization in budding yeast. *Nature* 2003;425:686-91.

46. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;6:397-9.

47. Reed JL, Palsson BO. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol* 2003;185:2692-9.

48. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. A genome-scale metabolic reconstruction for

Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007;3.

49. Lee SY, Woo HM, Lee D-Y, Choi HS, Kim TY, Yun H. Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioproc. Eng.* 2005;10:425-431.

50. Klamt S, Saez-Rodriguez J, Gilles ED. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol* 2007;1:2.

51. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5:101-13.

52. Kwast KE, Lai LC, Menda N, James DT, 3rd, Aref S, Burke PV. Genomic analyses of anaerobically induced genes in Saccharomyces cerevisiae: functional roles of Rox1 and other factors in mediating the anoxic response. *J Bacteriol* 2002;184:250-65.

53. Neidhardt FC, Umbarger HE. Chemical Composition of *Escherichia coli*. In: Neidhardt FC, ed. Escherichia coli and Salmonella : cellular and molecular biology. Washington, D.C.: ASM Press, 1996:13-16.

54. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. *J Bacteriol* 2006;188:8259-8271.

55. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D. SGD: Saccharomyces Genome Database. *Nucleic Acids Research* 1998;26:73-9.

56. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 2007;3.

57. Knorr AL, Jain R, Srivastava R. Bayesian-based selection of metabolic objective functions. *Bioinformatics* 2007;23:351-7.

58. Breitling R, Vitkup D, Barrett MP. New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol* 2008;6:156-61.

59. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. *Biotechnology and Bioengineering* 1995;45:69-79.

60. Shlomi T, Herrgard M, Portnoy V, Naim E, Palsson BO, Sharan R, Ruppin E. Systematic condition-dependent annotation of metabolic genes. *Genome Res* 2007.

61. Harrison R, Papp B, Pal C, Oliver SG, Delneri D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A* 2007;104:2307-12.

62. Saito N, Robert M, Kitamura S, Baran R, Soga T, Mori H, Nishioka T, Tomita M. Metabolomics approach for enzyme discovery. *J Proteome Res* 2006;5:1979-87.

63. Chiang KP, Niessen S, Saghatelian A, Cravatt BF. An enzyme that regulates ether lipid signaling pathways in cancer annotated by multidimensional profiling. *Chem Biol* 2006;13:1041-50.

64. Fuhrer T, Chen L, Sauer U, Vitkup D. Computational prediction and experimental verification of the gene encoding the NAD+/NADP+-dependent succinate semialdehyde dehydrogenase in Escherichia coli. *J Bacteriol* 2007;189:8073-8078.

65. Popescu L, Yona G. Automation of gene assignments to metabolic pathways using high-throughput expression data. *BMC Bioinformatics* 2005;6:217.

66. Rodionov DA, Kurnasov OV, Stec B, Wang Y, Roberts MF, Osterman AL. Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. *Proc Natl Acad Sci U S A* 2007;104:4279-84.

67. Cakir T, Efe C, Dikicioglu D, Hortacsu A, Kirdar B, Oliver SG. Flux balance analysis of a genome-scale yeast model constrained by exometabolomic data allows metabolic system identification of genetically different strains. *Biotechnol Prog* 2007;23:320-6.

68. Vemuri GN, Eiteman MA, McEwen JE, Olsson L, Nielsen J. Increasing NADH oxidation reduces overflow metabolism in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2007;104:2402-7.

69. Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering* 1997;56:398-421.

70. Pramanik J, Keasling JD. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and Bioengineering* 1998;60:230-238.

71. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 2002;99:15112-7.

72. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A* 2005;102:7695-700.

73. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2006;2:1-14.

74. Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* 2004;14:2367-76.

75. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 2004;427:839-843.

76. Burgard AP, Vaidyaraman S, Maranas CD. Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol Prog* 2001;17:791-7.

77. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 2003;5:264-76.

78. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin, II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;19:524-31.

Chapter 4

The metabolic reconstruction and computational analysis of the bacterium *Escherichia coli* K-12 MG1655, *i*AF1260

4.1 Abstract

An updated genome-scale reconstruction of the metabolic network in *E. coli* K-12 MG1655 is presented. This updated metabolic reconstruction includes; 1) an alignment with the latest genome annotation and the metabolic content of EcoCyc leading to the inclusion of the activities of 1260 ORFs, 2) characterization and quantification of the biomass components and maintenance requirements associated with growth of *E. coli*, and 3) thermodynamic information for the included chemical reactions. The conversion of this metabolic network reconstruction into an *in silico* model is detailed. A new step in the metabolic reconstruction process termed thermodynamic consistency analysis is introduced, in which reactions were checked for consistency with thermodynamic reversibility estimates. Applications demonstrating the capabilities of the genome-scale metabolic model to predict high-throughput experimental growth and gene-deletion phenotypic screens are presented. The increased scope and computational capability using this new reconstruction is expected to broaden the spectrum of both basic and applied systems biology studies of *E. coli* metabolism.

4.2 Introduction

The process of extracting biochemical content from genome annotations and literature sources to computationally catalog and interconnect the metabolic pathways available to the cell (i.e., metabolic reconstruction) is well established and has been carried out for a growing number of organisms on the genome-scale[1]. This network reconstruction process ultimately results in the generation of a biochemically, genomically and genetically (BiGG) structured database, which can be further utilized for both mathematical computation and analysis of high-throughput data sets. Goals of such computation and data integration efforts are to gain a better understanding of the observable phenotypes and coordinated functions of the cell, as well as to apply developed *in silico* models for biological discovery and engineering applications. For mathematical computation, a number of methods have been developed to characterize models built from a metabolic reconstruction[2,3], and reconstructions are becoming increasingly important in understanding high-throughput experimental data[4]. Thus, a well-curated metabolic reconstruction has a variety of uses and is of common interest to those studying systems biology relating to cellular metabolism.

The Gram-negative rod-shaped bacterium, *Escherichia coli*, has been an ideal target for metabolic reconstruction since it is arguably the most studied and best characterized microorganism in terms of its genome annotation, functional characterization and knowledge of growth behavior[5,6]. Reconstruction of the metabolic network of *E. coli* has been progressing since 1990 (reviewed in[7]). This network reconstruction has gone through a series of expansions and refinements[8-15] with each iteration building on previous work while incorporating new knowledge.

Applications utilizing the *E. coli* reconstruction have had implications in a number of fields (for a list of applications and references, see http://gcrg.ucsd.edu/organisms/ecoli/ecoli_others.html). For metabolic engineering applications, modeling enables examination and simulation of metabolism as a whole, circumventing the possible shortcomings of methods that rely on manual assessment of a limited number of interactions and possibly fail to detect non-intuitive causal interactions[16,17]. For studies of bacterial evolution, a reconstruction serves as a highly-curated database and model to examine and simulate evolutionary hypotheses[18,19]. Network analyses have been applied to genome-scale reconstructions of *E. coli* to identify sets of reactions or metabolites whose activity is interdependent. These studies have obvious implications in aiding therapeutic interventions along with other systemic analyses[20,21]. Additionally, for the prospective goal of biological discovery, genome-scale reconstructions drive discovery by identifying specific areas where knowledge is lacking, or disagreements with observations, and provide a framework for the integration of high-throughput data[22,23].

In this study, we expand and refine the reconstruction of the metabolic network in *E. coli*. The new additions include: 1) an up to date accounting for ORFs in *E. coli* that have metabolic annotations and an alignment of the content in EcoCyc[24] leading to the inclusion of 1260 ORFs (an increase of 356 ORFs over the previous reconstruction[14]), 2) an improved breakdown of the biomass composition, the maintenance requirements for growth and sustenance and a sensitivity analysis on the parameters used in computational modeling, and 3) thermodynamic information about the chemical transformations accounted for in the reconstruction. The thermodynamic properties estimated for the model reactions and compounds were utilized to test the thermodynamic consistency of the reactions included in the

reconstruction[25]. This expanded version of the *E. coli* metabolic network will allow for additional and more comprehensive computational and experimental studies of the systems properties of *E. coli* metabolism. We give several such examples that use the new network reconstruction.

4.3 Results

The results of the present study are presented in three parts. First, we describe the new content added to form the updated genome-scale *E. coli* metabolic reconstruction. Second, we detaile the conversion of the metabolic reconstruction into a computational model for physiological studies. Third, we present a series of applications and detailed biochemical studies that the new computational model enables.

4.3.1 Reconstruction Content and Enhancements

We generated a metabolic reconstruction consisting of the chemical reactions that transport and interconvert metabolites within *E. coli* K-12 MG1655. This network reconstruction, termed *i*AF1260, was based on a previous reconstruction, *i*JR904[14], the current functional annotation of the *E. coli* genome[26], content characterized in the EcoCyc Database[24] and specific biochemical characterization studies on the metabolic machinery and capabilities of *E. coli* (see Supplementary Data[27] for a complete list of references). The general features of *i*AF1260 are given in **Table 4.1**. When possible, enzymatically catalyzed reactions were linked to their corresponding open reading frames (ORFs) through gene-protein-reaction (GPR) assignments (see Methods and[1]). A full list of all reactions and metabolites for the reconstruction is

given in the supplementary information in spreadsheet and SBML formats[27] and is also available on the web on the BiGG database (http://bigg.ucsd.edu).

Table 4.1: Properties of *i*AF1260 and *i*JR904.

	*i*AF1260 this study	*i*JR904 Reed *et al*, 2003
Included genes	1260 (28%)[d]	904 (20%)[d]
Experimentally-based function	1161 (92%)	838 (93%)[e]
Computationally predicted function	99 (8%)	58 (6%)[e]
Unique functional proteins	1148	817
Multigene complexes	167	105
Genes involved in complexes	415	289
Instances of isozymes[a]	346	149
Reactions	2077	931
Metabolic reactions	1387	747
Unique metabolic reactions[b]	1339	745
Cytoplasmic	1187	745
Periplasmic	192	0
Extracellular	8	2
Transport reactions	690	184
Cytoplasm to periplasm	390	0
Periplasm to extracellular	298	0
Cytoplasm to extracellular	2	184
Gene—protein—reaction associations		
Gene associated (metabolic/ transport)	1294/625	706/166
Spontaneous/diffusion reactions[c]	16/9	2/9
Total (gene associated and no association needed)	1310/634 (94%)	708/175 (95%)
No gene association (metabolic/ transport)	77/56 (6%)	37/9 (5%)
Exchange reactions	304	143
Metabolites		
Unique metabolites[b]	1039	625
Cytoplasmic	951	618
Periplasm	418	0
Extracellular	299	143

[a] tabulated on a reaction basis, not counting outer membrane non-specific porin transport

[b] reactions can occur in or between multiple compartments and metabolites can be present in more than one compartment

[c] diffusion reactions do not include facilitated diffusion reactions and are not included in this total if they can also be catalyzed by a gene product at a higher rate

[d] Overall genome coverage based on 4453 total ORFs in *E. coli*[26], 2403 of these ORFs have been experimentally verified

[e] Eight ORFs included in *i*JR904 (1% of the total) have since been removed from the genome annotation[26]

The major areas of expansion of *i*AF1260 over previous *E. coli* network reconstructions fall into five categories: i) increased scope, ii) compartmentalization, iii) increased pathway detail, iv) incorporation of reaction thermodynamics, and v) alignment with EcoCyc.

i) *i*AF1260 is significantly larger in scope than *i*JR904, containing 356 additional ORFs, 1146 additional reactions and 414 additional metabolites (**Table 4.1**)[14]. Furthermore, 289 reactions were removed from *i*JR904, of which, 254 were replaced with similar reactions whereas 35 were totally removed because they were decomposed into more discrete enzymatic steps (27 reactions, see below) or found to be incorrect (8 reactions). Most of the replacements stem from the partition of the model into three distinct subcellular compartments (discussed further below). In order to capture a complete picture of metabolism, certain proteins (e.g., acyl carrier protein) and tRNAs, which function as substrates or products, were also included in the metabolic reconstruction. The ORFs that encode the included proteins were integrated into the GPRs for the reactions in which they participate. It is worthwhile to note that 1161 ORFs (92%) included in *i*AF1260 have been experimentally verified[26]. This number (1161) accounts for 48% of the total 2403 ORFs in *E. coli* that have been experimentally verified.

ii) The reconstruction presented here was separated into three distinct cellular compartments: the cytoplasm, periplasm and extracellular space. Each metabolite in the reaction network was explicitly assigned to one or more of these three compartments (see **Table 4.1**). This representation allowed for the inclusion of transport systems in both the inner and outer membrane and more

accurately represented the metabolic machinery available to *E. coli* in each compartment. Previous *E. coli* reconstructions have not considered the periplasm as a distinct compartment.

iii) *i*AF1260 was generated to minimize the number of grouped, or lumped, reactions in the network reconstruction. Previous versions included many lumped reactions, which simply represent a summation of two or more discrete enzymatically catalyzed reactions, in metabolic processes such as membrane lipid and lipopolysaccharide (LPS) biosynthesis. Although *i*AF1260 includes a smaller total number of lumped reactions than previous reconstructions, some cases remain in which the reaction mechanism(s) has yet to be fully characterized in *E. coli* (e.g., biotin synthase[28]).

iv) The standard Gibbs free energy change of formation, $\Delta_f G'^{\circ}$, and reaction, $\Delta_r G'^{\circ}$, were estimated for most metabolites and reactions in *i*AF1260; 872 (84%) and 1996 (96%), respectively. All $\Delta_f G'^{\circ}$ and $\Delta_r G'^{\circ}$ values were estimated using a new implementation of the group contribution method (MD Jankowski and V Hatzimanikatis, personal communication). A 1M reference state for the metabolite concentrations on which $\Delta_r G'^{\circ}$ is based does not accurately reflect the metabolite concentrations found in the cell (approximately 1mM). Thus, we computationally adjusted all estimated $\Delta_r G'^{\circ}$ to the free energy change of reaction at 1mM concentrations for all species, $\Delta_r G'^{m}$. The distribution of $\Delta_r G'^{m}$ values for reactions in *i*AF1260 indicates that 84% of estimated $\Delta_r G'^{m}$ values are less than or equal to zero in the predicted direction of flux (see below), meaning most reactions are thermodynamically feasible at 1 mM metabolite concentrations (**Figure 4.2A**). Because intracellular metabolite concentrations

can differ significantly from 1mM (typically, 0.00001 - 0.02 M[29]), the actual free energy change of a reaction, $\Delta_r G'$, can differ significantly from $\Delta_r G'^m$. This deviation of $\Delta_r G'$ from $\Delta_r G'^m$ due to metabolite concentrations is shown in **Figure 4.2B** (blue error bars). Uncertainties in the estimated $\Delta_r G'^\circ$ that arise from the group contribution method were also included in the calculation of the $\Delta_r G'$ ranges (purple error bars, **Figure 4.2B**). Thermodynamic estimates were further utilized in the reconstruction process (see below).

v) The content of *i*AF1260 and the EcoCyc[24], release 10.6, and MetaCyc[30] databases were compared to obtain a more accurate and comprehensive reconstruction. EcoCyc and MetaCyc possess a separate curation history from the database from which *i*AF1260 was built and are each extensively curated. A detailed comparison of these resources has resulted in a more thorough analysis and inclusion of metabolic content in *i*AF1260, and in EcoCyc and MetaCyc. A mapping between the reactions and compounds of EcoCyc, MetaCyc and *i*AF1260 was generated in the course of this process. Overall, 945 metabolites in *i*AF1260 (91%) were computationally and manually mapped to EcoCyc and Metacyc compounds (see Supplementary Data[27]). Similarly, 1308 reactions in *i*AF1260 (63%) were computationally mapped to reactions from EcoCyc and MetaCyc using the compound mappings. The results of these mappings are provided in Supplementary Data[27]. A key difference identified from the comparison lies in the use of generic reactions in which enzymes exhibit broad substrate specificity. In EcoCyc, many reaction equations were obtained from the IUBMB [(NC-IUBMB), 2006 #3784]. Accordingly, EcoCyc defined compound classes to represent groups of related substrates, and those compound classes were used as reaction substrates to

represent the fact that a given enzyme could act on several different substrates (i.e., compounds), without necessarily enumerating all these compounds explicitly. Since *i*AF1260 was converted into a computational model, all compounds in its reactions need to be explicitly instantiated. Accordingly, no compound classes or generic reactions were included *i*AF1260.

A breakdown of ORFs, reactions and metabolites included in *i*AF1260 and earlier reconstructions[8-14] are given in **Figure 4.1** and Supplementary Data[27]. **Figure 4.1** was generated using the functional categories assigned through the clusters of orthologous groups (COGs) ontology (http://www.ncbi.nlm.nih.gov/COG/) to classify the reactions included in the *E. coli* metabolic reconstruction. **Figure 4.1A** details the number of ORFs from each COG functional class that were included in *i*AF1260, as well as five previous versions of the *E. coli* reconstruction, to indicate the areas in which the network reconstruction has matured with each successive release. The largest increase in coverage compared to *i*JR904[14] is found in inorganic ion transport and metabolism (26% to 56%, respectively, 73 ORFs). Overall, amino acid and nucleotide transport and metabolism have the highest number of ORFs and percent coverage in *i*AF1260 (256 and 89%, respectively). Ion transport and utilization was recognized as an underrepresented area of metabolism in previous reconstructions and was specifically expanded and incorporated into simulations using *i*AF1260. **Figure 4.1B** and **Figure 4.1C** depict the classification of reactions and metabolites in *i*AF1260 tied to each COG functional class. The largest number of reactions and metabolites associated to ORFs in one COG functional class is in amino acid transport and metabolism and cell wall/membrane/envelope biosynthesis, respectively; furthermore lipid transport and metabolism has the highest reaction to

ORF ratio (5.8), followed by secondary metabolites biosynthesis, transport and catabolism (4.6) and cell wall/membrane/envelope biogenesis (3.4). The large reaction to ORF ratio highlights the fact that the proteins in these classes act on a large number of molecules that only differ slightly in structure. Furthermore, the highest number of unique metabolites that participate in reactions from one class was from coenzyme transport and metabolism. This finding points out the specialized nature of the proteins in coenzyme transport and metabolism pathways.

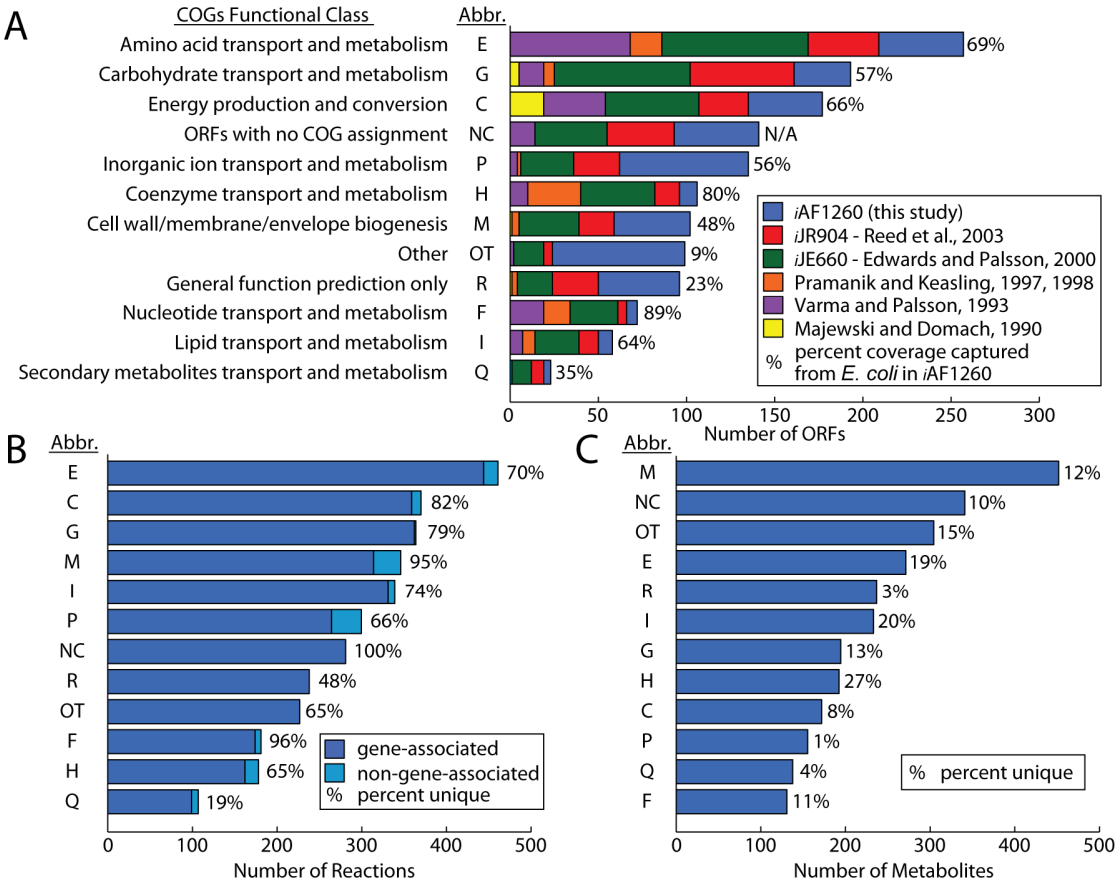


Figure 4.1: Classification of the ORFs, Reactions and Metabolites Included in *i*AF1260. (A) Coverage of characterized ORFs from each of the clusters of orthologous groups (COGs) functional classes included in *i*AF1260 and five previous reconstructions. The total number of ORFs in each functional class from the *E. coli* genome [26] is shown. Some ORFs included in the reconstructions did not have a COG functional class assignment (see Supplemental Data[27]). (B) The number of reactions (both gene-associated and non-gene-associated) that are associated to ORFs from each COG functional class. Since ORFs can belong to multiple classes, the percent unique in each class is listed. Non-gene-associated reactions were assigned to a class manually. (C) The number of metabolites that participate in reactions from each functional class and the percent unique in each class. Other (OT) includes classes J, K, L, O, T, U, V (see Supplementary Data[27]). NC – No COG assignment.

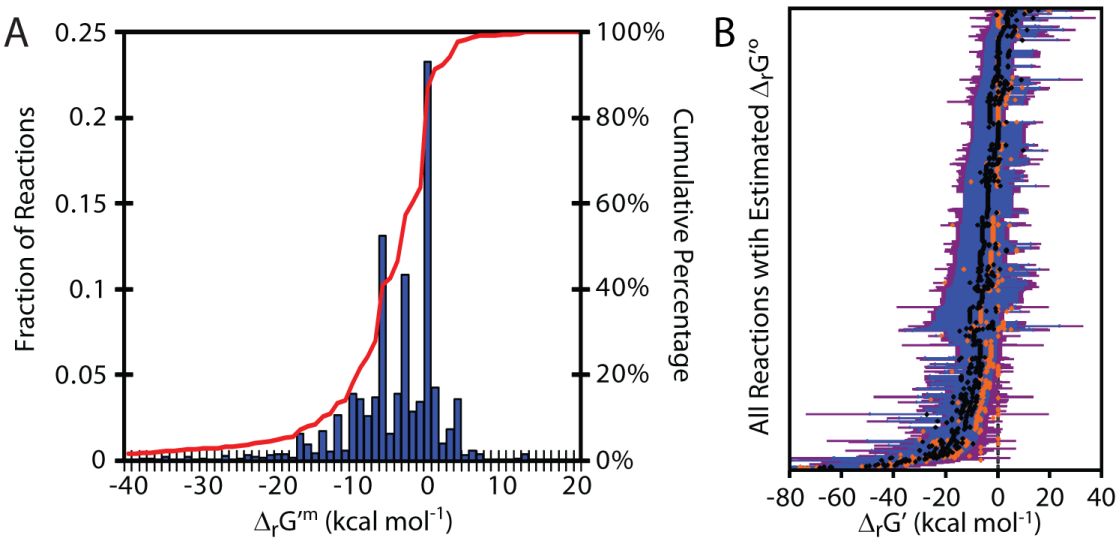


Figure 4.2: Thermodynamic Properties of the Reactions in iAF1260. (A) The distribution of estimated $\Delta_rG'^m$ values for the reactions in *i*AF1260. $\Delta_rG'^m$ could be estimated for 1996 reactions (96%) in the reconstruction. 64% of the represented reactions have a negative $\Delta_rG'^m$, and 20% of the reactions have a $\Delta_rG'^m$ of approximately zero. This distribution of $\Delta_rG'^m$ values indicates that most reactions in the model are thermodynamically favorable at mM concentration conditions. (B) The range of possible Δ_rG' values for the reactions in *i*AF1260. Δ_rG' differs from $\Delta_rG'^\circ$ (orange diamonds) and $\Delta_rG'^m$ (black diamonds) due to variations in metabolite concentrations from the 1M and 1mM reference states, respectively. Metabolite concentrations typically range from 0.02M to 0.00001M, resulting in a wide range of values for Δ_rG' (blue error bars). Taking uncertainty into account, the range of possible values for Δ_rG' can be extended (purple error bars). The Δ_rG' ranges were used to assess the feasibility and reversibility of the reactions in *i*AF1260; reactions for which a positive Δ_rG' is not possible are thermodynamically irreversible.

4.3.2 Conversion to a computational model

Network reconstructions of the type presented herein effectively represent 2-D genome annotations[31] defining the metabolic network that is specific to a particular organism. That is, reconstructions describe both the set of components in a network and the respective interactions between them; two layers of information. It is easily accessible and transferable once developed and can be queried for content, such as genes, proteins, reactions and metabolites. These network reconstructions can be

further converted through a defined series of steps into a computational model that can be used for phenotypic simulations.

The following steps are necessary to convert a network reconstruction to a predictive computational model:

i) *Explicit assignment of the metabolites participating in a reaction*. Some enzymes can act on a number of different metabolites. For modeling purposes, each of these potential metabolites needs to be explicitly defined as participating in a distinct reaction in order to outline a complete picture of metabolism. This step was incorporated in the reconstruction process of *i*AF1260, but is necessary for computational use of a reconstruction based on non-specific metabolites.

ii) *Definition of a system boundary.* Here, the system boundary was defined around the entire reaction network and an exchange reaction (i.e., a reaction that allows a metabolite to enter and exit the system) was made for each of the metabolites in the extracellular space compartment immediately surrounding the cell. Constraints were assigned to each of these exchange reactions during the modeling simulations to restrict the inputs and outputs of the system, depending on the simulated growth environment.

iii) *Conversion of the defined system into a mathematical format that forms the basis for a computational model.* After detailing all GPRs and defining a system boundary, the reconstruction was represented in mathematical terms. The system was represented in the form of a stoichiometric matrix (see Methods) and utilized in the available software platforms SimPheny, and LINDO or TOMLAB in conjunction with MATLAB[32]. The dimension of the

stoichiometric matrix for *i*AF1260 was 1668 × 2381 (# of metabolite × # of reaction species).

iv) *Curation: Filling gaps in the reconstruction.* In order to produce essential biomass components (amino acids, nucleotides, etc.) from minimal media components, there needed to be continuous pathways from media substrates to the required metabolites for biosynthesis. In some cases, the biosynthetic pathways to produce these metabolites were incomplete. A good example is in the biosynthetic pathway for the amino acid L-proline. After reconstruction of the enzymatically catalyzed reactions in the pathway, there was no continuous pathway for the *de novo* generation of L-proline. As a result, the spontaneous reaction L-glutamate 5-semialdehyde dehydratase was needed to complete the pathway and was added to the model with no gene-association[33]. In addition to spontaneous reactions, there were also essential reactions for which the catalytic enzyme has yet to be identified (see Supplementary Data[27] and http://ecocyc.org/enzymes.shtml). Flux-balance analysis (FBA) in conjunction with a biomass objective function (BOF), see below, was used to aid in filling the gaps in *i*AF1260 and results from this analysis are given in the see Supplementary Data[27] section.

v) Determining strain specific parameters. In order to examine the networks ability to fulfill the biomass requirements needed for cellular growth, we generated a set of biomass objective functions (BOFs). The BOFs were linear combinations of experimentally measured metabolites (along with quantities) commonly present in cellular biomass (see **Table 4.2** and see Supplementary Data[27]) and the included metabolites and amounts of each were further judged

for inclusion in this equation through interpretation of gene essentiality data (see Methods). The process of determining maintenance requirements is outlined in detail below.

After the conversion of the reconstructed network into a computational model, a constraint-based approach was used in the context of generating essential biomass components to predict cellular phenotypes under different genetic and environmental conditions.

4.3.3 Application of *i*AF1260 to predict cellular phenotypes

A computational model can be used predict and quantify the active pathways and probable system outputs during growth given a set of inputs that represent growth medium conditions. Analyzing metabolic models in the context of generating maximal amounts of biomass precursors (i.e., simulated optimal growth) from available media substrates using FBA can generate results that are consistent with experimental data[22,34-36]. Thus, we used *i*AF1260 to predict the physiological state of *E. coli* in selected growth conditions using this constraint-based approach (see Methods). It is worthwhile to note that the constraint-based computations performed in this section can be readily reproduced utilizing the *i*AF1260 SBML files see Supplementary Data[27]) and available implemented algorithmic methods[32].

Although *i*AF1260 contains a comprehensive picture of *E. coli* metabolism, there are also other events that need to be accounted for to computationally predict growth capabilities. Three specific issues arose in computational simulations using *i*AF1260. They were:

i) *Transcriptional regulatory events.* A transcriptional regulatory network can be used to determine which ORFs are being transcribed under a given condition[22,37], thus reducing the number of available active pathways under a given growth condition. The events can also limit the rate at which certain enzymes are transcribed, therefore, they are important to apply to a given simulation[22].

ii) *Maintenance costs.* Additional energetic requirements exist for growth beyond what is needed to generate the macromolecular content of the cell (beyond the metabolic costs, which are accounted for directly in the reaction network)[38,39]. These energetic maintenance requirements are for growth associated maintenance (GAM, e.g., protein polymerization costs) and non-growth associated maintenance (NGAM, e.g., membrane leakage) and can be estimated through ATP utilization costs (see Methods).

iii) *Reaction kinetic effects.* Kinetic issues affect metabolism. A potential result from kinetic limitations is that the cell does not always use the most efficient pathways during growth[40,41]. Currently, reaction kinetics are infeasible to incorporate on the genome-scale primarily because of the large number of unknown *in vivo* kinetic parameters and concentrations. However, we know that kinetic effects can influence the utilization of certain pathways, such as the electron transport system (ETS) in *E. coli*[40].

Table 4.2: The biomass composition of the average wild type *E. coli* cell.

Typical 'wild-type' composition

Protein (55.0%)			*Lipid (9.1%)*			
L-alanine	L-arginine	L-asparagine	*structure*			
L-aspartate	L-cysteine	L-glutamine	phosphatidylethanolamine	phosphatidylglycerol[a]	cardiolipin	
L-glutamate	glycine	L-histidine	*acyl chain length: number of unsaturated bonds*			
L-isoleucine	L-leucine	L-lysine	16:0	16:1	18:1	
L-methionine	L-phenylalanine	L-proline				
L-serine	L-threonine	L-tryptophan	*LPS (3.4%)*			
L-tyrosine	L-valine		inner/outer core KDO_2 lipid A			
RNA (20.5%)			*Cofactors, Prosthetic Groups and Other (<2.9%)*			
ATP	CTP	GTP	S-adenosylmethionine	FAD	coenzyme A	NAD(P)
UTP			thiamine diphosphate	riboflavin	undecaprenyl pyrophosphate	
			pyridoxal 5′-phosphate[b]	folates	quinones	hemes
DNA (3.1%)			chorismate	enterobactin	glutathione	putrescine
dATP	dCTP	dGTP	spermidine	vitamin B_{12}		
dTTP						
			Murein (2.5%)			
Inorganic ions (1.0%)			*structure*			
ammonium	calcium	chlorine	murein disaccharide			
cobalt	copper	iron	*peptide chain length*			
magnesium	manganese	molybdate	pentapeptide			
phosphorous	potassium	sulfate		tetrapeptide	tripeptide	
zinc			*Glycogen (2.5%)*			
			glycogen			

'Core' biomass composition substitutes

inner/outer core KDO_2 lipid A: substituted with KDO_2 lipid (IV) A
quinones: substituted with 2-octaprenyl-6-hydroxyphenol
hemes: protoheme; siroheme included
folates: tetrahydrofolate; 10-formyltetrahydrofolate; 5,10-methylenetetrahydrofolate included

The average *E. coli* wild type macromolecules (and the weight percentage for each) are listed along with their corresponding network metabolites or metabolic precursors. The non-essential wild type metabolites were determined using gene essentiality data[2,43] and are shown in red. Metabolites listed in blue were determined to have a reduced 'core' structure different from the wild type metabolite(s) and these are listed in the 'core' biomass substitutions.
[a] was determined nonessential from[44].
[b] determined to be essential under minimal media conditions and was not essential under the rich media condition examined.

Figure 4.3 demonstrates how we addressed the three modeling issues outlined above when using FBA with *i*AF1260 to predict the physiological state of *E. coli* growing aerobically on glucose. Initially, all of the pathways characterized in *i*AF1260 were represented in a computational framework (**Figure 4.3A**). We then constrained the reactions that correspond to ORFs that are not transcribed under aerobic glucose conditions to zero allowable flux in the network using the Boolean gene regulatory rules based on 104 transcription factors established by Covert and colleagues[22], **Figure 4.3B**, effectively eliminating 152 reactions (see Supplementary Data[27]). Using the reduced network, we then constrained the maximum allowable P/O

ratio of the ETS by using observations and predictions from previous studies. *E. coli* possesses two NADH dehydrogenase components (NDH-1 (*nuo*) and NDH-2 (*ndh*)) and two terminal oxidases (bo-type (*cyo*) and bd-type (*cyd*) oxidase) in the system[41,45]. Different combinations of these respiratory components can result in an overall translocation that can range from 2 H^+/ $2e^-$ to 7 H^+/ $2e^-$ in *i*AF1260. The specific constraint we placed on the system was to split the flux ratio between the two NADH dehydrogenases 1:1 (NDH-1:NDH-2) allowing for a P/O ratio between 0.5 – 1.375 (**Figure 4.3C**)[45,46].

Using chemostat data for *E. coli* growing aerobically on glucose (see Supplementary Data[27]), we estimated the GAM and NGAM costs (**Figure 4.3D**). We found a NGAM value of 8.39 mmol ATP gDW^{-1} hr^{-1} and a GAM value of 59.81 mmol ATP gDW^{-1} best fit the experimental data. Using these values and no restriction on pathway choice for the ETS, we calculated the line of optimal growth using FBA for aerobic growth on succinate. This line was plotted against the measured values for wild type batch growth determined by Edwards and colleagues[34]. The calculated line of optimality corresponds to the conditions (substrate uptake and product generation rates) which can maximize the biomass yield. The results show that most of the measured values lie very near the line of optimality in the experimental range examined (see **Figure 4.3E**).

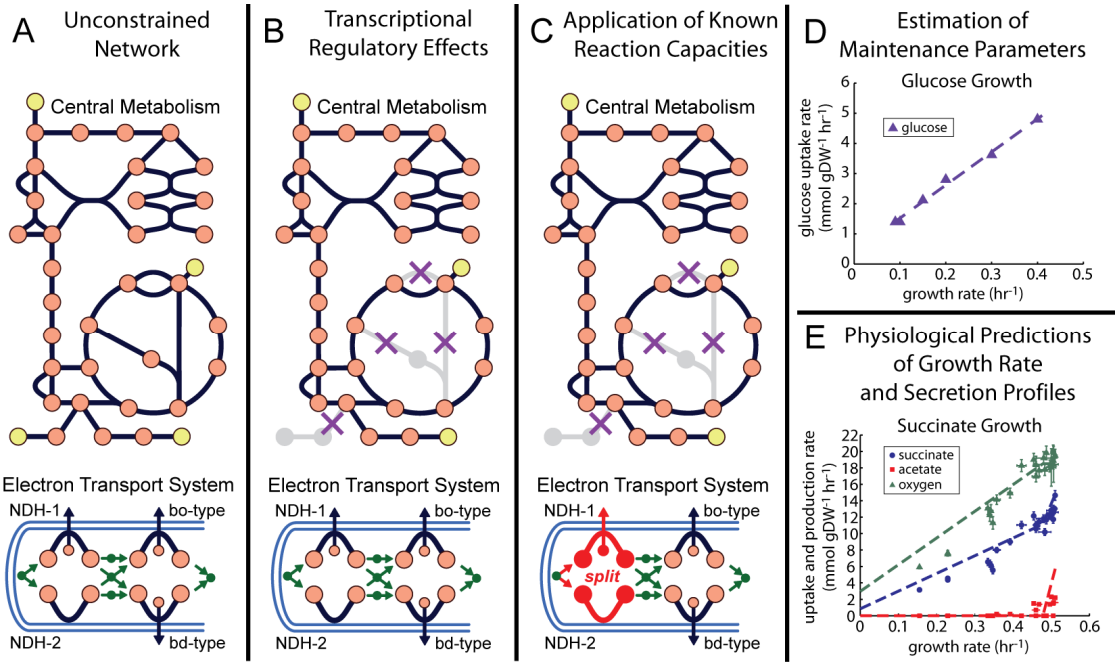


Figure 4.3: Utilizing *i*AF1260 as a Predictive Model. (A) A drawing of central metabolism and the electron transport system included in *i*AF1260. Originally, the whole network is unconstrained. (B) Application of transcriptional regulatory effects restricts the total number of pathways, or routes, flux can pass through in the network (C) Further application of known reaction capacities can result in more accurate predictions. For example, the flux through the NADH dehydrogenase enzymes is split in a 1:1 ratio during a simulation to produce an optimal P/O ratio of approximately 1.4 [41,46] (D) The non-metabolic activity of the cell can be accounted for through maintenance parameters and these were approximated using experimental data under known media conditions. Chemostat data (see Methods) was used (triangles) and the dotted line shows the modeling predictions with the appropriate maintenance parameters. (E) After the parameters are approximated, the model can then be used to predict the growth rate (circles), product formation (acetate, squares) and additional uptake rates (oxygen, triangles) under different environmental conditions (for succinate growth in this case).

To further examine the agreement between modeling simulations and experimental data, computationally predicted flux values, product formation rates and growth rates were compared to experimentally determined values derived from ^{13}C-labeling experiments[47]. Using measured glucose and oxygen uptake rates as modeling constraints[47], FBA was used to examine the predicted network flux distribution when optimizing for flux through the BOF_{CORE} reaction (see Methods). The

produced flux distribution accurately predicted both the growth and acetate secretion rate using the measured average glucose and oxygen uptake rates from triplicate ^{13}C-labeled experiments. Additionally, the CO_2 production rate was accurately predicted when considering the standard deviation on the reported uptake values. Both the experimental and computational results suggest that no other carbon containing products were generated in measurable amounts. Examining the flux distribution in central metabolism, there was complete agreement in the direction of flux through the glycolytic, pentose phosphate, TCA and pyruvate metabolism pathways between modeling and experimental results. For the Entner-Doudoroff pathway, the experimentally determined flux was equal to or less than 4% of the total glucose flux entering the system, whereas no flux was predicted for this pathway for an optimal growth solution using *i*AF1260. Looking at the quantitative values for 22 individual central metabolism fluxes[47], the experimentally reported and computationally predicted values were in good agreement (mean of the difference = 8 ± 1.4 % (SE), R^2 = .96; where fluxes were normalized to glucose uptake rates being 100%). The most notable discrepancy when comparing the computational and experimentally reported values was in the pentose phosphate pathway, where 26% of the total glucose flux entering the system was calculated to be shuttled through this pathway when analyzing the ^{13}C-labeling data, and the model predicted a value of 46% during an optimal growth solution. All of the flux values predicted for other central metabolism pathways agreed well with the experimental flux data.

4.3.4 Thermodynamic Consistency Analysis

Previous metabolic network reconstructions have focused on the chemistry of the reactions that take place and their genetic basis. The physicochemical

characteristics of the reactions, namely thermodynamic and kinetic properties, have not been incorporated. The kinetics are hard to obtain and change with organism adaptation and evolution[48]. Conversely, the thermodynamic properties represent physicochemical limitations that can be estimated and taken into account[49-51]. In forming *i*AF1260, we incorporated thermodynamic information (see **Figure 4.2**) to provide another means of assessing reaction reversibility beyond what is stated in the primary literature and assignments made using general heuristic rules (see Methods).

Through a process termed thermodynamic consistency analysis, the thermodynamic estimates were utilized to evaluate the reversibility and directionality assigned to the reactions in the reconstruction based on the primary literature and heuristic rules. First, flux variability analysis (FVA)[52] was utilized in combination with the calculated $\Delta_r G'$ ranges to identify reactions that operated in a thermodynamically infeasible direction during near optimal growth on at least one carbon source (see Methods). The co-substrates and cofactors involved in these inconsistent reactions were adjusted (either the participants or stoichiometries) with guidance from the literature so that the reactions in the final version of the reconstruction were not thermodynamically infeasible in any of the directions in which they must operate for near optimal growth on 174 carbon sources.

One example of an initially thermodynamically inconsistent reaction that was altered is the hydrogenase 3 catalyzed reaction, formate-hydrogen lyase. This reaction initially powered the transport of 1.3 protons across the cell membrane while oxidizing formate to hydrogen and carbon dioxide[53]. Thermodynamic analysis of this reaction indicated that the intracellular portion of this reaction is already unfavorable with a $\Delta_r G'^{\circ}$ and $\Delta_r G'^{m}$ of 2.1 ± 1.7 kcal/mol, in agreement with reported values[54,55].

Given the concentration gradients achievable *in vivo*, it was found to be highly improbable that this already unfavorable reaction could power the transport of 1.3 protons across the cell membrane. As a result, the transmembrane transport portion of this reaction was removed.

Some of the other thermodynamically inconsistent reactions identified prompted the adjustment of the reversibility for network reactions and also, expansion of the reconstruction content. For example, we identified thermodynamic infeasibilities in the reactions involved with the transport of the inorganic ions (i.e., Fe^{2+}, Cu^{2+}, etc.). Initially, the only reactions in the model allowing for the transport of these ions across the cytoplasmic membrane were reversible diffusion reactions of the form:

$$\mathrm{Ion}^{+2}_{\mathrm{periplasm}} \leftrightarrow \mathrm{Ion}^{+2}_{\mathrm{cytosol}}$$

According to FVA, during growth on some carbon sources, these metal ions could be exported from the cell. However, based on our thermodynamic calculations, we determined that these reactions are only thermodynamically feasible in the direction of import where the transmembrane electrochemical potential contributes energy to the transport process[25]. The literature confirms that while the import of these ions proceeds via diffusion through a regulated ion channel, the export of these ions requires a separate mechanism that utilizes ATP hydrolysis or proton antiport as a source of energy to drive the reaction[56,57]. These alternative export reactions were consequently added to the reconstruction.

The FVA performed as part of the thermodynamic consistency analysis further allowed for the reactions in the reconstruction to be functionally classified as essential (requiring a nonzero flux), substitutable (capable of carrying zero or nonzero flux) or blocked (zero flux) during growth on each of the 174 carbon source studied.

Interestingly, a large number of the reactions in the reconstruction behaved uniformly regardless of the carbon source being utilized (**Table 4.3**). Many reversible reactions only operated in a single direction despite being reversible, while many other reactions were predicted not to operate in any of the FVA studies performed. These reactions are potentially involved in dead-ends in the reconstruction or conversely, were limited because of the objective function used to examine the network.

Table 4.3: Classification of *i*AF1260 reactions based on a FVA for 174 different carbon sources[a].

	Number of reactions (All thermodynamically feasible in all directions of flux)
Essential for all 174 carbon sources, with flux always in the same direction	183
Essential for all 174 carbon sources, with flux in different directions depending on the carbon source	3
Substitutable for all 174 carbon sources, with flux always in the same direction	863
Substitutable for all 174 carbon sources, with flux in different directions depending on the carbon source	41
Essential, substitutable or blocked depending on the carbon source with flux always in the same direction whenever flux is present	502
Essential, substitutable or blocked depending on the carbon source, with flux in different directions also depending on the carbon source	182
Irreversible reactions blocked for all 174 carbon sources	227
Reversible reactions blocked for all 174 carbon sources	76

[a] see text for a definition of essential, substitutable and blocked. Exchange and demand reactions were not considered.

Once the reactions that operated in thermodynamically infeasible directions according to the FVA were identified and adjusted to remove all thermodynamic inconsistencies, we examined the $\Delta_r G'$ ranges calculated for all of the reactions defined as reversible based on the primary literature or heuristic rules. Through comparison to predicted $\Delta_r G'$ values, we identified many reactions that were originally specified as reversible and were actually thermodynamically irreversible (i.e., reactions being incapable of achieving both negative and positive values of $\Delta_r G'$ under physiological conditions). We corrected these reversible reactions to be consistent with our thermodynamic estimates (i.e., made them irreversible). In total,

after checking for consistency, 553 reactions in *i*AF1260 were assigned as reversible and 1524 reactions were assigned as irreversible.

Looking at the reversibility of the reactions predicted using thermodynamic estimates alone, 1673 (84%) of the reactions for which $\Delta_r G'^{\circ}$ could be estimated were predicted to be reversible whereas 323 (16%) were predicted to be irreversible. This finding indicated that reaction reversibility specified in the reconstruction was more restrictive than what is called for by thermodynamic analysis alone. The primary reason for this more restrictive property is that the reversibility set forth for the reactions in the reconstruction is often based on the physiological behavior of the reactions in the cell, not using the relatively broad concentration range achievable for metabolites along with the uncertainty inherent in the utilized method. Comparing these values to another approached, the method used in this work recognized 323 (16%) reactions as being irreversible. Whereas Kümmel and colleagues[58], recognized 130 (14%) reactions as being irreversible in the *i*JR904 network[14] utilizing a similar thermodynamic based assignment and an additional assignment through heuristic rules.

4.3.5 Sensitivity Analysis

To determine the sensitivity of the computational results (e.g., optimal product formation rate) generated using FBA with *i*AF1260, we varied independently the: i) constraints imposed by transcriptional regulation on the network, ii) metabolites included in the BOF, iii) macromolecular content of the cell, iv) effective P/O ratio in aerobic growth, and v) the maintenance costs associated with growth (i.e., NGAM and GAM). This analysis was performed using simulations of optimal growth under aerobic glucose-limited conditions.

In order to examine the overall regulatory effects on the computational results, we performed simulations both with the constraints outlined earlier[22] and with no transcriptional regulatory constraints (see Methods and **Figure 4.3**). The resulting optimal growth rate (GR) and oxygen uptake rate (OUR) for a given glucose uptake rate (GUR) (i.e., the line of optimality) predicted using FBA was found to be insensitive to the regulatory constraints placed on the system under these conditions (**Figure 4.4**). However, the regulated network was less flexible in terms of the number of reactions that could possess a non-zero flux for an optimally predicted growth rate (33 less reactions). This result is not altogether surprising given that glucose is a preferred substrate for growth on one carbon source and thus, regulation had likely evolved to limit the uptake of additional carbon sources[36]. Similarly, comparing the use of the BOF_{WT} and the BOF_{CORE} for predicted optimal growth using FBA, the line of optimality produced was essentially identical in both cases (**Figure 4.4**). However, there were 95 more reactions that required a non-zero flux value for an optimal solution in the network using the BOF_{WT} in FBA simulations. This result is expected since the BOF_{WT} is comprised of more metabolites requiring more active fluxes for their synthesis.

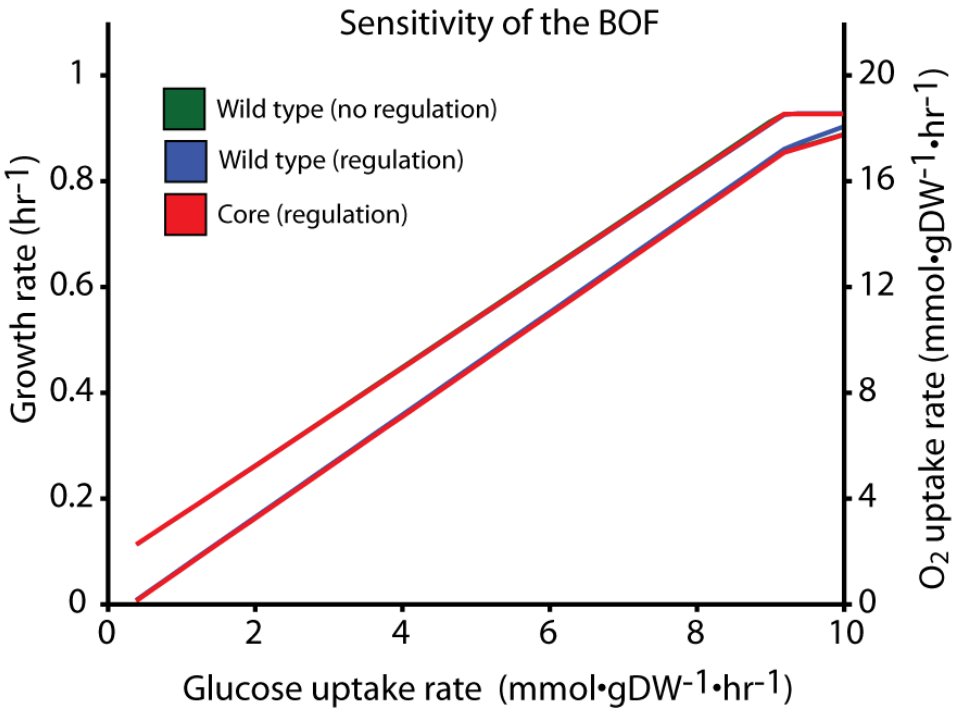


Figure 4.4: Sensitivity analysis varying the biomass objective function. The relationship between the glucose uptake rate (mmol gDW^{-1} hr^{-1}) (bottom axes, the dependant variable) and the resulting 1) growth rate (hr^{-1}) (left axes) and 2) oxygen uptake rate (mmol gDW^{-1} hr^{-1}) (right axes) produced during the sensitivity analysis using *i*AF1260. Using FBA and *i*AF1260, optimal growth was simulated under glucose aerobic conditions while varying which biomass objective function (BOF) was used along with the number of reactions available to the network due to transcriptional regulation. Two different BOFs were used, a core biomass objective function and wild-type biomass objective function, and regulation was imposed by not allowing any flux through reactions unavailable to the network due to transcriptional regulation[22]. The results show that the predicted optimal growth rate and O2 uptake rate are insensitive to the BOF used or level of transcriptional regulation imposed under these conditions.

To examine the effect of changing the macromolecular composition represented in the BOF, we varied the weight percentage of the three largest macromolecules in the cell (**Table 4.2** and **Figure 5A-5C**) in FBA simulations. We thus generated new BOFs varying the protein content from 50 – 80 wt%, the RNA content from 10 – 25 wt% and the lipid content from 7-15 wt% of the cell based on recorded experimental values[11]. While the macromolecular composition of the BOF_{CORE} had some effect on the overall optimal GR and OUR, the most extreme

variance was, at most, 5% and 8% at the median GUR in the range examined, respectively. Previously, Pramanik and Keasling[11] evaluated a BOF that was growth rate dependant to determined that the building blocks that make up the macromolecular content of the cell (e.g., the amino acids that make up the protein content) are essentially constant when *E. coli* is grown under different growth conditions and any small changes in these compositions do not significantly affect calculated reaction flux values[12]. Therefore, this variable was not examined.

The P/O ratio of the ETS in *E. coli* was varied to determine its effect on optimal solutions produced using *i*AF1260 and FBA. The maximum value that the P/O ratio can achieve under aerobic conditions is 1.75 based on the stoichiometry of the ETS enzymes in *i*AF1260. Since there is some debate on the possible overall stoichiometries of the ETS in *E. coli* (see above), we further increased the potential maximum to 2.7 in our analysis and tested the effect of a P/O ratio ranging from 1.0 – 2.7. Specifically, a P/O ratio of 2.7 could be achieved if the most energy efficient pathway was used exclusively and the ETS possessed an ATP synthase with a stoichiometry of 3 H^+/ATP and a NDH-1 with a stoichiometry of 4 $H^+/2e^-$. A P/O ratio of 1.0 is an estimated low-end value for aerobic growth. The analysis indicated that the modeling results are most sensitive to the P/O ratio than any other variable examined in this analysis. Optimal GR and OUR predictions varied, at most, 37% and 71% at the median GUR in the range examined, respectively.

Finally, we analyzed the effects of maintenance energy on optimal growth predictions. We varied the values of the NGAM and GAM ± 50% of the most consistent values of 8.39 mmol ATP gDW^{-1} hr^{-1} and 59.81 mmol ATP gDW^{-1}, respectively. The region that the line of optimality could posses for the varying

maintenance energies was plotted in **Figure 5E** and **Figure 5F**. The NGAM can affect the optimal GR and OUR predictions, at most, 8% and 15% and the GAM 16% and 31% at the median GUR in the range examined, respectively. Thus, these are also important variables to consider in FBA simulations of optimal growth under these conditions. Looking at the specific effect that each variable inflicts on the system; the NGAM shifts the intercept of the line of optimality with the GUR and OUR axes, whereas the GAM values change the slope of the line of optimality. The impact of the sensitivity analysis is addressed in the discussion.

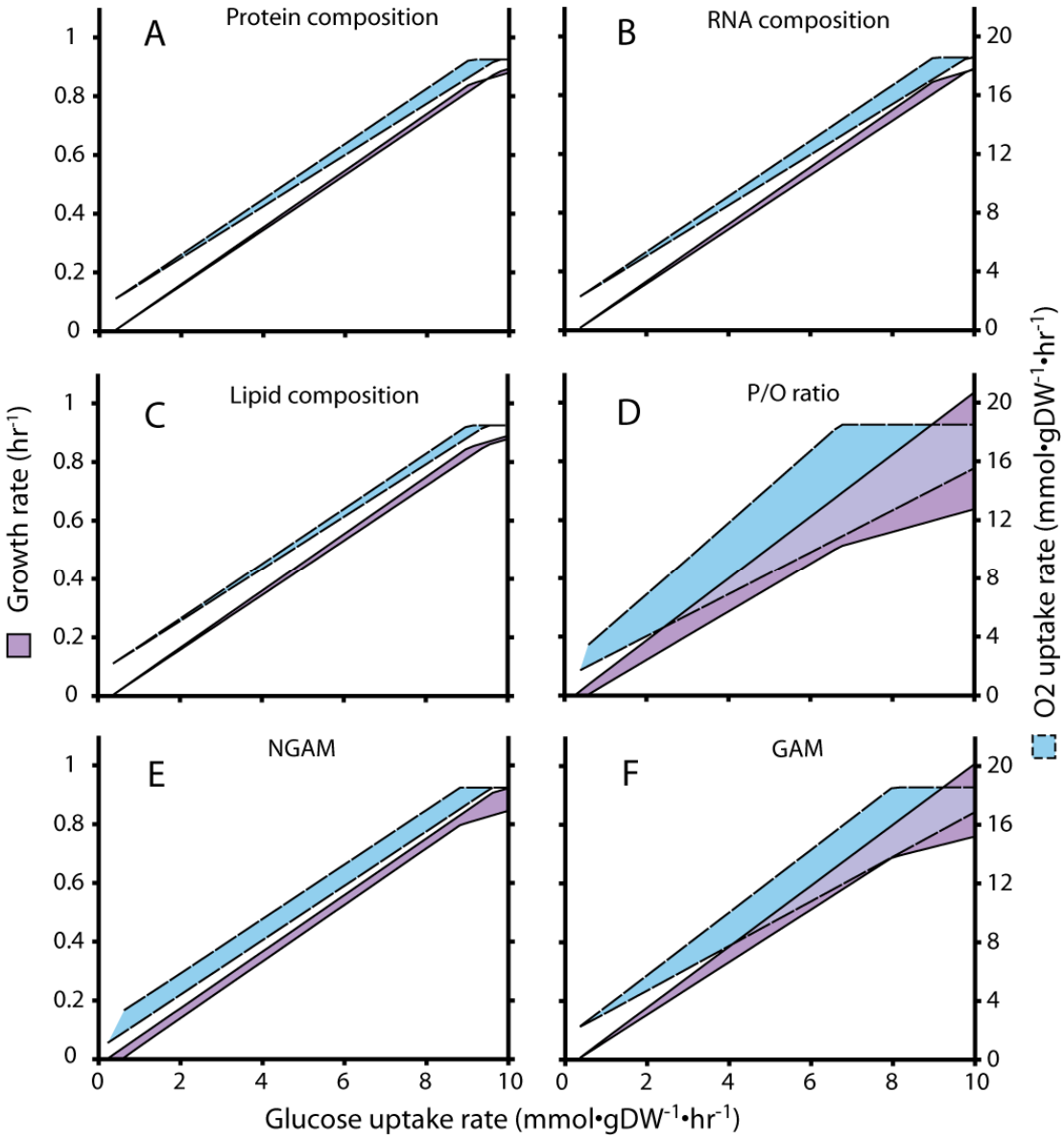


Figure 4.5: Sensitivity Analysis on the Modeling Parameters used in Analyzing *i*AF1260. The relationship between the glucose uptake rate (mmol gDW^{-1} hr^{-1}) (bottom axes, the dependant variable) and the resulting 1) growth rate (hr^{-1}) (left axes) and 2) oxygen uptake rate (mmol gDW^{-1} hr^{-1}) (right axes) produced during the sensitivity analysis using *i*AF1260. Using FBA and *i*AF1260, optimal growth was simulated under glucose aerobic conditions while varying (A) the dry weight percentage of protein (50-80%), (B) RNA (10-25%) and (C) lipid (7-15%) in the BOF_{CORE} using physiologically measured values [11]. Also analyzed was (D) potential P/O ratios (1.0 – 2.7) in the network as well as the (E) NGAM (± 50%) and (F) GAM (± 50%) determined for these conditions.

4.3.6 Context for Content

As high-throughput data becomes available for a number of organisms[4] there is a need for an underlying platform to analyze this data by placing it in a biological context. Genome-scale metabolic reconstructions, such as *i*AF1260, offer such a basis since they are biochemically and genetically structured databases. As a result, they can be utilized to interpret high-throughput data in analyses looking at specific reactions, pathways or even genome-wide trends.

4.3.7 Context for Content: Analysis of alternate growth conditions

Similar to our previously described application of *i*AF1260 to predict the physiological state of *E. coli* growing under an aerobic glucose or succinate limiting condition, we also performed a broader analysis to determine all of the additional carbon, nitrogen, phosphorus and sulfur sources that could support simulated growth in minimal medium and compared this to findings using *i*JR904[14]. Overall, there were 174 carbon, 78 nitrogen, 49 phosphorous and 11 sulfur sources that were predicted to support growth using FBA (see **Table 4.4** and Supplementary Data[27]); an increase over *i*JR904 by 84 carbon, 44 nitrogen, 45 phosphorous and 9 sulfur sources. We compared the computational results to a high-throughput experimental screen using the Biolog platform (http://www.biolog.com). **Table 4.4** details the comparison between the computational and experimental predictions. The overall agreement is approximately 76% using *i*AF1260, compared to 60% for *i*JR904. This result reflects the increased scope of *i*AF1260 to analyze a wider range of growth conditions and helps validate the content of *i*AF1260.

Table 4.4: Growth condition analysis.

	Computational		Experimental	Agreement (*i*AF1260/*i*JR904)			Disagreement (*i*AF1260/*i*JR904)		
Source	Potential substrates	Support growth[a]	Total possible comparisons	E-G C-G	E-NG C-NG	% Total	E-NG C-G	E-G C-NG	% Total
Carbon	262	174/90	87	54/46	11/15	75%/70%	22/18	0/8	25%/30%
Nitrogen	163	78/34	51	28/24	8/12	71%/71%	8/4	7/11	29%/29%
Phosphorous	63	49/4	20	20/3	0/0	100%/15%	0/0	0/17	0%/85%
Sulfur	25	11/2	12	8/2	0/0	67%/17%	0/0	4/10	33%/83%

[a] results using the *i*AF1260 / *i*JR904 computational model
G – growth, NG – no growth, E – experimental, C – computational

Disagreements between the computational and experimental data fall into two main categories and, going forward, will be resolved by different approaches. Cases in which computational growth is predicted and not observed experimentally indicate possible areas where there are either errors in the reconstruction or alternatively, where regulation limits the utilization of pathways needed for growth. This type of false positive for growth increased with *i*AF1260 since the network increased in total reactions available to support growth. In contrast, instances where experimental growth is observed and no growth is predicted computationally points to areas where further biochemical characterization is needed for *E. coli* and defines targeted areas for biological discovery[23]. These false negatives for growth were significantly reduced from *i*JR904 to *i*AF1260 and clearly demonstrate the effect of the expanded content on computational simulations. Additional targets for model-driven expansion can be found in Supplementary Data[27].

4.3.8 Context for Content: Gene essentiality analysis in *i*AF1260

We used the reconstruction as a framework to analyze the conditionally essential ORFs identified for *E. coli* K-12[42,43]. A comparison between the computationally predicted essential ORFs and the experimental data[42,43] is provided

in **Table 4.5** and **Figure 4.6**. Gene essentiality predictions under glucose aerobic conditions using *i*AF1260 show an overall increase in the number of ORFs that can be examined and correctly predicted when compared to *i*JR904 (an increase of 356 and 357 ORFs, respectively). There is also a modest improvement in overall accuracy (92% compared to 88%, see **Table 4.5**, see Supplementary Data[27]). These findings provide confidence for using *i*AF1260 to investigate previously unstudied conditions and examining the specific functionality that an essential (or non-essential) ORF provides for the system under that given condition. The agreement between the experimental and computational results, on the whole, validates the content of the reconstruction and the modeling procedure (assuming a low error rate in the experimental data).

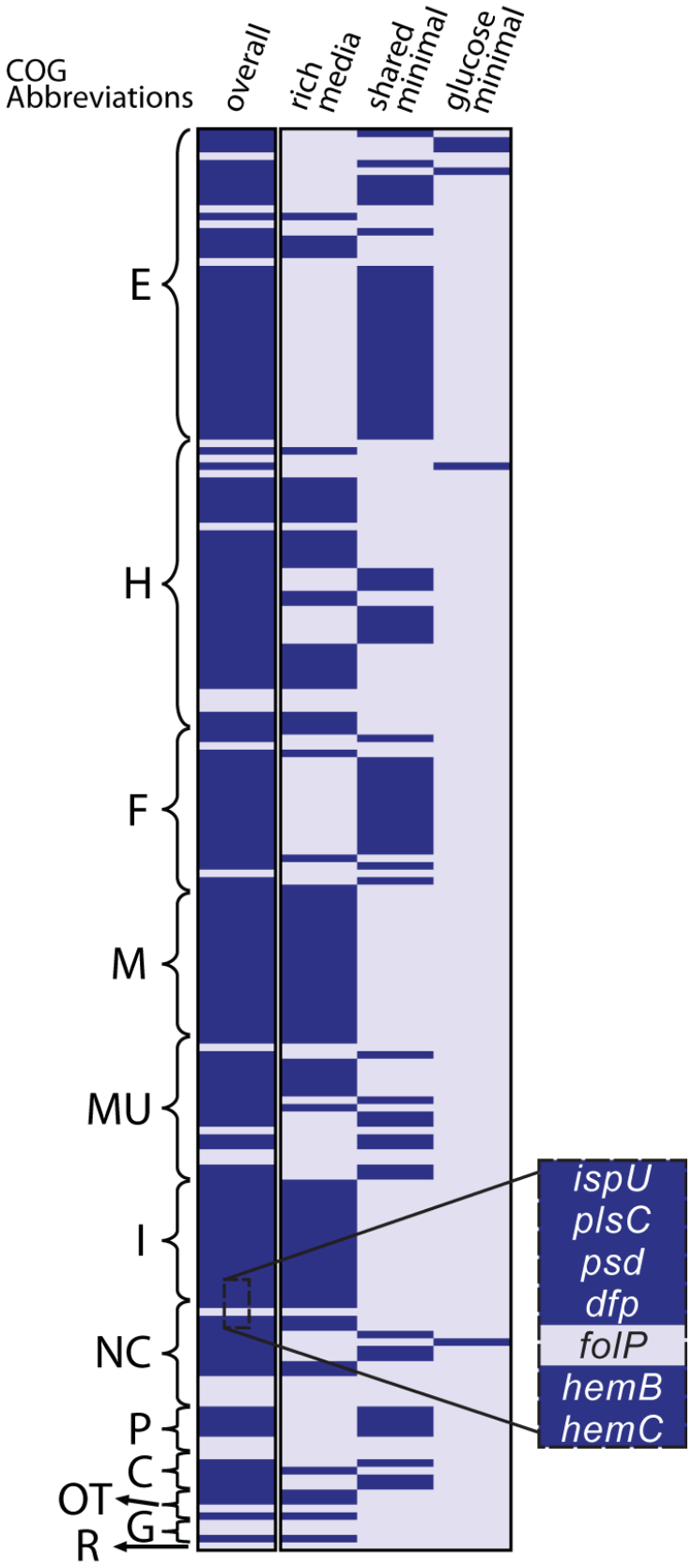
Table 4.5: Computational essentiality predictions.

	Experimental	
	Essential	Non-essential
Computational		
Essential	159 (13%)	29 (2%)
Non-essential	79 (6%)	993 (79%)

Disagreements between the experimental and computational data point to further areas of refinement and expansion of the metabolic and regulatory networks know for *E. coli*, as well as possible errors in the experimental data and model. The disagreements where ORFs were found to be computationally essential, but experimentally non-essential point to specific areas where additional intracellular and transport reactions can be examined to rectify the disagreements (29 cases, see **Table 4.5**). For example, the *ubiC* gene was predicted to be essential for its involvement in the ubiquinone biosynthesis pathway. This finding points to the fact that additional work is needed to characterize the full complement of genes

responsible for the aerobic and anaerobic production of ubiquinone[59]. Additionally, 8 of the 29 cases were predicted to be essential for thiamin biosynthesis, an essential cofactor in *E. coli*. This result suggests a likely error in the experimental data and is supported by Vander Horn and colleagues[60]. ORFs that are found to be experimentally essential but computationally non-essential suggest potential regulatory effects on the system and possible inaccuracies in the metabolic network (79 cases, see **Table 4.5**). Transcriptional regulation limits network pathways under a given condition, therefore computational disagreement could arise if such pathways are computationally utilized. Disagreements in this class also identify a current limitation of the model. The action of 18 tRNA charging reactions are contained in the reconstruction, but are not currently accounted for in the modeling scheme. The resulting computational disagreements will likely be resolved through expansion of the network to include transcription and translation processes in the cell. For a total list of the computational and experimental disagreements, see Supplementary Data[27].

Figure 4.6: ORF essentiality predictions using *i*AF1260. (facing page) This heatmap characterizes the agreement between ORFs predicted to be essential using *i*AF1260 and those experimentally determined from Baba and colleagues [42] and Joyce and colleagues [43]. The enlarged region details how each row corresponds to a computationally predicted essential ORF (188 total). The overall agreement between *i*AF1260 predictions and those found to be experimentally essential (total, column 1) is shown along with a breakdown for ORFs found to be essential under rich media conditions (rich, column 2), under both glucose and glycerol minimal media conditions (shared, column 3) and under just glucose minimal medium conditions (glucose, column 4). ORFs are further grouped by their COG functional class (see **Figure 4.1** for abbreviations; MU - ORF belongs to multiple COG classes). Dark blue indicates the condition under which each ORF was found to be essential. For example, *folP* was predicted to be an essential ORF for the biosynthesis of folate in *i*AF1260 under these conditions, but was not identified as essential by Baba and colleagues [42]. The suggested the possibility of an alternative pathway for this step in *E. coli* that has yet to be characterized.


COG
Abbreviations
overall
rich media
shared minimal
glucose minimal
E
H
F
M
MU
I
NC
P
C
OT
G
R
ispU
plsC
psd
dfp
folP
hemB
hemC

4.4 Discussion

Metabolic reconstruction and subsequent mathematical computation has become a useful tool in the post-genomic era by aiding both biological computation and experimentation. In this work, we present, characterize and utilize the *i*AF1260 metabolic reconstruction of *E. coli* K-12 MG1655. The reconstruction serves as both a BiGG database containing the current knowledge of *E. coli* metabolism, as well as a framework for mathematical analysis. Accordingly, the major contributions from this work are: (1) an expansion in size, scope and detail of the metabolic network of *E. coli*, effectively exhausting the available literature, (2) an enumeration and description of the parameters and methods needed to utilize the reconstruction as a predictive model; examples of simulation results compared with high-throughput experimental data are presented, and (3) the inclusion of thermodynamic information and a novel thermodynamic consistency analysis for chemical transformations accounted for in the reconstruction.

*i*AF1260 represents the largest metabolic reconstruction of any unicellular organism and accounts for 1260 ORFs (28%) in the current *E. coli* genome annotation[26]. Furthermore, 1161 of the included ORFs (92%) have experimentally based functions, conferring a high degree of confidence in the corresponding interactions. Just as gene annotation and sequence databases are used to identify and characterize genes in newly sequenced genomes, *i*AF1260 will similarly serve as a primary reference for future metabolic reconstructions. Because of its curation history and size, future reconstructions, especially those for closely related organisms, will draw directly from this content. This process will further be aided by the synchronization and mapping with the EcoCyc database. The next step in the

expansion of the *E. coli* metabolic network will require further discovery of metabolic functions and computational methods are needed that can facilitate this process[23].

In addition to expanded content, significant advancements in reconstruction techniques and methods used to determine network capabilities were presented. Thermodynamic consistency analysis represents a novel way to flag or highlight highly improbable intracellular and transport reactions for further evaluation. This approach can be added to future metabolic reconstruction and modeling projects. It effectively constitutes a QC/QA test that should improve the utility and scope of modeling predictions. Additionally, the use of a core biomass objective function (BOF_{CORE}) has identified an improved strategy to probe gene essentiality for growth. Previous analyses examining gene essentiality have utilized a BOF which is based on measurements from a specific growth environment and is also constant in the type and relative proportion of metabolites. A common problem that arises when using a BOF based on wild-type measurements is that potential false positives can be generated when conditionally essential metabolites are inappropriately included in the objective function[61,62]. The BOF_{CORE} presented here, with continual refinement guided by experimentation, should increase the accuracy and utility of computational predictions with respect to mutant phenotype predictions.

The approach taken to evaluate reaction reversibility in *i*AF1260 was to prevent the inclusion of reactions that were highly unlikely to be reversible. This approach was carried out by using the thermodynamic consistency analysis and subsequent analysis of reaction thermodynamic estimates. Due to the thermodynamic coupling of reactions operating simultaneously, reactions that are individually thermodynamically reversible under physiological conditions will not necessarily be

reversible when operating in concert with the other reactions in the cell. In line with this, using reaction reversibility determined from the thermodynamic analysis of individual reactions alone with FBA will result in improper model behavior due to the operation of thermodynamically infeasible pathways and cycles. Only if thermodynamic constraints are used in conjunction with the mass balance constraints of FBA to prevent the operation of these thermodynamically infeasible pathways (for example,[51]), can the reaction reversibility determined for individual reactions be used. Therefore, utilization of the thermodynamic information presented to fully assign reversibility and irreversibility in modeling simulations automatically, requires additional implementation of methods which consider thermodynamics on the systems level.

With the increasing use of network reconstruction and the constraint-based modeling approach, a need has emerged to clearly define and demonstrate the steps required to computationally utilize a reconstruction. By outlining these steps and examining the sensitivity of modeling parameters used in computations, we have both explicitly defined the protocol and revealed the impact of modeling parameters on predictions. A computational software package is also available to efficiently implement such metabolic modeling[32]. A sensitivity analysis of key strain-specific parameters using an early version of the reconstructed *E. coli* network[63] found that the P/O ratio significantly affects the growth rate and flux predictions, whereas varying the BOF had relatively little effect. However, our analysis shows a greater dependence on the maintenance parameters calculated for these conditions. This result is primarily due to our testing of a broader maintenance value range (± 50% of the calculated values as opposed to 20% by Varma and Palsson[63]). This larger value was selected because it is approximately the amount of the GAM that is difficult to

quantify (i.e., unknown maintenance that accounts for gradient maintenance, protein turnovers and so forth[38]) and it produced a range that can be justified by examining different *E. coli* growth data (results not shown). Future projects should take into account the impact of the influential parameters (i.e., P/O ratio, growth maintenance) when designing their computational studies.

The culmination of the increased size and expanded coverage of the reconstruction, in combination with the improved reconstruction techniques, has broadened the scope and accuracy of computational predictions. Comparisons of *i*AF1260 simulations with experimental data for gene-essentiality and growth phenotypes showed an overall increase of 4% and 16% over *i*JR904 predictions, respectively. Specifically, *i*AF1260 is markedly improved in analyzing and predicting a wider range of minimal media growth conditions (see **Table 4.4**). It can also better predict and screen the essential genes needed for viability in *E. coli* (see **Table 4.5**). The one area where it appears that the model's ability to match experimental data decreased was where ORFs were found to be experimentally essential, but computationally non-essential. This area can be addressed through further expansion of the reconstruction's scope (e.g., by including the transcriptional and translation machinery in *E. coli* as well as transcriptional regulatory effects) and targeted experimentation (e.g., elucidating the entry step into the *de novo* biosynthesis of biotin).

Future directions for improvement of the metabolic reconstruction of *E. coli* remain. As previously stated, the scope of the reconstruction will continually increase. Dead-ends and lumped reaction in the reconstruction point to specific areas of *E. coli* metabolism that can be further characterized in this expansion effort. A computational

approach to resolve these dead-ends that utilizes constraint-based methods can be used in this effort. Additionally, an area for further compartmentalization of metabolites in the reconstruction is for metabolites located in the lipid bilayers. For example, a lipid on the inner leaflet of the outer membrane is different than one on the outer leaflet of the inner membrane, but currently in the reconstruction, they are both located in the periplasm. Further advancements in modeling will also be achieved through the acquisition of additional experimental gene essentiality studies under different minimal media conditions to better define the core metabolites needed for viability and improve overall computational accuracy. Advancements are also likely to arise from additional incorporation of reaction and system thermodynamics.

In summary, *i*AF1260 represents a significantly expanded and comprehensively verified reconstruction of the *E. coli* metabolic network with broadened and enhanced predictive capabilities. With the growing number of studies based on previous versions of this reconstruction appearing, this work will enable a wider spectrum of studies focused on both proximal (i.e., immediate) and distal (i.e., over time) causation in biology. As the field of systems biology expands to incorporate cellular interactions from multiple core functions (e.g., regulation, signaling, etc.) on the genome-scale, *i*AF1260 will serve as a key component for the study of *E. coli* by providing an extensive picture of cellular metabolism.

4.5 Materials and methods

4.5.1 Network reconstruction

The reconstruction process has also been previously outlined[1,64]. Here we provide certain details specific to this work. Starting from the metabolic network for

*i*JR904[14], additional reactions were added to the network based on *E. coli* specific biochemical characterization studies (see Supplementary Data[27] for a full list of references) and other reactions were removed (see Results). This process was aided by comparing the content of *i*JR904 with the EcoCyc database (see below). The *E. coli* genome annotation[26] was used as a citation source for biochemical characterization studies and a framework upon which translated metabolic proteins, and subsequently reactions, were assigned to form gene to protein to reaction (GPR) assignments. Some reactions were also removed from *i*JR904 (see Results). The SimPheny™ (Genomatica Inc., San Diego, CA) software platform was used to build the reconstruction. For each reaction entered into the reconstruction, the participating metabolites were characterized according to their chemical formula and charge determined using their pKa value for a pH of 7.2. Metabolite charge was determined using its pKa value(s). When the metabolite pKa was not available, charge was determined using the pKa of ionizable groups present in a metabolite (http://www.chemaxon.com/product/pka.html). All of the reactions entered into the network were designated as enzymatically catalyzed reactions or spontaneous reactions, were both elementally and charged balanced and are either reversible or irreversible. Reversibility was determined first from primary literature for each particular enzyme/reaction, if available (see Supplementary Data[27] for references). Additionally, general heuristic rules like those applied by Kümmel and colleagues[58] were used to enter reversibility using knowledge about the physiological direction of a reaction in a pathway (sometimes including regulatory knowledge) and/or basic thermodynamic information (such as, reactions hydrolyzing high energy phosphate bonds are almost always irreversible). Furthermore, a thermodynamic analysis of reversibility was utilized to assign the directionality of some reactions (see above).

4.5.2 Comparison of *i*AF1260 and the EcoCyc and MetaCyc Databases

The comparison between the content of the *i*AF1260 and the EcoCyc[24] and MetaCyc[30] databases was performed in three phases. Initially, a list of metabolic ORFs contained in EcoCyc and not in *i*JR904 (the previous reconstruction) were manually evaluated for inclusion in *i*AF1260 in an effort to merge content. 176 out of 308 ORFs from this list were included into *i*AF1260 from manual analysis of this list or were included prior to this analysis from primary literature in a separate effort. Many of the inclusions in this phase were transporter encoding ORFs. A common type of ORF that was not included were those acting on non-specific metabolites (e.g., nonspecific drugs), proteins or RNA molecules.

The second phase of the comparison consisted of generating a full mapping of the metabolites contained in *i*AF1260 and EcoCyc or MetaCyc. This phase permitted the inclusion of compounds in each database that were missing from the other and identified possible errors in enzyme substrate specificity and metabolite structure. It also provided a future reference for linking of the metabolite content between the two resources. In an initial automated effort, mappings between metabolites in *I*AF1260 and EcoCyc/MetaCyc were established computationally using textual matching between the official name in *i*AF1260 to the common name and/or synonyms of metabolites in EcoCyc/MetaCyc, version 10.6. In addition, when available in both datasets, KEGG identifiers and CAS numbers were used to double-check matches or to make additional matches. After this computational step, 871 out of 1039 metabolites in *i*AF1260 were mapped to EcoCyc/MetaCyc. The remaining metabolites were mapped manually and changes to the content of *i*AF1260 made during this mapping process were facilitated by cross-referencing the ORFs that encoded for the

proteins that acted on specific metabolites in *i*AF1260 with their annotation in EcoCyc (see Results for findings and see Supplementary Data[27] for the mapping).

The final phase of the comparison was an automated mapping between reactions contained in *i*AF1260 and EcoCyc/MetaCyc. This phase generated a list of high-confidence reactions that both *i*AF1260 and EcoCyc contained, and provides a future reference for a full merging of the reaction content between the two resources. The automated reaction mapping was performed with software written specifically for this task, to accommodate frequently occuring types of differences between the models. The matcher parses the equations of every reaction *R* in *i*AF1260 and uses the previously described metabolite mappings to find the reaction object in EcoCyc/MetaCyc that contains the same set of metabolites as does *R*. Numerous reactions in *i*AF1260 contain protons in the equation that do not appear in EcoCyc, and the matcher can take into account this and other similar differences. The matcher also tries to find a generic reaction in EcoCyc that is specified in terms of compound classes, if the metabolite instances used in the equation in *i*AF1260 did not yield a direct match.

4.5.3 Generation of the biomass objective function (BOF)

The biomass objective function (BOF) was generated by defining all of the major and essential constituents that make up the cellular biomass content of *E. coli*. To determine these metabolites and their quantity, we used the dry weight composition data for an average *E. coli* B/r cell growing exponentially at 37°C under aerobic conditions in glucose minimal medium with an approximate doubling time of 40 min having a dry cell weight of 2.8×10^{-13} grams[38] (**Table 4.2**, see Supplementary Data[27]). Each cellular biomass macromolecule (i.e., protein, RNA, DNA, etc.) was

divided into its corresponding metabolic precursors present in the reconstruction (for example, L-alanine, UTP or dTTP, respectively). Each of the precursor metabolites was assigned a value that it contributes to the total percentage of the macromolecule, except for the soluble pool metabolites (e.g., thiamine diphosphate). This process was followed so that if the overall quantities of macromolecules were changed, the corresponding precursor metabolite would be scaled appropriately (cite the sensitivity analysis on composition). The quantity of soluble pool metabolites (approximately 2.9% of the total biomass) was taken from experimentally measured values or alternatively, it was estimated as a 0.1 mM intracellular concentration (see Supplementary Data[27] for a full list of references). From this data, a linear biomass objective function was formulated based on the wild type cell composition for *E. coli* and an ATP maintenance approximation to account for non-metabolic processes (see location of full equation). Using FBA, the model was analyzed to determine if each BOF metabolite could be generated from the defined minimal medium under both aerobic and anaerobic conditions with D-glucose, D-ribose and glycerol as the carbon and energy source. Only metabolites identified as cofactors could not be generated from the glucose minimal medium (discussed in Supplementary Data[27]).

Using the BOF_{WT}, gene essentiality and published data, a 'core' BOF was formulated that was consistent with the minimal set of macromolecular molecules needed for cell viability. The twenty common amino acids, inorganic ions, and nucleotide metabolites were all considered essential[38]. For the other BOF_{WT} metabolites, each metabolite was evaluated individually to determine if the genes that were necessary to synthesize the metabolite from minimal media substrates (see Supplementary Data[27]) were essential[42,43]. One macromolecule, glycogen, was not essential for cell viability because there were no essential ORFs encoding for

enzymes in the synthesis or breakdown of glycogen. The essential metabolites were defined by identifying the end product from the closest essential reaction to the BOF_{WT} metabolite (**Table 4.2**) in the possible *de novo* pathway(s) for biosynthesis. Molecules in this group, such as riboflavin were determined to be essential, whereas the wild type outer membrane *E. coli* K-12 LPS molecule was not found to be essential. However, a precursor of the common wild type LPS molecule, KDO_2-Lipid A, was found to be essential for cell viability[65]. Alternatively, 'core' metabolites were also determined from specific published studies. For example, thiamine diphosphate was found to be essential[60] whereas phosphatidylglycerol was determined not to be essential[44].

The ATP maintenance approximation in the BOFs which account for non-metabolic processes were approximated with the ATP utilization equation,

$$ATP + H_2O \rightarrow ADP + Pi + H^+ \qquad (1).$$

Where the number of ATP equivalents hydrolyzed is characterized in the GAM variable. The entire BOF is given in mathematical terms in Supplementary Data[27].

Aside from the BOF maintenance, a NGAM (mmol ATP gDW^{-1} hr^{-1}) value was used as an energy "drain" on the system during the linear programming calculations and accounts for non-growth cellular activities[39]. The NGAM was represented as a defined flux in the reaction flux vector, v_{NGAM} (see below and Supplementary Data[27]).

4.5.4 Modeling simulations

A stoichiometric matrix, $\boldsymbol{S}$ ($m \times n$), was constructed for *i*AF1260 where m is the number of metabolites and n is the number of reactions. The corresponding entry in the stoichiometric matrix, S_{ij}, represents the stoichiometric coefficient for the

participation of the i^{th} metabolite in the j^{th} reaction. FBA was then used to solve the linear programming problem under steady-state criteria[3] represented by the equation:

$$\boldsymbol{S} \bullet \boldsymbol{v} = 0 \qquad (2),$$

where $\boldsymbol{v}$ ($n \times 1$) is a vector of reaction fluxes. Since the linear problem is normally an underdetermined system for genome-scale metabolic models, there exists multiple solutions for $\boldsymbol{v}$ that satisfy Equation 1. To find a particular solution for $\boldsymbol{v}$, the cellular objective of producing the maximal amount of biomass constituents, represented by the ratio of metabolites in the BOF, is optimized for in the linear system. Additionally, constraints that are imposed on the system are in the form of:

$$\alpha_i \leq v_i \leq \beta_i \qquad (3)$$

where α and β are the lower and upper limits placed on each reaction flux, v_i, respectively. For reversible reactions, $-\infty \leq v_i \leq \infty$, and for irreversible reactions, $0 \leq v_i \leq \infty$. The constraints on the reactions that allow metabolite entry into the extracellular space were set to $0 \leq v_i \leq \infty$ if the metabolite was not present in the medium, meaning that the compounds could leave, but not enter the system. For the metabolites that were in the medium, the constraints were set to $-\infty \leq v_i \leq \infty$ for all except the limiting substrate(s) (e.g., glucose and/or oxygen). The reaction flux through the BOF was constrained from $0 \leq v_{BOF} \leq \infty$.

Linear programming calculations were performed using SimPheny™ (Genomatica, San Diego, CA) and the LINDO (Lindo Systems Inc., Chicago, IL) or TOMLAB (Tomlab Optimization Inc., San Diego, CA) solvers in MATLAB®, (The MathWorks Inc., Natick, MA) with the COBRA Toolbox[32].

When comparing the flux distribution in central metabolism to experimentally reported values[47], all of the comparisons were performed using computational results when optimal growth is predicted using the BOF_{CORE}, the 152 regulated reactions under these conditions constrained to zero (see above), a split in the flux ratio between the two NADH dehydrogenases of 1:1, a NGAM value of 8.39 mmol ATP gDW^{-1} hr^{-1}, a GAM value of 59.81 mmol ATP gDW^{-1} and *i*AF1260. A flux variability analysis on the optimal flux distribution yielded no flexibility in the central metabolism pathways examined in this study. From the Fischer and colleagues[47] study, data from *E. coli* growth in reactor conditions were used because the oxygen uptake and CO_2 secretion rates were reported, and the flux values that were used were based off ^{13}C-constrained flux balancing.

4.5.5 Sensitivity Analysis

The sensitivity analysis was performed under aerobic glucose limiting minimal medium conditions. For each analysis, the parameter being examined was varied while the glucose uptake rate was sequentially set between 0 - 10 mmol gDW^{-1} hr^{-1} for a series of simulation with the maximum oxygen uptake rate set to 18.5 mmol gDW^{-1} hr^{-1}. This maximum uptake rate was chosen since it closely matched the maximum uptake rate of oxygen observed *in vivo* (e.g., see[34,47]). All other modeling parameters were set to those determined in the physiological predictions section. The BOF_{CORE} objective function was used in all simulations (except those stated otherwise) since the predicted growth rate and oxygen uptake rate was found to be insensitive to the use of either the BOF_{CORE} or BOF_{WT}.

4.5.6 Alternate growth condition analysis

To determine the carbon, nitrogen, phosphorus and sulfur sources that could support simulated growth, we screened all of the metabolites that could be exchanged with the environment (i.e, exchange reactions) in the *i*AF1260 and *i*JR904 models. The identified metabolites formed the potential substrate sets (**Table 4.4**). Through subsequent simulations, we set an arbitrary maximum flux of 20 mmol substrate gDW^{-1} hr^{-1} for each potential substrate tested (consistent with maximum observed substrate uptake rates *in vivo*) and optimized for flux through the BOF_{CORE} using FBA and either *i*AF1260 or *i*JR904. An oxygen uptake rate of 18.5 mmol gDW^{-1} hr^{-1}, the BOF_{CORE}, a NGAM of 8.39 mmol ATP gDW^{-1} hr^{-1}, a GAM of 59.81 mmol ATP gDW^{-1} and no regulatory constrains were used during the growth condition analysis of *i*AF1260 (for *i*JR904, see[14]). During the analysis, the reactions CAT, SPODM and SPODMpp were constrained to zero to prevent generation of cellular energy equivalents through reactions involved in *E. coli's* response to oxidative stress. If a positive flux could be generated through the BOF_{CORE} reaction ($v_{BOFcore} > 0$), then the substrate was considered a viable source. Experimental data used in the comparison was provided by Biolog (http://www.biolog.com) and both 'weak' and 'positive' readings from the biolog data were considered as a positive growth condition.

4.5.7 Gene essentiality analysis

To determine the effect of a gene deletion, the reaction(s) associated with each gene in *i*AF1260 were individually deleted from ***S*** and FBA was used to predict the mutation growth phenotype. The simulations were performed using glucose minimal medium conditions with a glucose uptake rate of 10 mmol gDW^{-1} hr^{-1}, an oxygen uptake rate of 20 mmol gDW^{-1} hr^{-1}, the BOF_{CORE}, a NGAM of 8.39 mmol ATP gDW^{-1} hr^{-1}, a GAM of 59.81 mmol ATP gDW^{-1} and zero flux through the 152 reactions

regulated under glucose aerobic conditions (see Supplementary Data[27]). The flux through the BOF_{CORE} was optimized in the mutated network, ***S'***, and a positive flux through the BOF ($v_{BOFcore} > 0$) was considered non-essential (equation 2). Experimental criteria for gene essentiality is described in detail in Joyce and colleagues[43].

4.5.8 Standard Conditions for all estimated $\Delta_r G'^{\circ}$ and $\Delta_f G'^{\circ}$

All $\Delta_f G_{est}'^{\circ}$ and $\Delta_r G_{est}'^{\circ}$ calculated for the reconstruction using the group contribution method are based upon the standard condition of aqueous solution with pH equal to 7, temperature equal to 298.15 K, zero ionic strength, and 1 M concentrations of all species except H^+, and water. In the cases where multiple charged forms of a molecule exist at pH 7 (i.e., ATP^{4-} and $HATP^{3-}$), the most abundant form is used. This is consistent with the form of the molecules used in the fitting of the group contribution molecules (MD Jankowski and V Hatzimanikatis, personal communication).

The charges of the molecules and the proton balances for the reactions included in the reconstruction are based on a reference pH of 7.2. In order for the $\Delta_r G_{est}'^{\circ}$ values included with the reconstruction to match the reference pH of the reconstruction, all $\Delta_r G_{est}'^{\circ}$ calculated using the group contribution method (based on a reference pH of 7) were adjusted to a reference pH of 7.2 using the method described in[54]. The adjusted $\Delta_r G_{est}'^{\circ}$ values were used in the calculation of $\Delta_f G'^{m}$ and for all other thermodynamic analysis performed on the reconstruction. The pKa values for the compounds in the reconstruction used in the transformation of $\Delta_r G_{est}'^{\circ}$ to a reference pH of 7.2 were estimated from the molecular structures of the compounds using the MarvinBeans software developed by ChemAxon.

4.5.9 Adjustment of $\Delta_r G'^{\circ}$ to $\Delta_f G'^{m}$

The $\Delta_r G'^{m}$ calculated for all reactions contained in the reconstruction is based on the reference state of 1 mM concentrations for all species except H^+, water, H_2, and O_2. The reference concentrations for H_2 and O_2 are the saturation concentrations for these species in water at 1 atm, and 298.15 K. All $\Delta_r G'^{m}$ values reported in this work also include the energy contribution of the transmembrane electrochemical potential and proton gradient for all reactions involving transport across the cytoplasmic membrane assuming a periplasmic pH of 7.7 and a cytoplasmic pH of 7. All $\Delta_r G'^{m}$ calculated for reactions in the *i*AF1260 model are listed in Supplementary Data[27].

We also determined the direction of flux required in the reactions contained in *i*AF1260 to achieve near optimal growth (90% - 100%) on each of 174 carbon sources using FVA[52] and the BOF_{CORE}. It is worthwhile to note that the same set of reactions can or cannot be utilized in FVA simulations when examining approximately 5% – 95% of the optimal flux value achievable for the BOF_{CORE} under glucose aerobic conditions (one exception is the cytochrome oxidase bo and oxygen transport reactions which are needed for generating the necessary energy to achieve approximately 80% or greater of the BOF_{CORE} flux). During the FVA of conditions corresponding to glucose aerobic growth, the reactions CAT, SPODM and SPODMpp were constrained to zero to prevent generation of cellular energy equivalents through reactions involved in *E. coli's* response to oxidative stress, and the reaction formate hydrogenlyase, which appears to be involved in regulating cytosolic pH[66], was also constrained to zero to prevent the production of significant amounts of hydrogen gas that is not typically observed for most buffered experiments around pH 7. The results of the FVA indicated that some of the reactions in the reconstruction consistently

operated in the reverse direction. During the calculation of $\Delta_r G'^m$ for these reactions, the forward direction of each reaction was redefined to be in the direction of flux required for near optimal growth to occur. Because of this adjustment, all negative $\Delta_r G'^m$ and $\Delta_r G'$ values reported (see **Figure 4.2**) indicate reactions that are thermodynamically feasible in the direction of flux while positive values indicate thermodynamically infeasible reactions.

4.5.10 Estimation of achievable range of values for $\Delta_r G'$

The range of possible values for the $\Delta_r G'$ of a reaction depends not only on $\Delta_r G'^\circ$, but also the uncertainty in the estimated $\Delta_r G'^\circ$ ($U_{r,est}$), the activities of the metabolites involved in the reaction, and for transport reactions, the energy contribution of the electrochemical potential and proton gradient across the cytoplasmic membrane ($\Delta G_{Transport}$)[25]. $\Delta_r G'$ can deviate from $\Delta_r G'^{mM}$ because the activity of a metabolite can deviate from the reference value of 1 mM. The maximum and minimum values for $\Delta_r G'$ were calculated using the following equations.

$$\Delta_r G'_{\max} = \Delta_r G'^\circ + \Delta G_{\text{Transport}} + RT \sum_{i=1}^{\text{Products}} n_i \ln(x_{\max}) + RT \sum_{i=1}^{\text{Reactants}} n_i \ln(x_{\min}) + U_{r,est} \quad (4)$$

$$\Delta_r G'_{\min} = \Delta_r G'^\circ + \Delta G_{\text{Transport}} + RT \sum_{i=1}^{\text{Products}} n_i \ln(x_{\min}) + RT \sum_{i=1}^{\text{Reactants}} n_i \ln(x_{\max}) - U_{r,est} \quad (5)$$

where x_{min} is the minimal metabolite activity assumed to be 0.00001 M, and x_{max} is the maximum metabolite activity assumed to be 0.02 M. The physiological range of activities for the dissolved gasses H_2, O_2, and CO_2 is much lower than the range of activities for other metabolites involved in metabolism. For this reason all of the x_{min} values for H_2, O_2, and CO_2 were set to 10^{-8} M, which is approximately equivalent to one molecule per cell, and the x_{max} values for H_2, O_2, and CO_2 were set

to the saturation concentrations for these gasses in water at 298.15 K and 1 atm, 0.000034 M, 0.000055, and 0.0014 M respectively. The activity terms for H^+ and H_2O were left out of equations 4 and 5 because these activities have already been lumped into the $\Delta_rG'^{\circ}$.

The Δ_rG' ranges encompassed by $\Delta_rG_{min}'^{\circ}$ and $\Delta_rG_{max}'^{\circ}$ calculated for the reactions in the reconstruction were used to assign reversibility and directionality to the reactions based on the thermodynamic estimates. Reactions with exclusively negative Δ_rG' values were identified as thermodynamically irreversible in the forward direction; reactions with exclusively positive Δ_rG' values were identified as thermodynamically irreversible in the reverse direction; and reactions with both positive and negative Δ_rG' values were identified as thermodynamically reversible. FVA was then utilized to determine the directions in which each of the reactions in the reconstruction operated during near optimal growth on 174 carbon sources. In this way, reactions for which the direction of operation indicated by FVA conflicted with the direction of thermodynamic feasibility indicated by the Δ_rG' ranges were identified.

Acknowledgements

We would like to thank Kenyon Applebee, Edward Chuong, Ingrid Keseler, Sean Nihalani, Alan Ruttenberg, Milton Saier, Jan Schellenberger and Jeremy Zucker for their help in the generation and analysis of the reconstruction.

Chapter 4, in full, is adapted from an article that appeared in Nature Molecular Systems Biology, volume 3, number 121, pages 1-18, published June, 2007. The dissertation author was the primary author of this paper, which was co-authored by Dr. Christopher S Henry, Dr. Jennifer L Reed, Markus Krummenacker, Dr. Andrew R

Joyce, Dr. Peter D Karp, Dr. Linda J Broadbelt, Dr. Vassily Hatzimanikatis, and Dr. Bernhard Ø. Palsson.

References

1. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

2. Stelling J. Mathematical models in microbial systems biology. *Curr Opin Microbiol* 2004;7:513-518.

3. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

4. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 2006;7:198-210.

5. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;6:397-9.

6. Elena SF, Lenski RE. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* 2003;4:457-69.

7. Reed JL, Palsson BO. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol* 2003;185:2692-9.

8. Varma A, Boesch BW, Palsson BO. Biochemical production capabilities of *Escherichia coli*. *Biotechnology and Bioengineering* 1993;42:59-73.

9. Varma A, Palsson BO. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. *Journal of Theoretical Biology* 1993;165:477-502.

10. Majewski RA, Domach MM. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering* 1990;35:732-738.

11. Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering* 1997;56:398-421.

12. Pramanik J, Keasling JD. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnology and Bioengineering* 1998;60:230-238.

13. Edwards JS, Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A.* 2000;97:5528-5533.

14. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology* 2003;4:R54.1-R54.12.

15. Lee SY, Woo HM, Lee D-Y, Choi HS, Kim TY, Yun H. Systems-level analysis of genome-scale *in silico* metabolic models using MetaFluxNet. *Biotechnol. Bioproc. Eng.* 2005;10:425-431.

16. Alper H, Miyaoku K, Stephanopoulos G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* 2005;23:612-6.

17. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91:643-8.

18. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 2005;37:1372-5.

19. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. *Nature* 2006;440:667-70.

20. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 2004;427:839-843.

21. Nikolaev EV, Burgard AP, Maranas CD. Elucidation and structural analysis of conserved pools for genome-scale metabolic reconstructions. *Biophys J* 2005;88:37-49.

22. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429:92-6.

23. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 2006;103:17480-4.

24. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* 2005;33:D334-7.

25. Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys J* 2006;90:1453-61.

26. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, Kosuge T, Mori H, Perna NT, Plunkett G, 3rd, Rudd KE, Serres MH, Thomas GH, Thomson NR, Wishart D, Wanner BL. *Escherichia coli* K-12:

a cooperatively developed annotation snapshot-2005. *Nucleic Acids Res* 2006;34:1-9.

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

28. Lotierzo M, Tse Sum Bui B, Florentin D, Escalettes F, Marquet A. Biotin synthase mechanism: an overview. *Biochem Soc Trans* 2005;33:820-3.

29. Albe KR, Butler MH, Wright BE. Cellular concentrations of enzymes and their substrates. *J Theor Biol* 1990;143:163-95.

30. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier C, Zhang P, Karp PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucl. Acids Res.* 2006;34:D511-516.

31. Palsson BO. Two-dimensional annotation of genomes. *Nat Biotechnol* 2004;22:1218-9.

32. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

33. Williams I, Frank L. Improved chemical synthesis and enzymatic assay of delta-1-pyrroline-5-carboxylic acid. *Anal Biochem.* 1975;64:85-97.

34. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001;19:125-130.

35. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186-9.

36. Covert MW, Palsson BO. Transcriptional Regulation in Constraints-based Metabolic Models of *Escherichia coli*. *J Biol Chem* 2002;277:28058-64.

37. Barrett CL, Herring CD, Reed JL, Palsson BO. The global transcriptional regulatory network for metabolism in *Escherichia coli* attains few dominant functional states. *Proc Natl Acad Sci U S A* 2005;102:19103-19108.

38. Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the bacterial cell: a molecular approach. Sunderland, Mass.: Sinauer Associates, 1990:xii, 506.

39. Pirt SJ. The maintenance energy of bacteria in growing cultures. *Proc R Soc Lond B Biol Sci* 1965;163:224-31.

40. Helling RB. Speed versus Efficiency in Microbial Growth and the Role of Parallel Pathways. *J. Bacteriol.* 2002;184:1041-1045.

41. Gennis RB, Stewart V. Respiration. In: Neidhardt FC, ed. *Escherichia coli* and *Salmonella*. Washington, DC: ASM Press, 1996:217-261.

42. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;2:2006.0008.

43. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. *J Bacteriol* 2006;188:8259-8271.

44. Kikuchi S, Shibuya I, Matsumoto K. Viability of an *Escherichia coli* pgsA null mutant lacking detectable phosphatidylglycerol and cardiolipin. *J Bacteriol* 2000;182:371-6.

45. Calhoun MW, Oden KL, Gennis RB, de Mattos MJ, Neijssel OM. Energetic efficiency of *Escherichia coli*: effects of mutations in components of the aerobic respiratory chain. *Journal of Bacteriology* 1993;175:3020-5.

46. Noguchi Y, Nakai Y, Shimba N, Toyosaki H, Kawahara Y, Sugimoto S, Suzuki E. The energetic conversion competence of *Escherichia coli* during aerobic respiration studied by 31P NMR using a circulating fermentation system. *J Biochem (Tokyo)* 2004;136:509-15.

47. Fischer E, Zamboni N, Sauer U. High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived 13C constraints. *Anal Biochem* 2004;325:308-16.

48. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* 2006;38:1406-1412.

49. Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. *Biophys J* 2002;83:79-86.

50. Kümmel A, Panke S, Heinemann M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* 2006;2:2006.0034.

51. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-Based Metabolic Flux Analysis. *Biophys. J.* 2007;92:1792-1805.

52. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 2003;5:264-76.

53. Hakobyan M, Sargsyan H, Bagramyan K. Proton translocation coupled to formate oxidation in anaerobically grown fermenting *Escherichia coli*. *Biophys Chem* 2005;115:55-61.

54. Alberty RA. Thermodynamics of biochemical reactions. Cambridge, MA: Massachusetts Institute of Technology, 2003.

55. Thauer RK, Jungermann K, Decker K. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev.* 1977;41:100-80.

56. Grass G, Otto M, Fricke B, Haney CJ, Rensing C, Nies DH, Munkelt D. FieF (YiiP) from *Escherichia coli* mediates decreased cellular accumulation of iron and relieves iron stress. *Arch Microbiol* 2005;183:9-18.

57. Silver S. Bacterial resistances to toxic metal ions--a review. *Gene* 1996;179:9-19.

58. Kümmel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics* 2006;7.

59. Alexander K, Young IG. Alternative hydroxylases for the aerobic and anaerobic biosynthesis of ubiquinone in *Escherichia coli*. *Biochemistry* 1978;17:4750-5.

60. Vander Horn PB, Backstrom AD, Stewart V, Begley TP. Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in *Escherichia coli* K-12. *J Bacteriol* 1993;175:982-92.

61. Imielinski M, Belta C, Halasz A, Rubin H. Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities. *Bioinformatics* 2005;21:2008-16.

62. Ghim CM, Goh KI, Kahng B. Lethality and synthetic lethality in the genome-wide metabolic network of *Escherichia coli*. *J Theor Biol* 2005;237:401-11.

63. Varma A, Palsson BO. Parametric sensitivity of stoichiometric flux balance models applied to wild-type *Escherichia coli* metabolism. *Biotechnology and Bioengineering* 1995;45:69-79.

64. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2006;2:1-14.

65. Raetz CRH. Bacterial lipopopysaccharides: A remarkable family of bioactive macroamphiphiles. In: Neidhardt FC, ed. *Escherichia coli* and *Salmonella*. Washington, DC.: ASM Press, 1996:1035-1063.

66. Mnatsakanyan N, Bagramyan K, Trchounian A. Hydrogenase 3 But Not Hydrogenase 4 is Major in Hydrogen Gas Production by Escherichia coli Formate Hydrogenlyase at Acidic pH and in the Presence of External Formate. *Cell Biochem Biophys* 2004;41:357-66.

Chapter 5

The metabolic reconstruction and computational analysis of the archaeal methanogen *Methanosarcina barkeri* Fusaro, *i*AF692

5.1 Abstract

We present a genome-scale metabolic model for the archaeal methanogen *Methanosarcina barkeri*. We characterize the metabolic network and compare it to reconstructions from the prokaryotic, eukaryotic, and archaeal domains. Using the model in conjunction with constraint-based methods, we simulate the metabolic fluxes and resulting phenotypes induced by different environmental and genetic conditions. This represents the first large-scale simulation of either a methanogen or an archaeal species. Model predictions are validated by comparison to experimental growth measurements and phenotypes of *M. barkeri* on different substrates. The predicted growth phenotypes for wild type and mutants of the methanogenic pathway have a high level of agreement with experimental findings. We further examine the efficiency of the energy-conserving reactions in the methanogenic pathway, specifically the Ech hydrogenase reaction, and determine a stoichiometry for the nitrogenase reaction in *M. barkeri*. This work demonstrates that a reconstructed metabolic network can serve as an *in silico* analysis platform to predict cellular phenotypes, characterize

methanogenic growth, improve the genome annotation, and further uncover the metabolic characteristics of methanogenesis.

5.2 Introduction

Metabolic reconstruction is a process through which the genes, proteins, reactions and metabolites that participate in the metabolic activity of a biological system are identified, categorized and interconnected to form a network. Most often, the system is a single cell of interest and, by using the genomic sequence as a scaffold, reconstructions can incorporate hundreds of reactions that approximate the entire metabolic activity of a cell. With the growing availability of genome sequences for eukaryotic, prokaryotic and archaeal species, genome-scale metabolic reconstructions have been performed for organisms across all three of these domains (for a review, see[1]).

Within the eukaryotic and prokaryotic domains in particular, metabolic reconstructions have been analyzed using constraint-based methods, which simulate our current understanding of metabolism in an organism and drive experiments to verify modeling predictions[2]. Constraint-based methods enforce cellular limitations on biological networks such as physio-chemical constraints, spatial or topological constraints, environmental constraints, or gene regulatory constraints[3]. One specific example of metabolic modeling using a constraint-based approach is flux balance analysis (FBA). FBA uses linear optimization to determine the steady-state reaction flux distribution in a metabolic network by maximizing an objective function, such as ATP production or growth rate[4]. The effectiveness of metabolic modeling using constraint-based methods has been demonstrated in predicting the outcomes of gene deletions[5], identifying potential drug targets[6], engineering optimal production stains

for bioprocessing[7] and elucidating cellular regulatory networks[2]. Analytical methods are continually being developed to understand additional emergent properties of metabolic models and to expand their application; for a review see Price and colleagues[3].

Surprisingly, constraint-based analysis has not yet been applied to study the metabolism of methanogenesis or archaeal organisms. Although high quality organism-specific metabolic pathway databases are available for several archaea[8]; see also BioCyc website, http://biocyc.org/biocyc-pgdb-list.shtml#tier2), such databases have not yet been curated for constraint-based analysis which requires that the network i.) has been evaluated to produce biomass constituents, such as amino acids, nucleotides and lipids, ii.) has sufficient representation of how metabolites enter and leave the cell, and iii.) contains explicit substrates, products and reversibility for all reactions.

Among archaea, methanogens are an attractive model because of their utilization of low carbon substrates, metabolic diversity, and the availability of detailed information on their metabolism. Methanogens also have major environmental and economic importance. They serve as a key component of the carbon cycle by degrading low carbon molecules in anaerobic environments to generate methane. Because of this, methanogens have been used for processing of industrial, agricultural and toxic wastes rich in organic matter[9]. Methanogens contribute to the greenhouse effect and are a potential source of renewable energy[10]. Moreover, a number of methanogenic archaea can form syntrophic relationships with eubacteria, allowing for the study of metabolite and energy coupling across species[11]. While many pieces of methanogenic metabolism are understood, there are still many

questions to be answered about the biochemistry of methanogenesis and how these pieces work together in the context of the whole organism. Reconstruction and analysis at a genome scale would better determine the biochemical properties of key components and analyze methanogenic metabolism as a whole in its cellular context.

To better understand the general metabolic capabilities of the archaeal domain, and methanogenesis in particular, we reconstructed the metabolic network of the archaeal methanogen *Methanosarcina barkeri* and performed constraint-based analysis on the genome-scale model. *M. barkeri* is one of the most versatile methanogens and is capable of growing on all three of the major methanogenic substrates: methanol, acetate and H_2/CO_2[9]. An isolated strain was also shown to utilize the uncommon methanogenic substrate, pyruvate[12]. Our metabolic reconstruction, labeled *i*AF692 following a previously established naming convention[13], represents the first curated genome-scale model of an archaea generated specifically for constraint-based modeling. We use this model to determine the growth capabilities for *M. barkeri* for both wild type (WT) and mutant strains. We also examine the maintenance energy requirements for growth, minimal media requirements and the stoichiometry of energy-conserving proton and ion translocating reactions in the methanogenic process.

5.3 Results and Discussion

5.3.1 Reconstructing the *M. barkeri* model

The metabolic reconstruction of *M. barkeri*, *i*AF692, was generated and refined using an iterative model building procedure (see **§5.5** and **Figure 5.1**). The model contains 692 metabolic genes associated with 509 reactions and 558 distinct

metabolites (see **Table 5.1**). An additional 110 reactions were included because they have been reported in prior literature, or because they were required to fill a gap in the reconstructed network (see **§5.5**). However, these are currently unassociated with any gene product in the annotation. *i*AF692 and constraint-based optimization results are available as Systems Biology Markup Language (SBML) files (level 2, version 1, http://sbml.org/, see Supplementary Data[14]) or in spreadsheet form (see Supplementary Data[14]).

Table 5.1: Properties of the archaeal metabolic reconstructions of *M. barkeri* and *M. jannaschii.*

	*i*AF692 *M. barkeri*	Tsoka *et al* (2004) *M. jannaschii*
Genome size	4.8 Mb	1.7 Mb
ORFs	5072	1792
Included genes	692	436
Unique proteins	542	266
Multiple gene associations	96	67
Enzyme complexes	65	NR
Instances of isozymes	31	NR
Reactions	619	609
Gene associated	509 (82%)	297 (49%)
No gene association	110 (18%)	312 (51%)
Transport reactions	88	1
Metabolites	558	510
Transported metabolites	56	1
Freely diffusible	14	NR

NR, not reported

The reactions in *i*AF692 were subdivided into eight high-level functional categories based on the major metabolic roles of the cell (**Figure 5.2**). The largest number of reactions (153) was involved in the biosynthesis of vitamins and cofactors, probably because *M. barkeri* synthesizes many large molecular weight cofactors which require multiple enzymatic steps [15]. *M. barkeri* contains all of the *de novo*

pathways required to synthesize the 20 common amino acids[9] and these pathways, containing 141 gene-associated reactions, are well characterized in methanogenic archaea[16]. Transport reactions were another major functional class. Of the 88 transport reactions, 54 were included from the annotation, 31 were included from physiological data alone, while three were added from growth simulation requirements. The high number of transport reactions with no gene assignment in *M. barkeri* points to the fact that further work is needed to characterize the mechanisms and machinery involved in the transport of molecules in archaea.

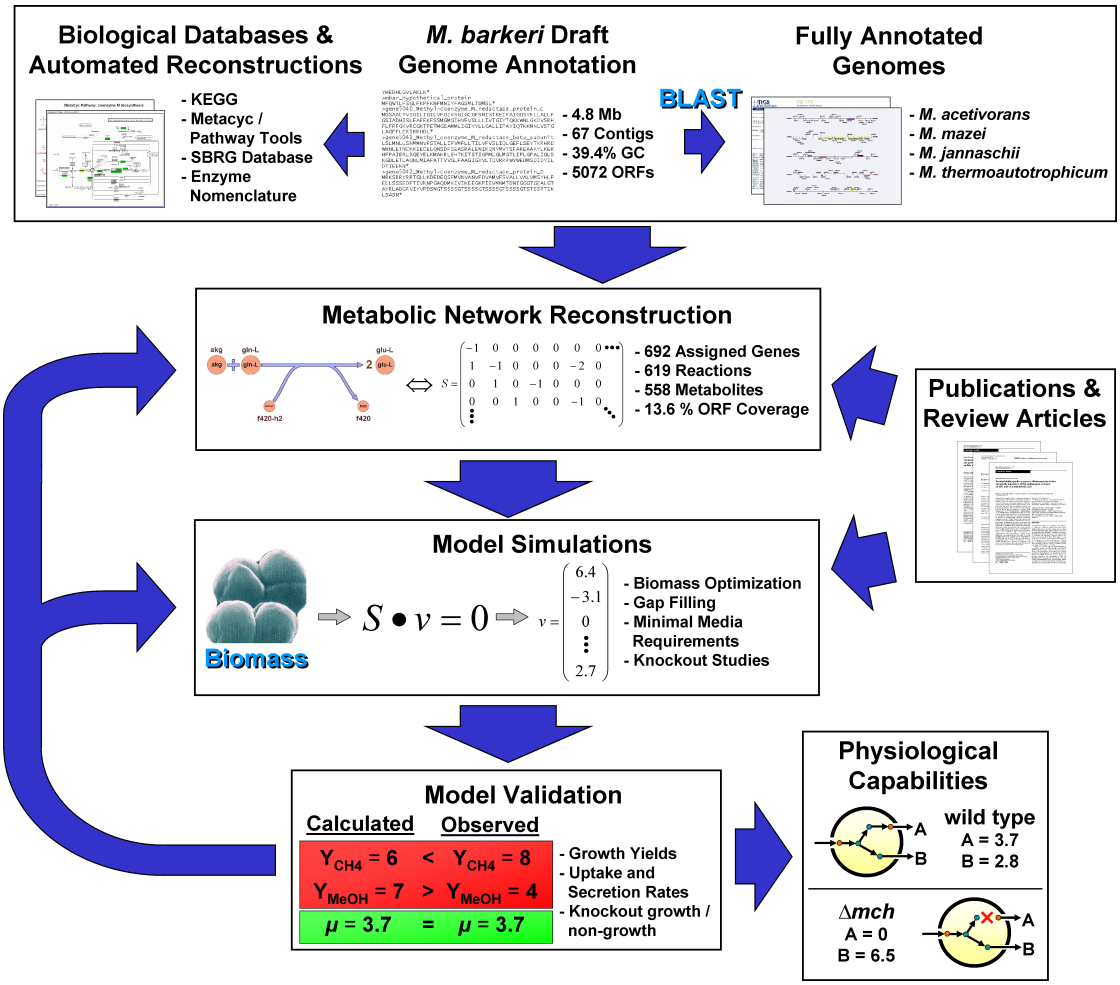
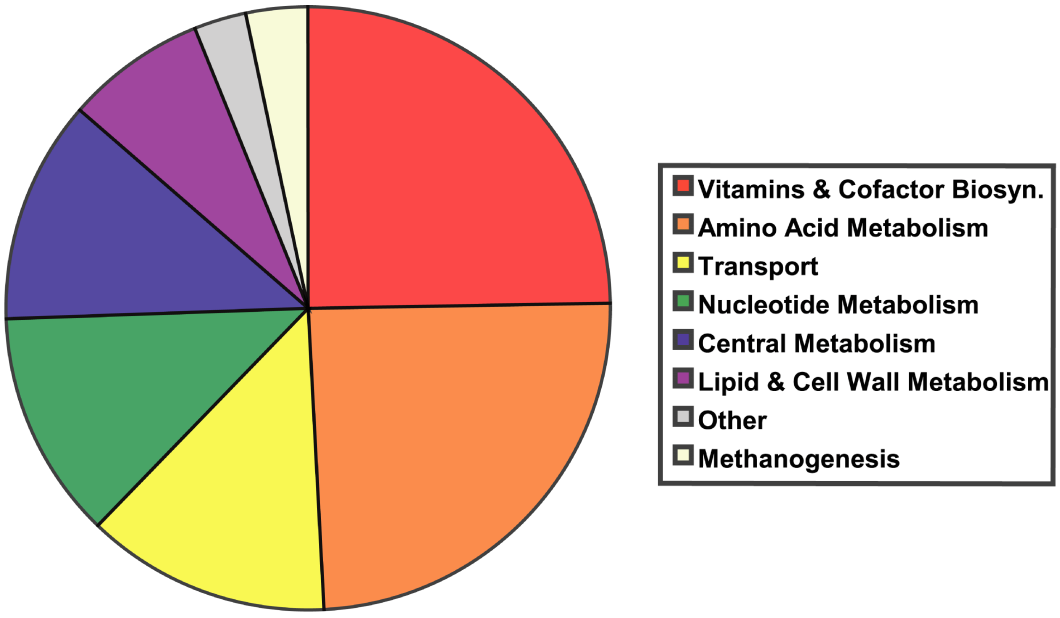


Figure 5.1: The iterative model building procedure used to generate *i*AF692. The draft genome annotation was used as a scaffold, on which gene-protein-reaction assignments were made. The reactions added to the model were taken from both biochemical databases and published data. Once a reaction was found to be in the network, it was manually curated and either associated to a potential ORF or added with no gene assignment. A biomass objective function was formulated to perform model simulations based on cellular composition. Modeling simulations were run under steady-state conditions to determine the reaction flux distribution in the network. The results from the simulations were interpreted and compared to experimental data. From the comparison, physiological capabilities of the cell were confirmed or the network was further refined or updated.



	All Reactions		Gene Association		No Association	
Pathways	No. rxns	% of total	No. rxns	% of total	No. rxns	% of total
Vitamins & Cofactor Biosyn.	153	25%	105	17%	48	8%
Amino Acid Metabolism	150	24%	141	23%	9	1%
Transport	82	13%	48	8%	34 (14[a])	5%
Nucleotide Metabolism	75	12%	74	12%	1	0%
Central Metabolism	72	12%	64	10%	8	1%
Lipid & Cell Wall Metabolism	46	7%	39	6%	7	1%
Other	18	3%	15	2%	3	0%
Methanogenesis	23	4%	23	4%	0	0%
Total	*619*		*509*	*82%*	*110*	*18%*

Figure 5.2: The distribution of reactions in *i*AF692. The table gives the pathway distribution for the total, gene-associated and non gene-associated reactions. Non gene-associated reactions were added on the basis of biochemical, physiological or modeling evidence.

[a] denotes that 14 reactions are diffusion reactions and would not require a gene association

5.3.2 Reconstruction as an annotation tool

The *i*AF692 model suggests 55 new functional annotations for predicted ORFs in the *M. barkeri* genome. These ORFs were either uncharacterized (30 genes) or likely misannotated in the draft annotation (25 genes). The model assists with functional annotation in cases in which a gene has multiple strong BLAST hits versus

other species, or has only weak sequence homologies to other genes. The model acts to filter these lists of ambiguous matches, by indicating which homologous genes fulfill a metabolic requirement of the cell or bridge a gap between metabolites in the network. A list of the potential ORFs annotated during the reconstruction is given in Supplementary Data[14].

One example of a functional prediction made during reconstruction is the case of 7,8-didemethyl-8-hydroxy-5-deazariboflavin (FO) synthase. FO is a choromophone that comprises part of the methanogenic cofactor coenzyme F_{420}[17]. *M. barkeri* has been verified to produce coenzyme F_{420} for use in the methanogenic process[18]. Although there are no genes annotated as FO synthase in *M. barkeri* or in any of the other *Methanosarcina* species, the enzyme has been characterized in *M. jannaschii* and is catalyzed by two different subunits, *CofG* and *CofH*[17]. Three sequential genes from contig 187 (*gene 838, 839, 840*) of the *M. barkeri* draft annotation were identified as orthologs to the biochemically verified genes *CofG* and *CofH* from *M. jannaschii* using BLAST. *Gene 838* is a predicted ortholog to the *CofG* gene, whereas *Genes 839* and *840* are two predicted paralogs that are othologous to the *CofH* gene of *M. jannaschii*. The sequential chromosomal location of the three genes on contig 187 also supports the gene-protein-reaction assignments in the model.

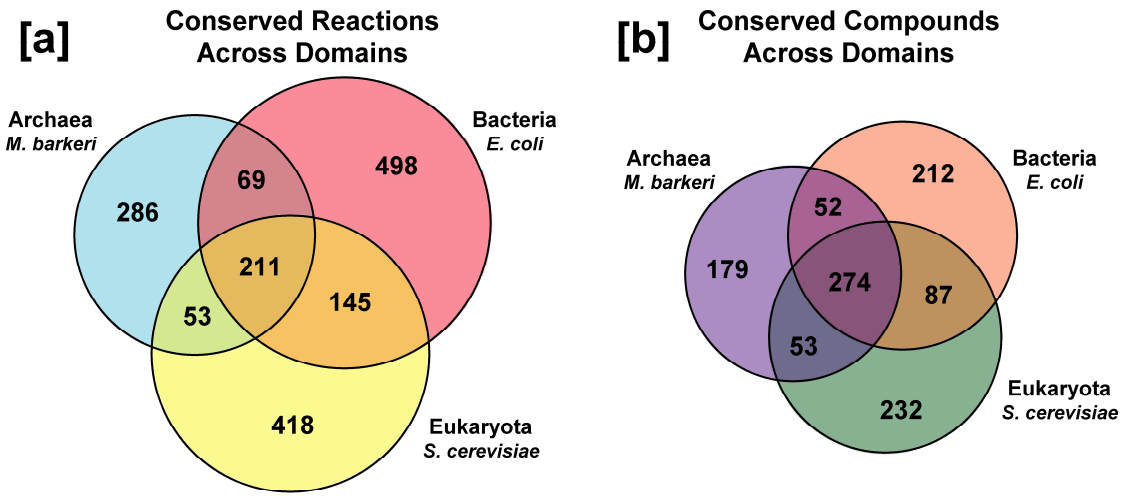
5.3.3 Comparison of *i*AF692 with previous metabolic reconstructions

The major differences between *i*AF692 and previous archaeal reconstructions[8]; see also BioCyc website, http://biocyc.org/) are: i.) the number of gene-associated reactions, ii.) the organism specificity of the reactions, iii.) defined reversibility and assurance of elementally and charge balanced reactions, iv.) the inclusion of sufficient transport reactions to support growth, v.) incorporation of

physiological information and vi.) further curation of the model after comparison of flux simulations to experimental data. For instance, in the reconstruction of *Methanococcus jannaschii* by Tsoka and colleagues[8], 49% of the reactions in the network were gene-associated after curation in comparison to 82% in *i*AF692 (**Table 5.1**). It is surprising that the metabolic model of *M. jannaschii* had roughly the same number of reactions as *i*AF692, but a much smaller genome. Although this result could indicate that many *M. barkeri* genes encode for functions that are non-metabolic, it was at least partially due to the fact that reactions involving DNA, proteins, and unspecified products/substrates were included in the *M. jannaschii* reconstruction, and that some predicted ORFs from the *M. barkeri* draft annotation may not be real genes.

We also systematically compared *i*AF692 to previous metabolic models from the prokaryotic and eukaryotic domains, the other two domains of life. **Figure 5.3** compares the content (reactions and metabolites) of the *M. barkeri* model with that of *Escherichia coli*, *i*JR904[13], and *Saccharomyces cerevisiae*, *i*ND750[5]. All of these models shared a core set of 211 reactions (12.6 % overall) and 274 metabolites (25.2% overall), indicating that the metabolites contained in the models are more highly conserved than the biochemical conversions between them (the reactions). This core set of reactions predominately involved biosynthesis and degradation of amino acids and nucleotides; a full list of conserved reactions is available in Supplementary Data[14]. Several pathways encoded by the model were primarily found only in or are specific to archaea. For instance, *i*AF692 contains all of the reactions in the methanogenic pathway necessary for growth on all known *M. barkeri* substrates (23 reactions associated with 125 distinct genes) and the biosynthetic pathways to generate all of the specific *Methanosarcina* species cofactors. Included are the

biosynthetic pathways for coenzyme M, coenzyme B, tetrahydrosarcinapterin (H_4SPT), coenzyme F_{420}, coenzyme F_{430}, coenzyme F_{390} and the anaerobic pathway for the synthesis of a vitamin B12 derivative (see Supplementary Data[14] for references). *i*AF692 also contains the biosynthesis pathways for the unique archaeal ether-linked lipids (46 reactions generating nine distinct lipids,[19]). Fifty-two transport reactions were unique to *M. barkeri*, probably because of the specialized nature of many of the methanogenic substrates[20]. A map of the complete metabolic network of *i*AF692 including the methanogenic pathway, the lipid biosynthesis pathway, vitamin and cofactor biosynthesis, amino acid metabolism and nucleotide metabolism is available in Supplementary Data[14].



	Similar Across All Domains		*M. barkeri* & *S. cerevisiae*		*M. barkeri* & *E. coli*		*E. coli* & *S. cerevisiae*		Unique to *M. barkeri*	
	No. of Reactions	% of Total	No. of Reactions	% of Total	No. of Reactions	% of Total	No. of Reactions	% of Total	No. of Reactions	% of Total
Amino Acid Metabolism	84	40%	110	42%	95	34%	105	29%	29	10%
Nucleotide Metabolism	53	25%	59	22%	56	20%	92	26%	13	5%
Vitamins & Cofactor Biosynthesis	32	15%	37	14%	44	16%	58	16%	104	36%
Central Metabolism	27	13%	34	13%	42	15%	52	15%	25	9%
Transport	7	3%	11	4%	25	9%	34	10%	52	18%
Lipid & Cell Wall Metabolism	4	2%	9	3%	11	4%	8	2%	30	10%
Other	4	2%	4	2%	7	3%	7	2%	13	5%
Methanogenesis	0	0%	0	0%	0	0%	0	0%	20	7%
Total:	*211*		*264*		*280*		*356*		*286*	

Figure 5.3: Conserved reactions and compounds among reconstructed metabolic models from the three phylogenetic domains. The models of the archaea *M. barkeri*, *i*AF692, the bacteria *E. coli*, *i*JR904, and the eukaryote *S. cerevisiae*, *i*ND750, were compared to determine identical reactions and compounds contained in the models. All of the models were decompartmentalized so that only the reactions and cytosolic transporters were compared and not their location inside of the cell. The table gives information on the distribution of reactions in their respective pathways.

A comparison of global topological properties of the metabolic networks is given in **Table 5.2**. Comparing *M. barkeri* to other models generated specifically for constraint-based analysis (see **Table 5.2**), *M. barkeri* and *S. cerevisiae* were more similar to each other than to *E. coli*. For instance, *M. barkeri* and *S. cerevisiae* had a longer average path length than *E. coli*, and also had a smaller average degree and network diameter. These findings show that the *E. coli* metabolic network is more connected than of *M. barkeri* and *S. cerevisiae* and suggest that these models have less redundancy in their network structure. All three networks followed a power law degree distribution implying that the models are scale-free networks (**Figure 5.4**) and also contained one large connected component of reactions (the giant strong component (GSC), see[21]) along with several isolated sub-networks composed of linear and significantly smaller connected pathways. As argued by Ma and Zeng[21], the GSC contains most of the core metabolites. The number of metabolites in each sub-network is given in **Table 5.2** and the metabolites present in each sub-network for each model are given in Supplementary Data[14]. *i*AF692 contained fewer links (reactions) and nodes (metabolites) than *i*JR904 or *i*ND750. This was not surprising given the level of genetic characterization of both *E. coli* and *S. cerevisiae*[22]. **Table 5.2** also shows that constraint-based models are more connected than those generated from biochemical databases and reversibility rules[21]).

Table 5.2: Network properties for selected metabolic reconstructions.

Organism	Reconstruction reference	Links[a] (irr/rev)	Mets[b]	APL	D	⟨k⟩	SC	Network structure subsets			
								GSC	S	P	IS
M. barkeri	This study	636/253	592	8.00	24	3.03	7	322	54	131	85
E. coli	Reed *et al* (2003)	1076/333	725	6.75	19	3.89	8	468	145	65	47
S. cerevisiae	Duarte *et al* (2004)	1073/536	972	8.00	31	3.36	11	629	98	125	120
E. coli	Ma and Zeng (2003)	NR	811	8.20	23	NR	29	274	93	161	283
S. cerevisiae	Ma and Zeng (2003)	NR	679	9.71	NR	NR	NR	206	54	164	255

The network properties for the metabolic reconstructions generated by Reed *et al*[13], Duarte *et al*[5], and those of the *M. barkeri* model were calculated in this study. These models were built specifically for use with constraint-based methods. The additional network properties were reported for reconstructions generated by Ma and Zeng[21]. irr, irreversible; rev, reversible; mets, metabolites; APL, average path length; D, network diameter; <k>, average degree; SC, strong components; GSC, giant strong component; S, substrate subset; P, product subset; IS, isolated subset; NR, not reported [a] Model compartmentalization was conserved[5]. [b] Currency metabolites were removed from each network.

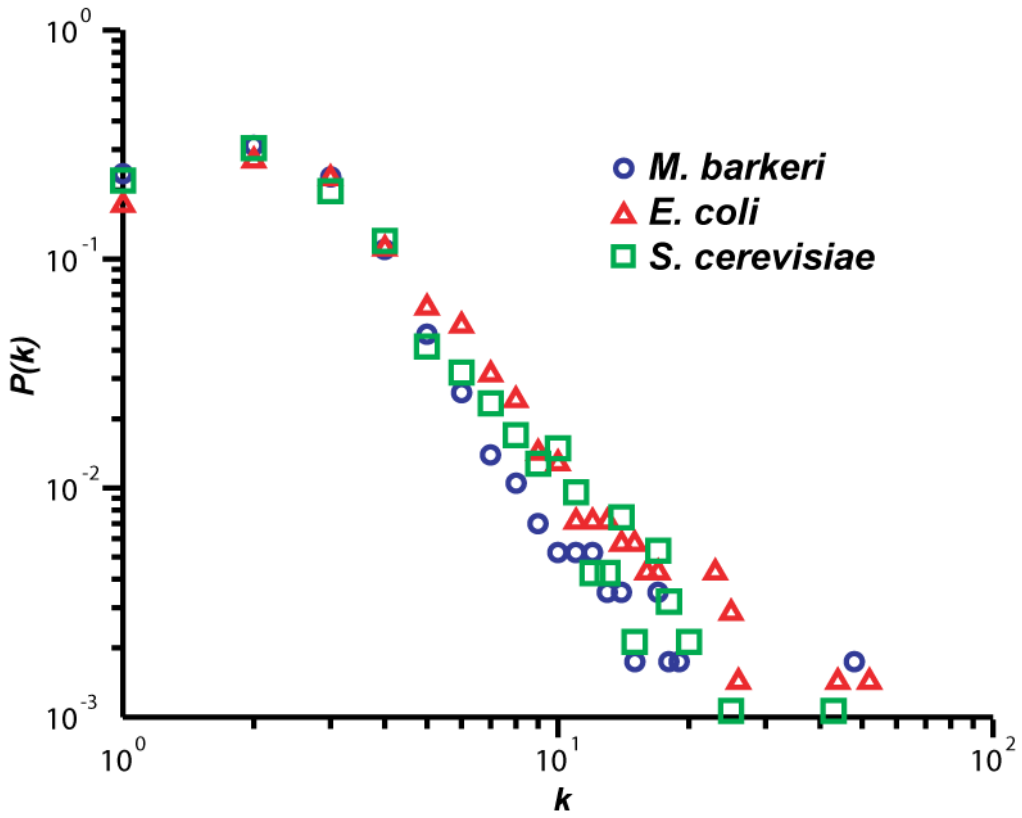


Figure 5.4: The degree distribution for the three metabolic models from each phylogenic domain. The degree distribution for the three metabolic models (*i*AF692, *i*JR904 and *i*ND750) indicate a scale-free topology for the networks[23]. The degree distribution, P(k), give the probability that a given metabolite has exactly k links and was calculated after removing currency metabolites from the network.

5.3.4 Computational analysis of minimal media for *M. barkeri*

Minimal media conditions were determined which could produce all of the biomass constituents found in the biomass objective function (BOF) for *M. barkeri*. The BOF is a linear equation consisting of the molar amounts of metabolites that comprise the dry weight content of the cell (**Table 5.3**) along with a growth maintenance reaction (see Supplementary Data[14]). Optimization of the network to maximize the reaction flux through the BOF simulates a cell which strives to maximize

the generation of biomass constituents from available media substrates. Beyond the primary source (methanol, acetate, CO_2 or pyruvate), two additional carbon-containing compounds were needed to generate the metabolites present in the BOF: p-aminobenzoic acid and nicotinic acid. P-aminobenzoic acid (pABA) is needed for the biosynthesis of folic acid[24] and H_4SPT[15], while nicotinic acid is used in the generation of nicotinaminde coenzymes.

However, Buchenau and Thauer[24] reported that pABA may not be necessary for the biosynthesis of H_4SPT. A possible alternate pathway in *M. barkeri* is discussed below. Other carbon-containing compounds commonly found in *M. barkeri* media[25] such as biotin, folic acid (from pABA), thiamine, pantothenate, and vitamin B12 can be synthesized by the network and thus were not essential for simulated growth. There is corroborating experimental evidence that *M. barkeri* was not dependent on these compounds for optimal growth[26]. On the other hand, Scherer and Sahm[26] stated that riboflavin (found to be non-essential using *i*AF692) was required for optimal growth. This finding suggests that the *de novo* pathway to synthesize riboflavin in *M. barkeri* (http://genome.ornl.gov/microbial/mbar/, described for similar archaea by Fischer and colleagues,[27]) may not be sufficient for optimal growth. In addition to carbon-containing compounds, the required compounds are metals, phosphate, sulfur and nitrogen. The stoichiometry of the nitrogenase reaction in *M. barkeri* was computationally determined (see below) and further details for the other media requirements are provided in Supplementary Data[14].

Table 5.3: Biomass composition of *M. barkeri*.

Metabolite	mmol gDW^{-1}	Metabolite	mmol gDW^{-1}	Metabolite	mmol gDW^{-1}
Protein (63%)[a]		*RNA (24%)*[a]		*Soluble pools (4%)*	
Alanine	0.5621	ATP	0.1846	Putrescine	0.0262
Arginine	0.3237	CTP	0.1379	Homospermidine	0.0047
Asparagine	0.2638	GTP	0.2222	Acetyl-Coenzyme A	0.0001
Aspartate	0.2638	UTP	0.1489	Coenzyme A	<0.0001
Cysteine	0.1002	*DNA (4%)*[a]		NAD^+	0.0022
Glutamine	0.2880	dATP	0.0331	NADH	0.0001
Glutamate	0.2880	dCTP	0.0215	$NADP^+$	0.0001
Glycine	0.6704	dGTP	0.0215	NADPH	0.0004
Histidine	0.1037	dTTP	0.0331	Succinyl-Coenzyme A	<0.0001
Isoleucine	0.3179	*Lipid (5%)*		Coenzyme M	0.0206
Leucine	0.4930	Archaetidylglycerol	0.0007	Coenzyme F420	0.0008
Lysine	0.3755	Hydroxyarchaetidylglycerol	0.0101	Tetrahydrosarcinapterin	0.0236
Methionine	0.1682	Archaetidylinositol	0.0027	Adenosylcobalamin-HBI	0.0047
Phenylalanine	0.2027	Hydroxyarchaetidylinositol	0.0135	Coenzyme F430	0.0020
Proline	0.2419	Archaetidylethanolamine	0.0007	Coenzyme B	0.0005
Serine	0.2361	Hydroxyarchaetidylethanolamine	0.0027	5,6,7,8-Tetrahydrofolate	0.0001
Threonine	0.2776	Archaetidylserine	0.0013	Coenzyme F390	<0.0001
Tryptophan	0.0622	Hydroxyarchaetidylserine	0.0115	ATP	0.0040
Tyrosine	0.1509	Glucosaminyl archaetidylinositol	0.0162	ADP	0.0020
Valine	0.4631	*Carbohydrates (<1%)*		AMP	0.0010
		Glycogen	0.0154		

[a] Relative ratios of wt % of protein, RNA, and DNA were estimated from the composition of a typical prokaryotic cell[28]. The values were taken from published data and converted to mmol gDW^{-1} and the wt% of each category is given (for references, see Supplementary Data[14]).

5.3.5 Estimation of the proton translocation efficiency of the Ech hydrogenase reaction

To demonstrate the predictive power of *i*AF692, we examined three different uncharacterized areas of *M. barkeri* metabolism using a model driven approach: i.) the stoichiometry of the Ech hydrogenase reaction, ii.) the stoichiometry of the nitrogenase reaction, and iii.) an alternate pathway for the biosynthesis of H_4SPT.

Although the methanogenic process is well defined and has been considerably reviewed (see Supplementary Data[14] for references), several aspects are still poorly understood. One of these aspects is the efficiency of the energy-conserving ion translocating reactions of the methanogenic pathway, specifically the Ech hydrogenase catalyzed reaction. These reactions couple conversion of metabolites with the transport of ions across compartmental membranes to create an electrochemical potential[29]. This electrochemical potential is used to generate ATP

through ATP synthase and to drive reactions that are otherwise energetically unfavorable[30]. Ech hydrogenase is one of the six energy-conserving ion translocating enzymes of the methanogenic pathway in *M. barkeri*, and the only one for which the stoichiometry is unknown (i.e., the number of protons translocated per electrons transferred)[30,31].

We examined the effect of the Ech hydrogenase reaction stoichiometry on the flux distribution and resulting growth yields for *M. barkeri* using *i*AF692. In this approach, we used FBA and BOF optimization to determine which choices of Ech hydrogenase stoichiometry resulted in the experimentally observed growth yields for various substrates (see **§5.5**). However, to calculate a flux distribution, additional parameters were needed. Two of these parameters, the growth associated maintenance (GAM) and non-growth associated (NGAM) maintenance, were also unknown due to the lack of experimental data. To reduce the number of unknown variables, the NGAM was set as 2.5% of the GAM based on previous analyses (see **§5.5**). This left the GAM and the stoichiometry of the Ech hydrogenase reaction as the remaining unknowns. A constraint on the Ech hydrogenase reaction stoichiometry was that it cannot exceed 2 protons translocated/1e^- due to thermodynamic reasons[32]. Given that two electrons are transferred in the Ech hydrogenase reaction, this provided a possible range of 0 - 4 protons translocated/2e^-. However, the value probably lies closer to that of another hydrogenase from the same family, hydrogenase 3 from *E. coli*, which had an apparent stoichiometry of 1.3 protons translocated/2e^-[32,33].

For proton translocation efficiencies in the feasible range (0 - 4 protons translocated/2e^-), FBA simulations were used to find the corresponding ranges of

GAM values that were consistent with observed growth yields. The variability in these ranges is shown in **Figure 5.5A** as a function of the Ech hydrogenase reaction stoichiometry (0 - 2.0 protons translocated/$2e^-$). Any stoichiometry > 2.0 protons translocated/$2e^-$ created an increasingly larger variability in the GAM values consistent with observed yields across all substrates (see **Figure 5.5** and **Figure 5.6**). **Figure 5.5B** and **Figure 5.5C** show the regions of experimentally observed growth yields for each substrate calculated at a given stoichiometry. A probable stoichiometry for the Ech hydrogenase reaction would be approximately 1.1 protons translocated/$2e^-$ if similar GAM values were found for growth on different substrates (**Figure 5.5C**). Using this dataset, a stoichiometry ≥ 2.0 or ≤ 0.2 protons translocated/$2e^-$ appeared unlikely because i.) GAM values typically vary less than 2.5 fold across all substrates for an extreme case[34] and ii.) the lowest value of GAM which produced a consistent yield is approaching the minimum theoretical cost for the polymerization of cellular macromolecules, 26 mmol ATP gDW^{-1} (see **Figure 5.5B** and **Figure 5.6**). The overall rates of end-product formation and product/substrate yields produced during FBA simulations were determined to be unique for each substrate and consistent with experimental data (see Supplementary Data[14]).

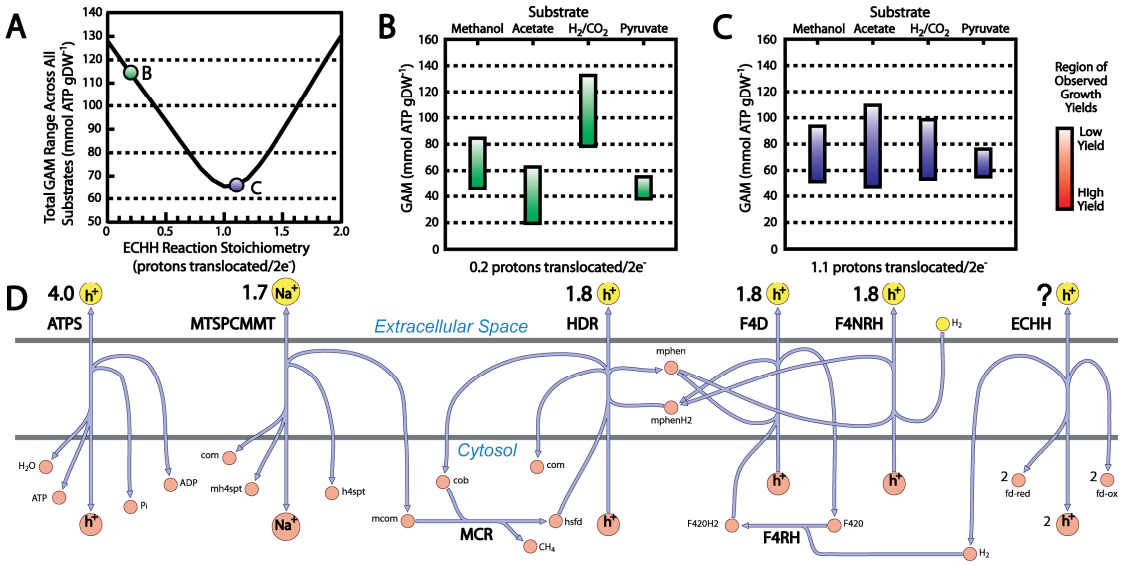


Figure 5.5: The effect of the Ech hydrogenase reaction stoichiometry on growth yields. Shown in panel A is the variability in the growth associated maintenance (GAM) that will produce growth yields consistent with experimental data. This variability is shown as a function of Ech hydrogenase reaction stoichiometry. Yields were calculated using *i*AF692 for four different substrates (methanol, acetate, H_2/CO_2 and pyruvate). Panel A shows the total GAM variability across all substrates. Panels B and C show the GAM variability for each substrate at a constant Ech hydrogenase reaction stoichiometry (0.2 protons translocated/$2e^-$ (B, green) and 1.1 protons translocated/$2e^-$ (C, purple)). The ranges calculated were constrained by experimental values. Panel D is a diagram of all energy-conserving ion translocating reactions in *M. barkeri*, each labeled with the stoichiometry of the translocated ion. Any value greater than 2 protons translocated/$2e^-$ for Ech hydrogenase created a larger total GAM range across all substrates. Abbreviations: ATPS, ATP synthase; MTSPCMMT, methyl-H_4SPT:coenzyme M methyltransferase; HDR, heterodisulfide reductase; F4NRH, F_{420}-nonreducing hydrogenase; F4D, $F_{420}H_2$ dehydrogenase; ECHH, Ech hydrogenase; MCR, methyl-coenzyme M reductase; F4RH, F_{420}-reducing hydrogenase; Pi, pyrophosphate; com, coenzyme m, mh4spt, methyl-tetrahydrosarcinapterin, h4spt, tetrahydrosarcinapterin, mcom, methyl-coenzyme M; cob, coenzyme B; hsfd, heterodisulfide; mphen, oxidized methanophenazine; mphenH2, reduced methanophenazine; $F_{420}H2$, reduced coenzyme F_{420}; F_{420}, oxidized coenzyme F_{420}; fd-red, reduced ferredoxin; fd-ox, oxidized ferredoxin

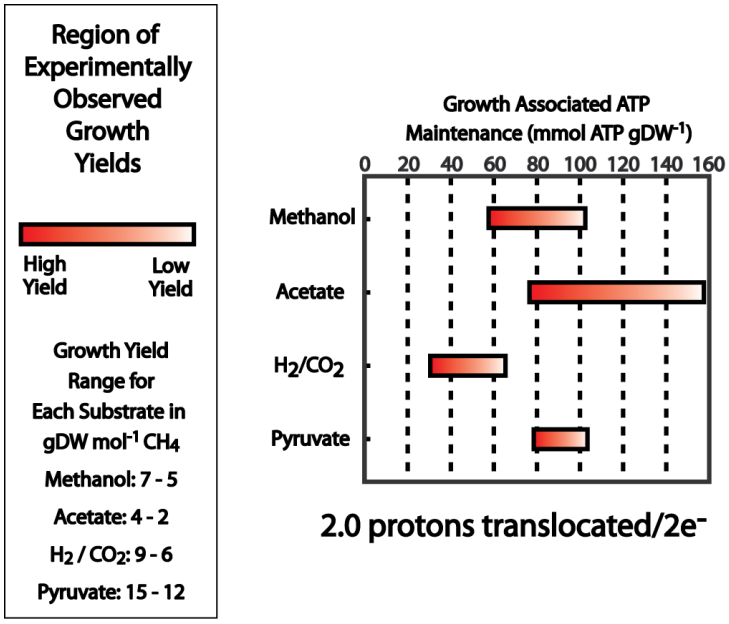


Figure 5.6: Analysis of growth yields using FBA for a variance of the Ech hydrogenase reaction in *M. barkeri*. Shown are the regions of experimentally observed growth yields for different substrates using a stoichiometry for the Ech hydrogenase reaction of 2.0 protons translocated/$2e^-$. Any stoichiometry > 2.0 protons translocated/$2e^-$ created an increasingly larger variability in the GAM values consistent with growth yields across all substrates (i.e., GAM values consistent with growth yields increased for simulated growth on acetate and decreased when H_2/CO_2 was the substrate).

5.3.6 Determination of the stoichiometry for the nitrogenase reaction in *M. barkeri*

M. barkeri was found to contain two different clusters of genes potentially encoding nitrogenases[35] and can fix molecular nitrogen at a lower yield per substrate than when supplied with ammonium[36]. The stoichiometry (efficiency of energy coupling) of the nitrogenase reaction in *M. barkeri* is unknown and we estimated this value using *i*AF692 and experimental data. The data used was growth data on methanol and either N_2 (which requires nitrogenase to convert N_2 to NH_4) or NH_4 (a control, since nitrogenase is not needed)[36]. Again, we were faced with a number of

unknown variables in our system (see previous analysis on Ech hydrogenase reaction stoichiometry and **§5.5**). For this analysis, the Ech hydrogenase translocation efficiency was approximated at 1 proton translocated/$2e^-$. For growth on methanol, the overall growth rate is less dependent on the translocation efficiency of Ech hydrogenase compared to acetate or H_2/CO_2 (at most ±10% for a one-fold change, see **§5.5**). The value of the GAM was determined from BOF optimization using the constraints from the control data (NH_4 as a nitrogen source). With all other variables defined for our system, the stoichiometry of the nitrogenase reaction was determined using BOF optimization constrained by the data from diazotrophic growth (growth using N_2 as a nitrogen source) and is given in the following balanced equation:

$$N_2 + 32\ ATP + 8\ ferredoxin_{reduced} + 32\ H_2O + \rightarrow$$

$$2\ NH_4 + 32\ ADP + 8\ ferredoxin_{oxidized} + 6\ H^+ + H_2 + 32\ Pi$$

This nitrogenase stoichiometry shows that energy coupling between ATP and the nitrogenase enzyme(s) was greater than the theoretical limit of ATP hydrolyzed per electron transferred[37]. This finding was not uncommon[37] and quantifies the energy needed for *M. barkeri* to grow diazotrophically.

This represents an improvement over the classical approximations used in Bomar and colleagues[36], in which overall growth yields and amounts of methanol assimilated were approximated using a standardized constant. *i*AF692 determines these values directly by computing the network flux through the physiological reactions available to the cell. The percentage of methanol assimilated for diazotrophic growth compared to growth on NH_4 indicates how much additional methanol is needed to generate ATP for fixing N_2 (43% and 56%, respectively).

These percentages were significantly different (6% and 14%, respectively) to those computed by Bomar and colleagues[36]. It is also worth mentioning that cells grown diazotrophically and those grown on NH_4 had similar nitrogen content (.069 and .066 gN/gDW, respectively,[36], which provides confidence in the use of the same BOF under both growth conditions.

5.3.7 Examination of a possible alternate pathway for the biosynthesis of H_4SPT

The possibility of an alternate pathway for the biosynthesis of H_4SPT in *M. barkeri* was proposed by Buchenau and Thauer[24] when they found that pABA was not required for H_4SPT synthesis. Unfortunately, the only pathway characterized for the synthesis of H_4SPT involved pABA and was characterized for a H_4SPT derivative in *M. jannaschii*[15,38]. This prevented a computational comparison of opposed pathways using *i*AF692 and experimental measurements. However, the model was used to identify a possible precursor metabolite, which may aid in the discovery of an alternate pathway[39]. In review of biochemical databases and metabolites present in *i*AF692, compounds were analyzed to determine their structural similarity to pABA (see **§5.5**). Chorismate was found to be the closest structurally related compound that could be synthesized from basic media substrates in simulations using *i*AF692. Chorismate can also be converted to pABA in some microbes[40], but this is unlikely in *M. barkeri* since this would contradict the finding that growth is dependent on either pABA or folic acid[24]. Another possible lead for an alternative pathway was found when a homology search indicated two genes in *M. barkeri* with strong sequence similarity to the gene which is responsible for the utilization of pABA in *M. jannaschii* for synthesis of the H_4SPT derivative[38]. A search of the *M. jannaschii* genome found

no probable paralogs to this gene, indicating that *M. barkeri* may have an additional metabolic capability similar to the function of the pABA utilizing gene. Characterization of the substrates for these probable enzymes in *M. barkeri* (gene5036 and gene4403) could lead to evidence supporting or refuting a possible alternate pathway.

5.3.8 Gene deletion analysis for the methanogenic pathways in *M. barkeri*

The essential genes and reactions in the methanogenic pathway needed for growth of *M. barkeri* on different methanogenic substrates were computationally determined (**Figure 5.7**). By using FBA with BOF optimization, it was possible to determine the active gene-encoded reactions and their flux values (**Figure 5.8**) that are essential for generating sufficient energy (or any at all) for growth under given substrate conditions (see **§5.5**). Each reaction in turn was deleted from the model, simulating a loss-of-function mutation of any single gene or group of genes associated with the reaction. Through interpretation of the computational results, it was possible to determine why certain mutation states fail to grow while others are still viable. The results were categorized into three different conditions: methanogenic growth, acetogenic growth, and no growth (see **§5.5** and **Figure 5.7**).

Simulation results were compared to experimental measurements on *M. barkeri* mutants (**Figure 5.7**). Three single reaction loss-of-function mutations in the methanogenic pathway (the *ech* operon, *mtr* operon and *mch* gene) have been generated for *M. barkeri* and characterized for growth on different substrates[41-43]. Also, Bock and Schonheit[44] characterized growth on pyruvate when the methyl coenzyme M reductase reaction was completely inhibited (similar to a loss-of-function mutation). The predictions made using *i*AF692 fully agree with the findings for a *M.*

barkeri mutant lacking the function of the *mtr* operon[43], the *mch* gene[42] or when the methyl coenzyme M reductase reaction was not available to the cell[44] (see Supplementary Data[14] for a discussion of selected substrate and genetic growth conditions).

Meuer and colleagues[41] characterized a mutant of *M. barkeri* lacking the functional *ech* operon encoded genes involved in methanogenesis. The predictions made using *i*AF692 agree with the experimental observations on single substrate cultures for the *ech* mutant (see **Figure 5.7**). Conversely, the model does not reproduce the finding that the *ech* mutant does not grow on methanol/H_2/CO_2. This is surprising in that the mutant would grow on methanol alone, but not with the addition of H_2/CO_2 to the medium. One possibility that has been proposed is that the *ech* mutant did not grow because of repression of the oxidative branch of methanogenesis (mh4spt to co2, see **Figure 5.8**) when H_2 was added to the medium[41]. Although an active oxidative pathway was determined to produce a higher growth rate, this pathway was not essential for simulated growth under these conditions. With only one false positive in this limited dataset, the reason for this disagreement will likely become evident once additional genetic mutants can be analyzed using *i*AF692. It is worthwhile to mention that FBA will never predict reduced growth with only the addition of substrates to medium (like the addition of H_2) unless the cell is forced to take them up.

Enzyme	Encoding Genes	Reaction Abbrev.	Substrate						
			methanol	acetate	H_2 / CO_2	methanol / acetate	acetate / H_2 / CO_2	methanol / H_2 / CO_2	pyruvate
acetate kinase	AckA	ACKr							
phosphotransacetylase	Pta	PTAr							
carbon monoxide dehydrogenase / acetyl-CoA synthase	CdhA, CdhB, CdhC, CdhD, CdhE, CooS	CODHr							
formylmethanofuran dehydrogenase (b)	FmdA, FmdB, FmdC, FmdD, FmdE, FmdF, FwdB, FwdD, FwdE, FwdG	FMFD(b)							
formylmethanofuran:H4SPT N-formyltransferase (b)	Ftr	FMFTSPFT(b)							
methenyl-H4SPt cyclohydrolase	Mch	MTSPC	⊕	⊕	⊕			⊕	
F420-dependent methenyl-h4spt dehydrogenase	Mtd	F4MTSPD							
F420-dependent methenyl-h4spt reductase	Mer	F4MTSPR							
methyl-H4SPT:coenzyme M methyltransferase	MtrA, MtrB, MtrC, MtrD, MtrE, MtrF, MtrG, MtrH	MTSPCMMT	⊕	⊕	⊕	⊕		⊕	
methanol:coenzyme M methyltransferase	MtaA, MtaB, MtaC	MCMMT							
methyl-coenzyme M reductase	McrA, McrB, McrG	MCR							⊕
heterodisulphide reductase	HdrA, HdrB, HdrC, HdrD, HdrE	HDR							
F420 deydrogenase	FpoA, FpoB, FpoC, FpoD, FpoF, FpoH, FpoI, FpoJ, FpoK, FpoL, FpoM, FpoN, FpoO	F4D							
F420-reducing hydrogenase	FrhA, FrhB, FrhD, FrhG	F4RHi							
Ech hydrogenase	EchA, EchB, EchC, EchD, EchE, EchF	ECHH	⊕	⊕	⊕			⊖	
F420-nonreducing hydrogenase	VhoA, VhoC, VhoG	F4NH							
ATP synthase	AhaA, AhaB, AhaC, AhaD, AhaE, AhaF, AhaH, AhaI, AhaK	ATPS4r							

Figure 5.7: Essential reactions and genes in the methanogenic pathway of *M. barkeri*. Listed are the enzymes of the methanogenic pathway, the protein encoding genes needed to produce the functional enzymes and the abbreviation for the reaction they perform in *M. barkeri*. Each of the reactions catalyzed by the enzyme listed was removed from the network and growth phenotypes were determined. For each computational prediction, green indicates methanogenic growth, blue indicates acetogenic growth, and red indicates no growth (0 net flux through the BOF). A plus symbol for a given condition indicates an agreement between model predictions and experimental characterization. A negative symbol indicates a disagreement. The colored enzyme and encoding gene sets (pink, yellow and green) indicate equal flux correlated reaction sets that possess the same reaction flux value under all growth conditions since they belong to a linear pathway. The simulation results can be used to determine the growth phenotypes of mutant strains and interpret the active pathways under each given condition.

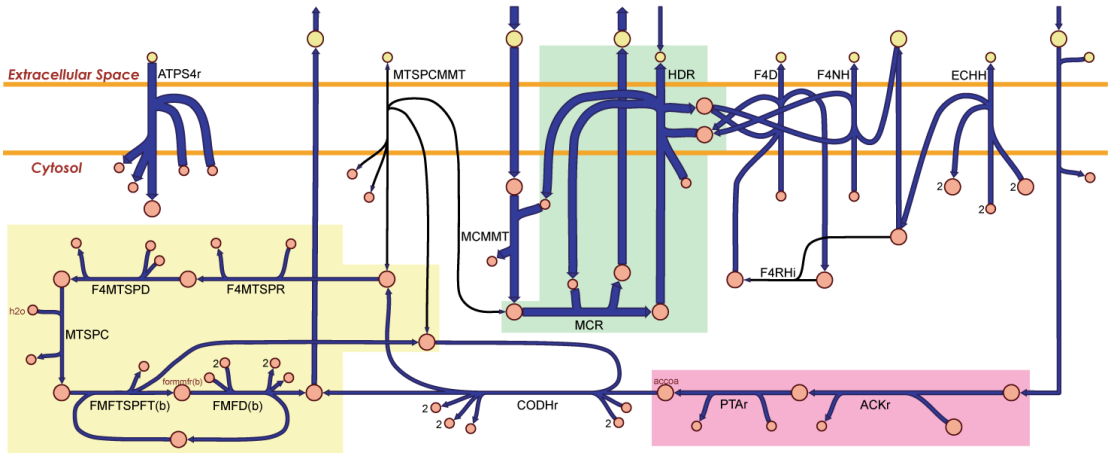


Figure 5.8: A flux map of the methanogenic pathway for growth of an mtr mutant on methanol and acetate. Shown is a reaction flux map of the methanogenic pathway reconstructed in *i*AF692 for growth of an *mtr* mutant of *M. barkeri*. The blue arrows indicate the direction of enzymatic activity and the arrow thicknesses are proportional to the flux through each reaction (a thicker arrow has a larger flux). The uptake of acetate is dependent on the uptake of methanol, which was constrained at 16 mmol gDW^{-1} hr^{-1}. The MTSPCMMT reaction encoded by the *mtr* operon is not available to the cell (because of the mutated state) and the F4RHi reaction is predicted not to be active for this optimal growth solution. All of the primary metabolites (large circles) are connected, except for ATP which participates in ATPS4r and ACKr, and the secondary metabolites (small circles) that appear more than once in the map are adp, pi, h, h2o, fdred, fdox, f420-2h2, f420-2, coa and com. The different colored regions correspond to the equal flux correlated reaction sets that were determined during the gene deletion analysis and are listed in Figure 5. All the reactions in each set will contain the same flux value for a given condition. The map was generated using the SimPheny™ software platform and used to visualize simulation results. The map does not display the stoichiometry for the ion translocating reactions (see **§5.3.5**). The reaction abbreviations are listed in Figure 5.7 and also in Supplementary Data[14] along with the metabolite abbreviations.

5.4 Conclusions

We have reconstructed the metabolic network of *M. barkeri* and analyzed the metabolic network using constraint-based analysis. Combined with strong sequence homology similarities, simulation data was used to interpret ambiguous results and assign probable functions for unknown genes. This method of 'functional annotation' will become increasingly useful for interpreting ambiguous homology searches as more enzymes and their genomic sequences are characterized.

The level of agreement between *i*AF692 predictions and experimental findings, especially for growth phenotypes, shows promise in the use of the model as a high-throughput analysis tool for studying growth of *M. barkeri*. The model can not only correctly predict growth phenotypes, but it can also determine i.) active reactions and pathways, ii.) the flux distribution in the network reactions, iii.) the redundancy or robustness of reactions in the network to a particular objective function, iv.) product formation and additional substrates for growth under a given state, such as predicted WT growth solely on cysteine and, v.) areas of disagreement between current knowledge (the model content) and experimental findings, as was seen when minimal media conditions were examined. The findings that *i*AF692 can accurately predict phenotypes on mixed substrate conditions are also of high interest because the pathways that are active during these conditions are poorly understood.

Although *i*AF692 is already a useful model, continual refinement and updating is necessary. If in reality the incorrect growth prediction of the *ech* mutant on methanol/H_2/CO_2 was caused by regulatory events, *i*AF692 will be instrumental in the interpretation of experimental data and characterization of metabolic regulation. With the ability to examine all aspects of metabolism, an iterative modeling process can

generate logical hypotheses and identify conditions (such as regulatory events) that would reconcile disagreements between experimental observations and simulation results. These hypotheses can then be further investigated. As the overall amount of data on *M. barkeri* increases (for instance, an updated annotation and specific growth maintenance values), *i*AF692 will continue to expand in its scope and accuracy to predict cellular phenotypes. In the future, *i*AF692 can serve as a starting point for future archaeal reconstructions and as an analysis platform for interpretation of experimental data and the study of methanogenesis.

5.5 Materials and methods

5.5.1 Network reconstruction

The reconstruction software SimPheny™, version 1.7.1.1 (Genomatica Inc., San Diego, CA), was the software platform in which the model was built. The ORF draft annotations for *M. barkeri* Fusaro, downloaded from the ORNL website (http://genome.ornl.gov/microbial/mbar/, Feb. 2004), were used as a framework to which translated metabolic proteins were assigned to form gene-protein-reaction (GPR) assignments. The draft genome consisted of 67 contigs of length 4.8 Mb and 5072 predicted candidate ORFs were examined. Most GPR assignments were made from the genome annotation and the model was constructed on a pathway basis manually. Biochemical databases such as KEGG (http://www.genome.jp/kegg/), the Enzyme Nomenclature Database (http://www.chem.qmul.ac.uk/iubmb/enzyme/) and the MetaCyc database (http://metacyc.org/) were used as general guides for pathways and sources for previous genome annotations. When a reaction was entered into the model, the participating metabolites were characterized according to

their chemical formula and charge determined for a cytosolic pH of 7.2, a value consistent with the intracellular range determined for methanogens[18,45,46]. Metabolite charge was determined using its pKa value. When the metabolite pKa was not available, charge was determined using the pKa of ionizable groups present in a metabolite. It should be mentioned that the charge of almost all metabolites in the network will not change for a pH increase or decrease of greater than ~1.5 pH units based on the pKa values of the ionizable groups (most frequently, carboxyl groups and amines, pKa ~4 and ~9, respectively). The BLAST algorithm[47] was implemented to infer gene function for enzymes needed to form complete pathways where no gene could be found in the annotation (see Supplementary Data[14] for detailed BLAST results). Operon structure was also considered when assigning function when multiple genes having identical annotations were found. GPR associations were also made directly from biochemical evidence presented in journal publications and reviews (see Supplementary Data[14]). The Pathway Tools software, version 8.5, (http://bioinformatics.ai.sri.com/ptools/), was used to generate an automated metabolic reconstruction and the pathways were analyzed and used to form or confirm GPR associations after manual inspection. Organism specificity of the reactions was achieved by including i.) the unique metabolites present in *M. barkeri*, such as H_4SPT[48], methanofuran-b[49], and coenzyme F_{420}[50], ii.) specific physiological cofactors, such as ADP for phosphofructokinase[51] and coenzyme F_{420} as an electron donor in glutamate synthase[50], iii.) the measured stoichiometric values for proton and ion translocation reactions in the electron transport chain of *M. barkeri*[30,31] and, iv.) the necessary metabolic transport reactions for substrates and products of metabolism. Transport reactions were added to the network from the genome annotation or alternatively from physiological data (these were added when a

metabolite was taken up into the cell or excreted into the media,[12,24,52]. All of the reactions entered into the network were both elementally and charged balanced and are either reversible or irreversible. Reversibility was determined first from primary literature if an enzyme was characterized and additionally from thermodynamic considerations, such as reactions that consume high energy metabolites (ATP, GTP, etc.) are generally irreversible.

ORFs in the draft genome annotation that were determined to be previously unannotated were genes assigned functionality in the model which contained the words "hypothetical" or similar. Genes that were deemed misannotated were determined to be assigned a function considerably different or more specific than what was given in the draft annotation.

5.5.2 Network comparison

For the comparison of model content, both *i*JR904 and *i*ND750 were decompartmentalized so that only the fundamental reaction and metabolite names were compared and not their location inside of the cells (transport reactions across the cytosolic membrane were compared). Reversibility was not considered in this comparison. Three reactions from the methanogenic pathway in *i*AF692 were changed to different high-level functional categories according to their role in *i*JR904 and *i*ND750 (PTA, ACK and ATPS).

For the comparison of network properties, currency metabolites (see Supplementary Data[14]) were removed from each network since they participate in several reactions and form links that do not represent real metabolic pathways[21] and model compartmentalization was conserved to maintain network structure. Irreversible and reversible links that appeared twice or more in a network were only

considered as one link. All of the network properties were calculated using the pajek software package[53].

5.5.3 Modeling simulations

A stoichiometric matrix, ***S*** ($m \times n$), was constructed for the *M. barkeri* metabolic network where m is the number of metabolites and n is the number of reactions. The corresponding entry in the stoichiometric matrix, S_{ij}, represents the stoichiometric coefficient for the participation of the i^{th} metabolite in the j^{th} reaction. FBA was then used to solve the linear programming problem under steady-state criteria[3,4,54]. The linear steady-state problem can be represented by the equation:

$\boldsymbol{S} \bullet \boldsymbol{v} = 0$ (1),

where ***v*** ($n \times 1$) is a vector of reaction fluxes. Since the linear problem is normally an underdetermined system for genome-scale metabolic models, there exists multiple solutions for ***v*** that satisfy Equation 1. To find a solution for ***v***, the cellular objective of producing the maximal amount of biomass constituents, represented by the ratio of metabolites in the BOF, is optimized for in the linear system. This is achieved by adding an additional column vector to ***S***, $\boldsymbol{S_{i,BOF}}$, containing the stoichiometric coefficients for the metabolites in the BOF and then subsequently maximizing the reaction flux through the corresponding element in ***v***, v_{BOF}, under the steady-state criteria. Additionally, constraints that are imposed on the system are in the form of:

$\alpha_i \leq v_i \leq \beta_i$ (2)

where α and β are the lower and upper limits placed on each reaction flux, v_i, respectively. For reversible reactions, $-\infty \leq v_i \leq \infty$, and for irreversible reactions, $0 \leq v_i \leq \infty$.

The constraints on the reactions that allow metabolites entry to the extracellular space were set to $0 \leq v_i \leq \infty$ if the metabolite was not present in the medium, meaning that the compounds could leave, but not enter the system. For the metabolites that were in the medium, the constraints were set to $-\infty \leq v_i \leq \infty$ for all except the limiting substrate and cysteine. When cysteine was a media component, it was allowed only for use as a source of sulfur by restricting hydrogen sulfide from exiting the system. Artificial transhydrogenase cycles in the network[1] were avoided by only allowing the net flow through a set of potential NAD(H)/NADP(H) cycling reactions in one direction. The reaction flux through the BOF was constrained from $0 \leq v_{BOF} \leq \infty$ and the BOF was generated as a linear equation consisting of the molar amounts of metabolic constituents that make up the dry weight content of the cell (**Table 5.3**) and a GAM (mmol ATP gDW^{-1}) reaction to account for non-metabolic growth activity,

$ATP + H_2O \rightarrow ADP + Pi + H^+$ (3).

The full BOF is included in Supplementary Data[14].

Aside from the BOF, a NGAM (mmol ATP gDW^{-1} hr^{-1}) value was used as an energy "drain" on the system during the linear programming calculations and accounts for non-growth cellular activities[55]. The NGAM was represented as a set flux in the reaction flux vector, v_{NGAM}. The corresponding reaction vector in the stoichiometric matrix, $\boldsymbol{S}_{i,NGAM}$, was in the form of an ATP maintenance reaction identical to Equation 3.

Linear programming calculations were performed using the previously mentioned SimPheny™ software platform and the MATLAB®, version 7.0.0.19920,

(The MathWorks Inc., Natick, MA) software platform in which the linear programming package LINDO (Lindo Systems Inc., Chicago, IL) was used as a solver.

5.5.4 Gap filling and determination of minimal media

To fill gaps in the network, biomass components were sequentially added to the BOF individually and FBA was used for BOF optimization under steady-state criteria. When a simulation resulted in a positive net flux through the BOF, a subsequent component was added to the BOF and the simulation was rerun. When a biomass component added to the BOF resulted in no flux through the BOF, the network was manually updated. This process was continued until all of the biomass constituents in **Table 5.3** were included, and FBA optimization produced a positive flux in the BOF. The initial gap filling procedure was performed with common media conditions[25] for all of the three major substrates: methanol, acetate and H_2/CO_2. Subsequent procedures were performed removing common media substrates to identify additional gaps that may have been overseen by using the complex common media.

After the gap filling procedure, nonessential compounds were individually removed from the common media conditions until a minimal set of compounds remained that produced a nonzero positive flux through the BOF for a simulation. When two compounds were found to fulfill the same metabolic requirement, the lesser carbon-containing compound was taken as a minimal component. All of the minimal media metabolites were manually analyzed to determine their necessity in producing the BOF constituents. An approximate value of 50 mmol ATP gDW^{-1} was used for the GAM and the NGAM was not considered for the gap filling procedure and determination of minimal media.

5.5.5 Estimation of the proton translocation efficiency of the Ech hydrogenase reaction

The proton translocation stoichiometry of the Ech hydrogenase reaction was varied in ***S*** for different simulations. For each ***S*** generated, there were two unknown variables needed for BOF optimization using FBA: the GAM and the NGAM. For growth simulations using FBA and BOF optimization on microbial cells, the NGAM has ranged from 1.0 - 7.6 mmol ATP gDW^{-1} hr^{-1} when calculated from experimental data[54,56,57]. It was shown that the NGAM and GAM will vary proportionally[57]. Therefore, the NGAM was estimated as 2.5% of the GAM since it produced values similar to those from the previous studies. Since no data were available to directly calculate the GAM in this analysis, a wide range of GAM values were considered. Minimal media conditions were used with the addition of cysteine aside from the main growth substrate. The main growth substrate uptake rates were set as the maximum uptake rates for each growth substrate and were taken directly or approximated using biomass yields (gDW mol^{-1} substrate) and the maximal specific growth rates (hr^{-1}) from the cited studies: 16 mmol methanol gDW^{-1} hr^{-1} and 8 mmol acetate gDW^{-1} hr^{-1}[58], 41 mmol H_2 gDW^{-1} hr^{-1}[59] (the amount of CO_2 was not constrained) and 5 mmol pyruvate gDW^{-1} hr^{-1}[12]. The maximal growth yields (gDW $mmol^{-1}$ CH_4) determined using linear optimization were calculated from the BOF reaction flux (hr^{-1}) and the reaction flux of methane (mmol CH_4 gDW^{-1} hr^{-1}) leaving the system. The substrate yields and molar yields of CO_2 and CH_4 were determined with a similar method. The observed growth yields that were used for comparison to simulation results were compiled from growth studies of *M. barkeri* on methanol (5 - 7 gDW mol^{-1} CH_4)[20,59,60], acetate (2 - 4 gDW mol^{-1} CH_4)[12,59], H_2/CO_2 (80:20 v/v) (6 - 9 gDW mol^{-1} CH_4)[59,60] and pyruvate (12 -15 gDW mol^{-1} CH_4)[12]. Data from growth on different substrates were

used since the Ech hydrogenase reaction operates in different directions depending on the substrate(s)[41]. The cost of the polymerization of the macromolecules was calculated using the cellular composition of *M. barkeri* (**Table 5.3**) and the common polymerization and processing costs for a typical prokaryotic cell[28]. Supplementary Data[14] contain the reaction flux distributions in *i*AF692 for optimized growth with the primary substrate of methanol, acetate, H_2/CO_2 and pyruvate, respectively. The flux distributions were generated using minimal media conditions with the addition of cysteine, the primary substrate uptake rates listed above, a GAM of 70 mmol ATP gDW^{-1}, a NGAM of 1.75 mmol ATP gDW^{-1} hr^{-1} and a stoichiometry of 1 proton translocated/$2e^-$ for the Ech hydrogenase reaction. These distributions are also provided as a Microsoft Excel worksheet in Supplementary Data[14].

5.5.6 Determination of the stoichiometry for the nitrogenase reaction in *M. barkeri*

The stoichiometry of the nitrogenase reaction was determined using the growth rates, total amounts of methanol consumed and growth yields for growth on NH_4 and diazotrophic growth[36]. Minimal media conditions were used in addition to methanol as the main substrate[36]. The value of 1 proton translocated/$2e^-$ was used for the Ech hydrogenase reaction stoichiometry since this was the approximate stoichiometry for a similar hydrogenase[33]. Using the NH_4 growth conditions, the stoichiometry of the Ech hydrogenase reaction was found to affect the predicted growth rate, at most, ±10% for a one-fold change above or below this value. The GAM used in the BOF was calculated by finding the value that produced the observed growth rate from the given methanol uptake rate[36] using BOF optimization when NH_4 was the nitrogen source in the simulation (the nitrogenase reaction was not

needed under these conditions). The value determined was 30 mmol ATP gDW^{-1}, the NGAM was set to 2.5% of this value and these values were used in all subsequent simulations. The stoichiometry of the nitrogenase reaction was then determined by finding the value of *p* in Equation 4 that produced the observed growth rate from the given methanol uptake rate[36] using BOF optimization when N_2 was the nitrogen source in the simulation (the nitrogenase reaction was needed under these conditions). Equation 4 is the balanced overall enzymatic reaction for nitrogenase presented by Rees and Howard[37]:

$$N_2 + 8p\ ATP + 8\ ferredoxin_{reduced} + 8p\ H_2O + \rightarrow$$

$$2\ NH_4 + 8p\ ADP + 8\ ferredoxin_{oxidized} + 6\ H^+ + H_2 + 8p\ Pi \quad (4)$$

5.5.7 Examination of a possible alternate pathway for the biosynthesis of H_4SPT

When searching for an alternate pathway for the synthesis of H_4SPT, gene MJ1427[38] was used as a query sequence for BLAST (gene5036 and gene4403 had e-values of 2e-63 and 4e-62, respectively). Structural similarity of metabolites was determined by finding compounds that could be converted to pABA in the smallest number of known enzymatic steps.

5.5.8 Gene essentiality

To determine the effect of a gene deletion, the reaction associated with each gene in the methanogenic pathway was individually deleted from ***S*** and FBA was used to predict the mutation growth phenotype. Using the same maximum uptake rates for each growth substrate as indicated in the examination of Ech hydrogenase reaction and minimal media conditions, the flux through the BOF was optimized in the

mutated network, ***S′***, for each substrate. The criteria used to determine growth *in silico* was a positive flux through the BOF ($v_{BOF} > 0$) using ***S′*** under steady-state (Equation 1). When $v_{BOF} > 0$ for a given ***S′***, a positive growth phenotype was categorized under two different types of growth: methanogenic growth or acetogenic growth. For methanogenic growth, methane was formed as a major end product (green boxes, **Figure 5.7**) and for acetogenic growth, acetate was formed as a major end product (blue boxes, **Figure 5.7**). When an optimization of the BOF resulted in $v_{BOF} = 0$, this was determined to be a no-growth phenotype (red boxes, **Figure 5.7**). All of the values of v_{BOF} for the positive mutation growth phenotypes were greater than 10% of the WT value under the same substrate conditions. The GAM was varied from 30 – 110 mmol ATP gDW^{-1} and the NGAM was 2.5% of the GAM for the gene deletion study. Growth phenotypes were also determined with a variable stoichiometry of 0.2 – 2.0 protons translocated/$2e^-$ for the Ech hydrogenase reaction. The GAM, NGAM or choice of Ech hydrogenase reaction stoichiometry did not have an effect on the predicted growth phenotypes in the range considered. Acetogenic growth was reported if it was the only means for growth and was possible under all Ech hydrogenase reaction stoichiometries. The gene deletion analysis was performed using the SimPheny™ software platform.

Acknowledgements

We would like to thank Jennifer Reed, Thuy Vo, Natalie Duarte, Sharon Wiback, Iman Famili, Radhakrishnan Mahadevan and Chris Workman for their invaluable insight in preparation of this chapter.

Chapter 5, in full, is adapted from an article that appeared in Nature Molecular Systems Biology, volume 2, number 2006.0004, pages 1-14, published January,

2006. The dissertation author was the primary author of this paper, which was co-authored by Dr. Johannes C. M. Scholten, Dr. Bernhard Ø. Palsson, Dr. Fred J. Brockman and Dr. Trey Ideker.

References

1. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

2. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429:92-6.

3. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

4. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;14:491-6.

5. Duarte NC, Herrgard MJ, Palsson B. Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. *Genome Res* 2004;14:1298-309.

6. Yeh I, Hanekamp T, Tsoka S, Karp PD, Altman RB. Computational analysis of Plasmodium falciparum metabolism: organizing genomic information to facilitate drug discovery. *Genome Res* 2004;14:917-24.

7. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003;84:647-57.

8. Tsoka S, Simon D, Ouzounis CA. Automated metabolic reconstruction for Methanococcus jannaschii. *Archaea* 2004;1:223-9.

9. Zinder SH. Physiological ecology of methanogens. In: Ferry JG, ed. Methanogenesis: Ecology, Physiology, Biochemistry and Genetics. London: Chapman & Hall, 1993:128-206.

10. Garcia JL, Patel BK, Ollivier B. Taxonomic, phylogenetic, and ecological diversity of methanogenic archaea. *Anaerobe* 2000;6:205-26.

11. Schink B. Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* 1997;61:262-80.

12. Bock A, Prieger-Kraft A, Schoenheit P. Pyruvate—a novel substrate for growth and methane formation in Methanosarcina barkeri. *Arch. Microbiol* 1994;161:33-46.

13. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology* 2003;4:R54.1-R54.12.

14. Feist AM, Scholten JCM, Palsson BO, Brockman FJ, Ideker T. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol* 2006;2:1-14.

15. Graham DE, White RH. Elucidation of methanogenic coenzyme biosyntheses: from spectroscopy to genomics. *Nat Prod Rep* 2002;19:133-47.

16. Peregrin-Alvarez JM, Tsoka S, Ouzounis CA. The phylogenetic extent of metabolic enzymes and pathways. *Genome Res* 2003;13:422-7.

17. Graham DE, Xu H, White RH. Identification of the 7,8-didemethyl-8-hydroxy-5-deazariboflavin synthase required for coenzyme F(420) biosynthesis. *Arch Microbiol* 2003;180:455-64.

18. de Poorter LM, Geerts WJ, Keltjens JT. Hydrogen concentrations in methane-forming cells probed by the ratios of reduced and oxidized coenzyme F420. *Microbiology* 2005;151:1697-705.

19. Nishihara M, Koga Y. Two new phospholipids, hydroxyarchaetidylglycerol and hydroxyarchaetidylethanolamine, from the Archaea Methanosarcina barkeri. *Biochim Biophys Acta* 1995;1254:155-60.

20. Hippe H, Caspari D, Fiebig K, Gottschalk G. Utilization of trimethylamine and other N-methyl compounds for growth and methane formation by Methanosarcina barkeri. *Proc Natl Acad Sci U S A* 1979;76:494-8.

21. Ma HW, Zeng AP. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 2003;19:1423-30.

22. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;6:397-9.

23. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000;407:651-4.

24. Buchenau B, Thauer RK. Tetrahydrofolate-specific enzymes in Methanosarcina barkeri and growth dependence of this methanogenic archaeon on folic acid or p-aminobenzoic acid. *Arch Microbiol* 2004;182:313-25.

25. Wolin EA, Wolin MJ, Wolfe RS. Formation of Methane by Bacterial Extracts. *J Biol Chem* 1963;238:2882-6.

26. Scherer P, Sahm H. Effect of trace elements and vitamins on the growth of Methanosarcina barkeri. *Acta biotechnologica.* 1981;1:57-65.

27. Fischer M, Schott AK, Romisch W, Ramsperger A, Augustin M, Fidler A, Bacher A, Richter G, Huber R, Eisenreich W. Evolution of vitamin B2 biosynthesis. A novel class of riboflavin synthase in Archaea. *J Mol Biol* 2004;343:267-78.

28. Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the bacterial cell: a molecular approach. Sunderland, Mass.: Sinauer Associates, 1990:xii, 506.

29. Thauer RK, Jungermann K, Decker K. Energy conservation in chemotrophic anaerobic bacteria. *Bacteriol Rev.* 1977;41:100-80.

30. Deppenmeier U. The membrane-bound electron transport system of Methanosarcina species. *J Bioenerg Biomembr* 2004;36:55-64.

31. Muller V. An exceptional variability in the motor of archael A1A0 ATPases: from multimeric to monomeric rotors comprising 6-13 ion binding sites. *J Bioenerg Biomembr* 2004;36:115-25.

32. Hedderich R. Energy-converting [NiFe] hydrogenases from archaea and extremophiles: ancestors of complex I. *J Bioenerg Biomembr* 2004;36:65-75.

33. Hakobyan M, Sargsyan H, Bagramyan K. Proton translocation coupled to formate oxidation in anaerobically grown fermenting *Escherichia coli*. *Biophys Chem* 2005;115:55-61.

34. Russell JB, Cook GM. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* 1995;59:48-62.

35. Chien YT, Auerbuch V, Brabban AD, Zinder SH. Analysis of genes encoding an alternative nitrogenase in the archaeon Methanosarcina barkeri 227. *J Bacteriol* 2000;182:3247-53.

36. Bomar M, Knoll K, Widdel F. Fixation of molecular nitrogen by Methanosarcina barkeri. *Fems Microbiol. Ecol.* 1985;31:47-55.

37. Rees DC, Howard JB. Nitrogenase: standing at the crossroads. *Curr Opin Chem Biol* 2000;4:559-66.

38. Scott JW, Rasche ME. Purification, overproduction, and partial characterization of beta-RFAP synthase, a key enzyme in the methanopterin biosynthesis pathway. *J Bacteriol* 2002;184:4442-8.

39. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Science* 2004;59:5051-5060.

40. Nichols BP, Seibold AM, Doktor SZ. para-aminobenzoate synthesis from chorismate occurs in two steps. *J Biol Chem* 1989;264:8597-601.

41. Meuer J, Kuettner HC, Zhang JK, Hedderich R, Metcalf WW. Genetic analysis of the archaeon Methanosarcina barkeri Fusaro reveals a central role for Ech

hydrogenase and ferredoxin in methanogenesis and carbon fixation. *Proc Natl Acad Sci U S A* 2002;99:5632-7.

42. Guss AM, Mukhopadhyay B, Zhang JK, Metcalf WW. Genetic analysis of mch mutants in two Methanosarcina species demonstrates multiple roles for the methanopterin-dependent C-1 oxidation/reduction pathway and differences in H metabolism between closely related species. *Mol Microbiol* 2005;55:1671-80.

43. Welander PV, Metcalf WW. Loss of the mtr operon in Methanosarcina blocks growth on methanol, but not methanogenesis, and reveals an unknown methanogenic pathway. *Proc Natl Acad Sci U S A* 2005;102:10664-9.

44. Bock AK, Schonheit P. Growth of Methanosarcina barkeri (Fusaro) under nonmethanogenic conditions by the fermentation of pyruvate to acetate: ATP synthesis via the mechanism of substrate level phosphorylation. *J Bacteriol* 1995;177:2002-7.

45. de Poorter LM, Geerts WG, Theuvenet AP, Keltjens JT. Bioenergetics of the formyl-methanofuran dehydrogenase and heterodisulfide reductase reactions in Methanothermobacter thermautotrophicus. *Eur J Biochem* 2003;270:66-75.

46. von Felten P, Bachofen R. Continuous monitoring of the cytoplasmic pH in Methanobacterium thermoautotrophicum using the intracellular factor F(420) as indicator. *Microbiology* 2000;146 Pt 12:3245-50.

47. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 1997;25:3389-402.

48. Grahame DA, DeMoll E. Partial reactions catalyzed by protein components of the acetyl-CoA decarbonylase synthase enzyme complex from Methanosarcina barkeri. *J Biol Chem* 1996;271:8352-8.

49. Bobik TA, Donnelly MI, Rinehart KL, Jr., Wolfe RS. Structure of a methanofuran derivative found in cell extracts of Methanosarcina barkeri. *Arch Biochem Biophys* 1987;254:430-6.

50. Raemakers-Franken PC, Brand RJ, Kortstee AJ, Van der Drift C, Vogels GD. Ammonia assimilation and glutamate incorporation in coenzyme F420 derivatives of Methanosarcina barkeri. *Antonie Van Leeuwenhoek* 1991;59:243-8.

51. Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM, Van Der Oost J. The unique features of glycolytic pathways in Archaea. *Biochem J* 2003;375:231-46.

52. Krzycki JA, Lehman LJ, Zeikus JG. Acetate catabolism by Methanosarcina barkeri: evidence for involvement of carbon monoxide dehydrogenase, methyl coenzyme M, and methylreductase. *J Bacteriol* 1985;163:1000-6.

53. Batagelj V, Mrvar A. Pajek: Program for large network analysis. *Connections* 1998;21:47-57.

54. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* 1994;60:3724-3731.

55. Pirt SJ. The maintenance energy of bacteria in growing cultures. *Proc R Soc Lond B Biol Sci* 1965;163:224-31.

56. Famili I, Forster J, Nielsen J, Palsson BO. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci U S A* 2003;100:13134-9.

57. Borodina I, Krabben P, Nielsen J. Genome-scale analysis of Streptomyces coelicolor A3(2) metabolism. *Genome Res* 2005;15:820-9.

58. Elferink O, H. SJW, Visser A, Hulshoff Pol LW, Stams AJM. Sulfate reduction in methanogenic bioreactors. *FEMS Microbiol. Rev.* 1994;15:119-136.

59. Smith MR, Mah RA. Growth and methanogenesis by Methanosarcina strain 227 on acetate and methanol. *Appl Environ Microbiol* 1978;36:870-9.

60. Weimer PJ, Zeikus JG. One carbon metabolism in methanogenic bacteria. Cellular characterization and growth of Methanosarcina barkeri. *Arch Microbiol* 1978;119:49-57.

Chapter 6

Model-driven metabolic engineering, Part 1: A computational evaluation of the production potential for growth-coupled products of *Escherichia coli*

6.1 Abstract

Integrated approaches utilizing *in silico* analyses will be necessary to successfully advance the field of metabolic engineering. Here, we present an integrated approach through a systematic model-driven evaluation of the production potential for the bacterial production organism *E. coli* to produce multiple native products from different representative feedstocks through coupling metabolite production to growth rate. In the analysis, designs were examined for eleven unique central metabolism and amino acid targets from three different substrates under aerobic and anaerobic conditions. Optimal strain designs were reported for each growth condition (where found) for designs which possess maximum yield, substrate-specific productivity, and strength of growth-coupling for up to ten reaction eliminations (i.e., knockouts). In total, growth-coupled designs could be identified for 36 out of the total 54 conditions tested corresponding to eight out of the eleven targets. There were 17 different substrate / target pairs for which over 80% of the theoretical maximum potential could be achieved; designs could be identified for each of the substrates examined. The developed method utilizes two different strain design

algorithms, OptKnock[1] and OptGene[2], the genome-scale metabolic reconstruction of *E. coli i*AF1260[3], and introduces a new concept of objective function tilting for strain design. Additionally, theoretical maximum production potential for each of the substrate / target pairs is reported based on native *E. coli* reactions and minimum growth requirements. Herein, this study provides specific metabolic interventions (i.e., strain designs) for production strains that can be experimentally implemented, characterizes the production potential for *E. coli* to produce native compounds, and outlines a strain design pipeline that can be utilized to design production strains for additional organisms.

6.2 Introduction

Metabolic engineering has been successful in generating biological strains for the production of compounds for a variety of purposes[4-8]. The earliest approaches to engineer microorganisms have consisted of mutating strains with a known mutagen and selecting for a strain with the desired phenotype. This process is largely dependent upon encountering a desired mutant after the mutagenesis and selection process. Over the past decade, new tools and strategies for engineering microorganisms have appeared including loss-of-function gene mutations (i.e., gene knock-outs)[9-14], overexpression of gene products[9,10,15,16], and introduction of homologous DNA for new or improved functionalities of cells. Today, an increasing amount of success has been achieved by rationally designing strains using these readily available methods[17-19]. However, it is difficult to identify an optimal strain by rational design alone, as several parameters that should be considered simultaneously in designs are difficult to predict without a computational approach.

Fueled by the ability to sequence and annotate genomes and the development of techniques such as flux balance analysis, systems biology can now play a role in the metabolic engineering process by guiding interventions to divert metabolite flux within a microbial cell. Genome-scale metabolic[3,20-26] models can be used as query platforms to examine new strategies and interventions as they contain parts lists for content in cells. Systems biology has been shown to be successful in predicting the outcomes (e.g., products, growth rates, etc.) of cellular growth utilizing a constraint based reconstruction and analysis (COBRA) approach[27]. Furthermore, there now exist several examples of model-driven metabolic engineering success stories for the production of various products[28-32].

The coupling of bacterial growth and target molecule production is an important feature to select for in strain design for several reasons. First, a strain in which target production is growth coupled to biomass production must produce this target, in order to produce biomass components. Therefore, in order for the cell to achieve faster growth, the target molecule must be secreted as well. This allows for evolving the strain towards higher target production rates by coupling it to the natural selection of the population towards faster growth. In this way, strains, or mutants, that produce the target molecule can be easily selected by selecting for the fastest growth through adaptive evolution and serial passage. Serial passage results in optimization of the strain for target production as well as growth rate, both desirable traits. This optimized strain can also be considered stable, as mutations that would result in lower target production would cause a decreased growth rate, and therefore would be outcompeted by the optimized strain in future passes. Adaptive evolution has been used for the experimental analysis of evolution of both wild type and knockout strains with success[33-36] therefore providing confidence that the knockout strains described

herein will be able to undergo the same adaptation when experimentally implemented.

In conjunction with COBRA methods and the development of organism-specific models, the experimental approach of adaptive evolution has been proven effective for the selection of strains that possess optimal growth phenotypes[33-36]. We now can combine these two approaches and use them as design principles to obtain production strains that agree with computationally predicted phenotypes. Initial computational efforts towards this goal have been performed[1,2,37,38] and an initial study has been experimentally verified for a selected case[33], however a rigorous computation of potential has not appeared. Therefore, we performed a large-scale computational study of growth-coupled production potential of a wide array of industrially relevant targets under several different substrate conditions, utilizing the OptKnock and OptGene algorithms. We then identified the best designs for each target/substrate pair under three desirable criteria: yield, substrate specific productivity (SSP), and strength of growth coupling (SOC). We also identify trends and common structural characteristics that enable certain targets to be growth-coupled with production, as well as genes that are commonly associated with growth-coupled designs under each substrate condition. The methods outlined in this paper, as well as the strain designs themselves, serve as fundamental approaches to and results of applying systems biology methods to metabolic engineering.

6.3 Results

6.3.1 Selection of substrate and products for analysis

Although *E. coli* can utilize a number of individual substrates for growth, we determined three primary *E. coli* substrates for this analysis that will be practical in terms of cost and availability, and possess unique design potential (e.g., glucose and fructose are not unique and are examples of interconverted substrates with little to no cost to the cell). These substrates are glucose, xylose, and glycerol. Hexoses (represented by glucose) and pentoses (represented by xylose) will be the substrates available from plant material and are expected to become more prevalently available with the push towards biofuels and bio-based products[39]. Glycerol is becoming more prevalently available as it is a byproduct of biodiesel production[40]. Additional substrates, such as 4-carbon substrates, can be added as design substrates seamlessly in future analyses as this computational platform is easily updatable. Additionally, both anaerobic and aerobic conditions were analyzed for this study to examine the range of products that can be coupled to growth under these two different respiratory conditions.

The products included in the study were chosen by analyzing metabolites in the network that possess either commercial value or are representative molecules from key points in metabolism. Compounds that contained commercial value were determined by evaluating, i.) large-scale reports released by federal agencies, such as the U.S. Department of Energy, which identify key platform chemicals[41,42], ii.) products that are currently under production on a large-scale (such as organic[43] and amino acids[44]), and ii.) metabolites that have had high profile attention in the literature[9]. The final list of metabolites examined as targets is listed in **Table 6.1**.

Table 6.1: Theoretical maximum production analysis.

	Substrate	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol
	Aerobicity	Anaerobic	Anaerobic	Anaerobic	Aerobic	Aerobic	Aerobic
product	no. of carbons	$Y_{p/s}$	$Y_{p/s}$	$Y_{p/s}$	$Y_{p/s}$	$Y_{p/s}$	$Y_{p/s}$
Ethanol	2	49%*	49%*	49%	49%	49%	54%
D-Lactate	3	95%*	95%*	13%	95%	95%	97%
Glycerol	3	37%	27%		75%	74%	
L-Alanine	3	95%*	76%	13%	95%	93%	96%
L-Serine	3	47%	35%	6%	115%	114%	116%
Pyruvate	3	71%	60%	6%	100%	99%	100%
Fumarate	4	54%	40%	5%	110%	108%	117%
L-Malate	4	63%	46%	5%	127%	125%	135%
Succinate	4	93%	81%	12%	104%	101%	111%
2-Oxoglutarate	5	40%	32%	3%	98%	96%	101%
L-Glutamate	5	44%	36%	3%	92%	90%	97%

* indicates anaerobic condition where homofermentation of product is possible (< 2% wt% other carbon products, CO_2 exempt), all aerobic conditions except hydrogen have homofermentation potential

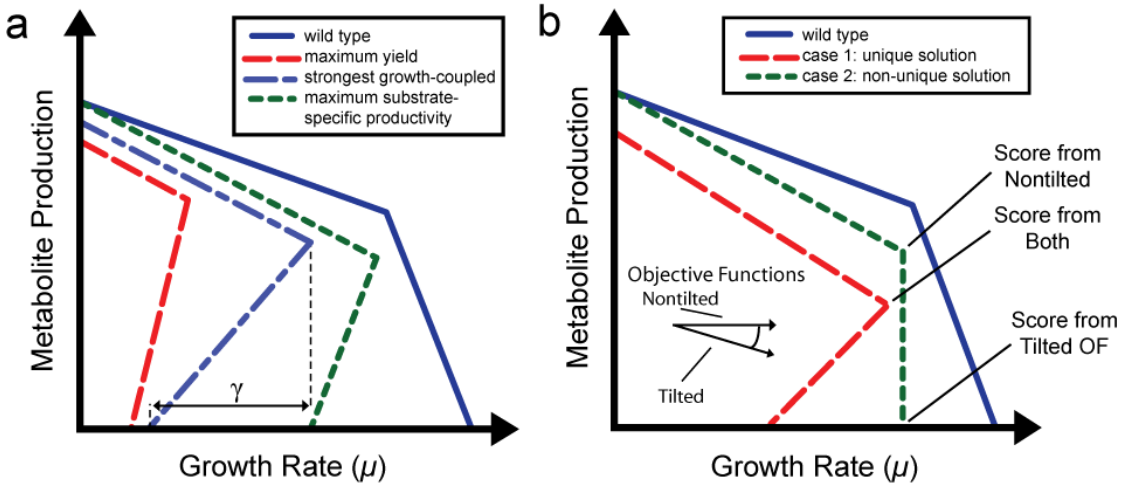


Figure 6.1: Strain Design Selection: Secondary objective criteria. Graphs showing (**a**) the production envelopes of the different secondary objective functions examined for designing strains, and (**b**) the different types of production envelopes encountered during this analysis. Also shown is a schematic of the direction of optimization for the 'tilted' and 'non-tilted' objective functions used in the analyses and points on the production envelopes each will score.

6.3.2 Theoretical analysis of the production potential in *E. coli*

To evaluate the efficiency of the strain design process and to identify target metabolites, an initial theoretical analysis was performed to define the maximum production potential in *E. coli* for the selected targets. These potentials were based from the content in the *i*AF1260 model[3], a comprehensive parts list of the cell detailing the available content for cellular transformation of the substrates. This analysis was performed by: i.) setting an uptake rate of 20 mmol gDW^{-1} hr^{-1} of each main carbon substrates and 20 mmol gDW^{-1} hr^{-1} O_2 when specified; values near experimentally measured maximum uptake rates[45], ii.) a minimal growth rate (μ) that *E. coli* must achieve to sustain growth, 0.1 hr^{-1}, and iii.) maximizing the flux through each of the exchange reactions in the model for the targeted products. The results from this analysis are given in **Table 6.1**.

From this analysis, a number of conclusions can be drawn to both understand the theoretical potential for production with *E. coli* and validate the analysis. The results are presented for anaerobic and aerobic conditions.

Anaerobic conditions: The maximum weight yield for each target (product) is found for glucose and, as expected, is successively less (or the same as) for xylose and glycerol, respectively. This is due to a higher percentage of the incoming carbon to the cell being necessary to account for the biomass production required to achieve the set minimum growth rate (0.1 hr^{-1}) for xylose and glycerol. The most obvious conclusion from analyzing anaerobic conditions is that production from glycerol anaerobically of any metabolite besides ethanol has a very poor potential yield (< 15% in each case). This result is in agreement with previous studies[46,47] . The trends for theoretical yield between glucose and xylose are very similar with the yields being,

on average, 10% higher on glucose. The compound L-alanine had a much higher than average yield on glucose over xylose as a substrate, whereas those for ethanol and lactate are approximately the same. For both glucose and xylose, lactate, L-alanine, and succinate showed the highest potential maximum yields. Alternatively, glycerol, 2-oxoglutarate, L-glutamate, and L-serine were predicted to have the lowest theoretical production potential.

Aerobic conditions: Aerobically, the theoretical yields are predicted to be highest on glycerol with the yields being similar from glucose and xylose. Different from anaerobic conditions, a number of products can potentially be made in higher titers on glycerol aerobically. The reason that many of the theoretical yields are over 100% is in part due to the fact that the phosphoenolpyruvate carboxylase reaction can carry a high flux in these theoretical calculations and fix carbon dioxide that is incorporated into many of the products. Overall, the average yields were 2% higher on glucose than on xylose and approximately 5% higher on glycerol than glucose or xylose. Looking at specific products, L-malate, L-serine, and fumarate showed the highest aerobic production potential, while ethanol and glycerol showed relatively poor theoretical yields.

Comparison of anaerobic and aerobic conditions: For each of the three substrates, the maximum theoretical yield increased, or stayed the same, in aerobic conditions. This is expected as in performing FBA with metabolic models, when adding more inputs to the system, the solution space can only expand. What is most interesting are the cases that stayed the same when allowing oxygen as an input, these are for ethanol and lactate production on glucose and xylose and for L-alanine production on glucose (all less than 0.1% increase). The average increases in

potential yield between the two conditions were 31%, 39%, and 83% for glucose, xylose, and glycerol, respectively. The higher yields for glycerol are due to the fact that the amount of oxygen incorporated into the product is greater per weight basis for glycerol than for glucose or xylose. Target products that were predicted to have the greatest increase in production potential aerobically included L-serine, L-malate, 2-oxoglutarate, fumarate, and L-glutamate.

Homofermenting strain designs: Homofermenting strains are those that produce a single product from fermentation of the main substrate. These strains are desirable in that they are not only likely to possess a high yield, but also offer an advantage as they reduce the necessity to separate different products after fermentation. The strains that have the potential to be homofermentors under anaerobic conditions are indicated in **Table 6.1**. From this analysis, all strain designs were predicted to be homofermentors aerobically. For the analysis, homofermenting strains were categorized as those that were predicted to have less than 2% wt% carbon-containing fermentation products besides the main product. Carbon dioxide was not considered as a byproduct in the analysis as it does not require a separation step due to its low solubility and its ability to escape as a gas fermentation conditions. For anaerobic fermentation, succinate an essential byproduct (at a very small amount, less than 1 wt% in most cases) when using the *i*AF1260 reconstruction in modeling simulations. Therefore, it will be a byproduct of all anaerobic condition predictions. Out of the 32 different theoretical maximum predictions anaerobically, five designs were predicted to possess the potential for homofermentation. This included three different compounds, and two different substrate conditions.

These results, in themselves, define the targets and conditions in which strain designs for which modification of wild-type *E. coli* content have the greatest potential. **Appendix A** contains additional results for this analysis in terms of molar yields. These results were used for comparison with yields achieved for growth-coupled stain designs for *E. coli*. As production of any compound besides ethanol on glycerol anaerobically resulted in a low theoretical yield, it was not included in the analysis. Furthermore, simulations for the growth-coupled production of compounds from glycerol aerobically were only examined for a higher number of allowable network knockouts. This was due to the assumption that growth-coupling of compounds to glycerol would require more metabolic interventions to produce compounds to this low carbon substrate.

6.3.3 Strain Design: Model pre-processing and selection of target reactions for elimination

In order to both decrease computation time and to limit the elimination of reactions in *E. coli* to those that are biologically relevant knock-out targets for diverting metabolite flux, the model was preprocessed as outlined in (see **Figure 6.2**). In short, these engineering assumptions were based on biologically relevant and practical principals (steps b-d) and computational approaches (steps a and f). An example of a biologically relevant reduction of targets for removal was that reactions that spontaneously occur inside of the cell were not considered as they have no catalyzing enzyme (step c). The reduction resulted in both a model with reactions and reaction bounds that were only relevant to the input conditions for which the analysis was run (**Table 6.2**), and a set of biologically relevant reactions for elimination with the strain design algorithms. The average number of reactions that were eliminated

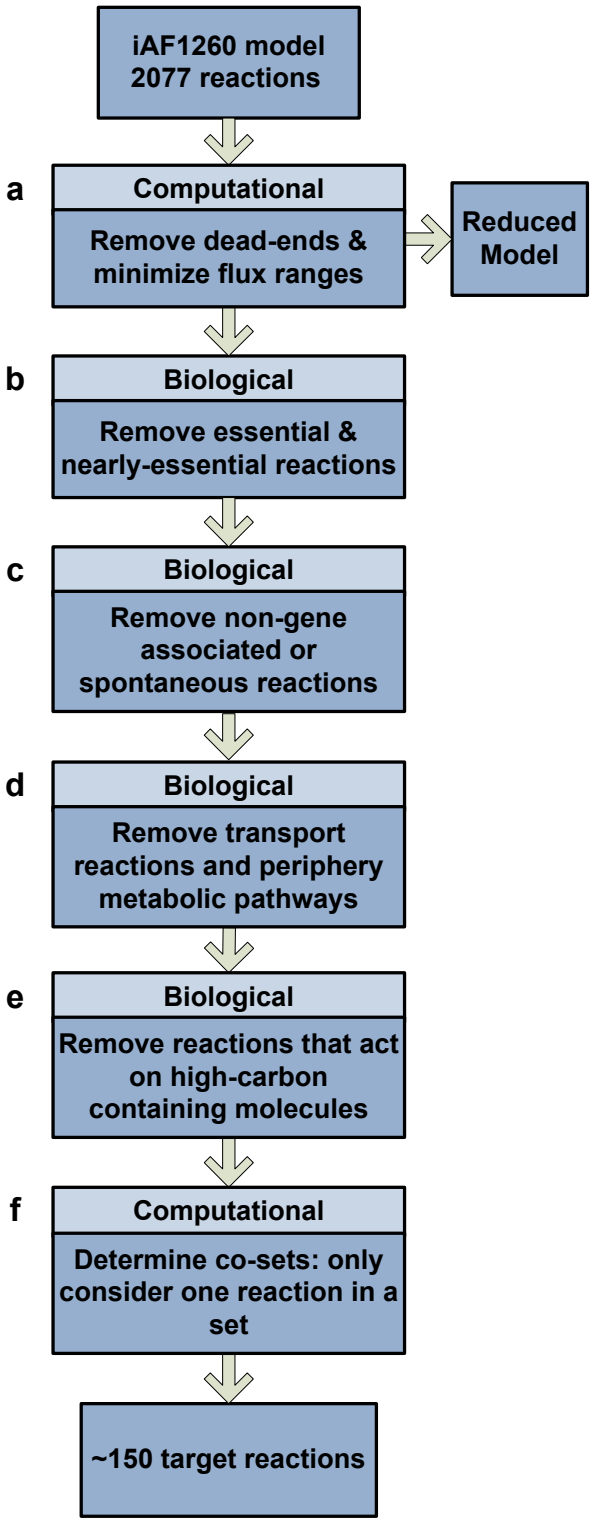
from the total 2077 in the *i*AF1260 reconstruction[3] was 92.6% across the different substrate conditions (**Table 6.2**).

Table 6.2: Substrate conditions.

Carbon Substrate(s)	Aerobicity	Wild Type Growth Rate	Target Reactions After Reduction of Scope
Glucose	Anaerobic	0.459	142
Xylose	Anaerobic	0.319	141
Glycerol	Anaerobic	0.119	140
Glucose	Aerobic	1.276	170
Xylose	Aerobic	1.131	165
Glycerol	Aerobic	0.983	166

Maximum uptake rates for primary carbon sources were set to 20 mmol/gDW^{-1} hr^{-1}. In aerobic simulations, maximum oxygen uptake rate was set to 20 mmol/gDW^{-1} hr^{-1}.

Figure 6.2: Problem Formulation: Reduction of model and selection of targeted reactions. (facing page) Method used to acquire target reactions for deletion from the *E. coli* genome and to reduce computation time. For the six steps, four are based off biological assumptions and two are computational approaches.


iAF1260 model
2077 reactions
a
Computational
Remove dead-ends & minimize flux ranges
Reduced Model
b
Biological
Remove essential & nearly-essential reactions
c
Biological
Remove non-gene associated or spontaneous reactions
d
Biological
Remove transport reactions and periphery metabolic pathways
e
Biological
Remove reactions that act on high-carbon containing molecules
f
Computational
Determine co-sets: only consider one reaction in a set
~150 target reactions

6.3.4 Strain Design: Algorithm computation and output

To design strains of *E. coli* that overproduced the defined target metabolites, we used a combination of the OptKnock[1] and OptGene[2] algorithms with the conditioned model of *i*AF1260[3]. The procedure utilized for this analysis is outlined in **Figure 6.3** (see also, Methods). First, OptKnock was utilized to design strains of *E. coli* for each substrate / target pair in **Table 6.1** for a maximum of 3 and 5 reaction knock-outs allowed (with the exception that glycerol was not selected as a substrate for OptKnock simulations, see above). Each simulation was allowed to run to completion, so that the entire solution space was searched. OptKnock was utilized first as it evaluates the maximum achievable yield given a set number of knock-outs and finds the global optimum set of knockouts as long as such a solution exists. Final and intermediate strain designs from OptKnock were summarized and evaluated for: i.) the category of production envelope they returned (see **Figure 6.1**), and ii.) the maximum production potential (e.g., yield) achieved. After completion of this step, the resulting strain designs were used as a base population for the appropriate substrate/target pair for the OptGene algorithm (except for glycerol, as mentioned). OptGene simulations were then conducted for the objectives of maximum yield (for verification of the OptKnock solutions and for higher knockout designs), substrate-specific productivity, and strength of growth coupling. Results for each analysis and for specific targets are now detailed.

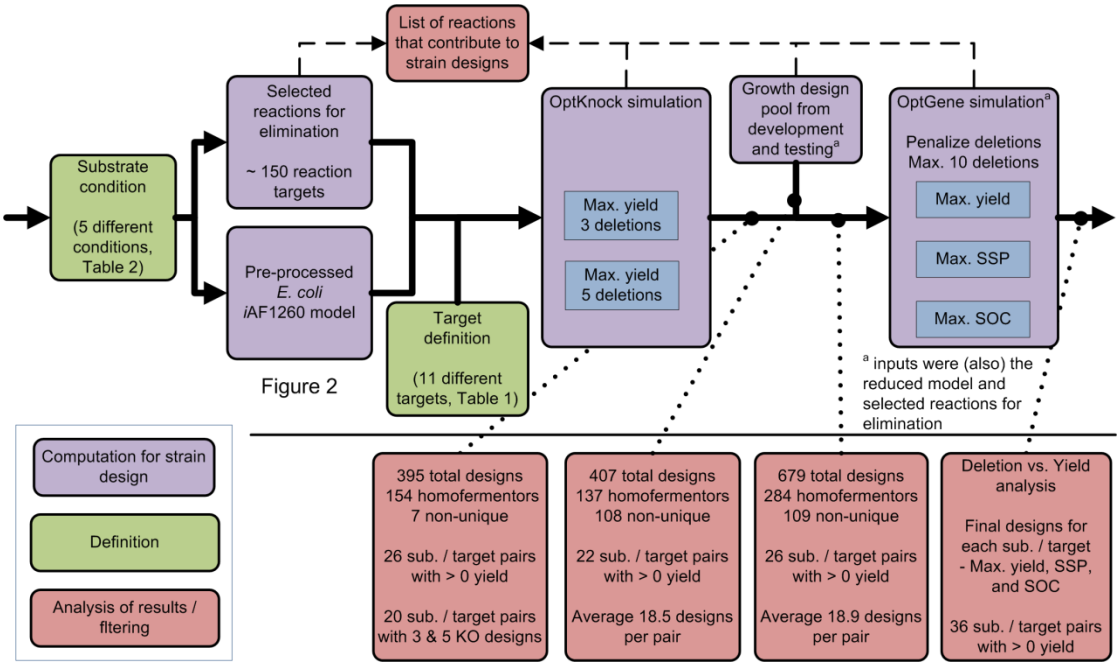


Figure 6.3: Strain design pipeline: the process used to compute strain designs for growth-coupled production in *E. coli*. This workflow outlines the process developed to generate the strain designs for the analysis and the results at various points in the process. Each colored box represents a computation (violet), substrate or target definition (green), or filtering or analysis of results generated during the procedure (red). Starting on the left, the substrate conditions were defined to produce substrate-specific model and the reactions targeted for elimination in the analysis. From here, targets were defined and the OptKnock[1] algorithm was first used to examine lower knockout number maximum yield designs. Using the results from this analysis, designs were fed into simulations with the OptGene[2] algorithm along with results from a testing design pool. OptGene simulation results examined maximum yield, substrate-specific productivity (SSP), and strength of growth coupling (SOC) for up to ten reaction knockouts. Results from different time points in the analysis are given on the bottom. Additionally, reactions that contributed to designs were compared to the initial targets for comparison.

6.3.5 OptKnock analysis of maximum yield for three and five knockout designs

After the first set of simulations was completed for analyzing maximum yield achievable for three and five knockout designs, growth-coupled designs could be found for 26 different substrate / targets pairs. The results of this analysis are given in **Table 6.3**. The production envelopes resulting from the analysis for five different

targets for the given substrates are given in **Figure 6.4** and in **Appendix A**. Overall, this number was 59% out of the potential 44 combinations examined and it demonstrates that target production could be coupled to growth in most of the cases examined. Furthermore, for the 26 different pairs, growth-coupled designs could be identified for both three and five knockout designs in 20 out of the 26 cases resulting in a total of 46 different designs. For five different targets, designs could be found in all of the conditions examined, for three targets, designs were found under some of conditions examined and for the last three, no designs could be identified. This indicates that for the scope of metabolites targeted, if a solution could be found for one condition for a given substrate / target pair, it could be found in the others given these knockout limits in most of the cases. When comparing the results to the theoretical maximum achievable, 13 different substrate / target pairs were at or above 80% of the calculated theoretical maximum achievable (five different targets). These targets were not directly correlated with the number of carbons of the target molecule as designs were found for all categories of metabolite carbon number except for compounds with five carbons (2-oxoglutarate and L-glutamate). The full set of final optimal and intermediate designs (found during the optimization in route to the maximum value) for maximum yield are given in Supplementary Data[48] (in this section, only the final optimal solutions were discussed).

Table 6.3: Strain design properties designed using the OptKnock algorithm.

	Glucose	Xylose	Glucose	Xylose
	Anaerobic	Anaerobic	Aerobic	Aerobic
product	3KO / 5KO / %TMP	3KO / 5KO / %TMP	3KO / 5KO / %TMP	3KO / 5KO / %TMP
Ethanol	36.9* / 38.4* / 100%	31.1* / 31.6* / 99%	25.5* / 36.2* / 94%	19.2* / 30.7* / 96%
D-Lactate	36.2* / 38.4* / 100%	30.7* / 31.0* / 97%	16.8* / 35.5* / 92%	11.9* / 30.5* / 96%
Glycerol	-	-	- / 13.4 / 45%	-
L-Alanine	- / 38.0* / 99%	-	- / 24.3* / 63%	- / 14.8* / 47%
L-Serine	-	-	-	-
Pyruvate	14.2 / 19.1 / 65%	13 / 15.6 / 75%	24.2 / 33.7* / 81%	19.8* / 27.2* / 80%
Fumarate	0.3 / 0.3 / 2%	0.2 / 0.2 / 2%	5.8 / 6.9 / 20%	9.1 / 10.1 / 36%
L-Malate	-	-	-	-
Succinate	18.2 / 25.9 / 90%	19.2 / 19.3 / 92%	2.6 / 17.8 / 55%	7.9 / 19.3 / 73%
2-Oxoglutarate	- / 3.0 / 30%	- / 1.9 / 28%	-	-
L-Glutamate	-	-	-	-

3KO – maximum optimal production rate achievable with 3 knock-outs, 5KO – maximum optimal production rate achievable with 5 knock-outs, %TMP – percentage of the theoretical maximum achievable optimal production rate for the 5KO design, '-' – no design found

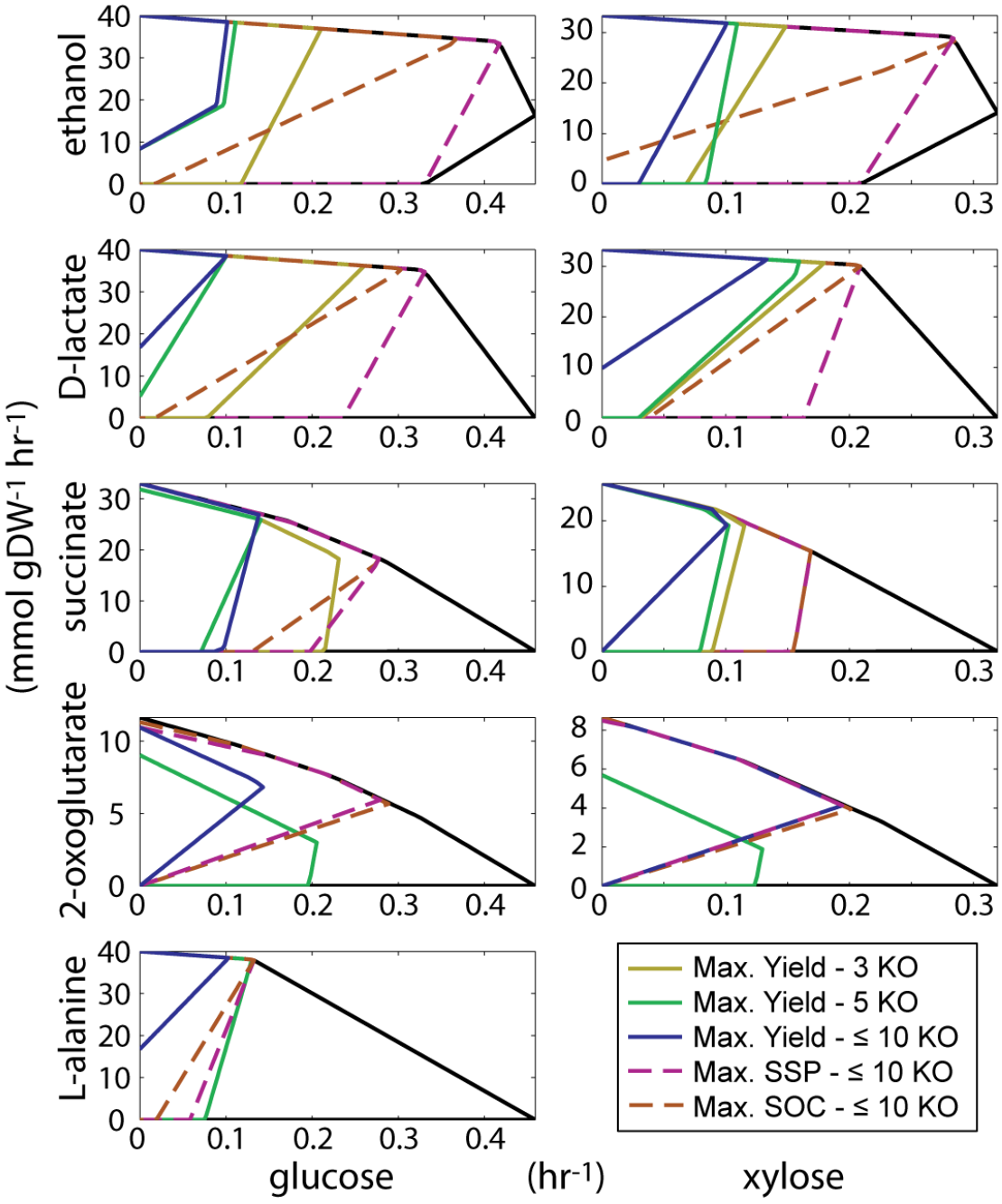
* indicates condition where homofermentation of product is possible (< 2% wt% other carbon products, CO2 exempt)

Conclusions can be drawn for the extent of interventions (i.e., number of knockouts) necessary to couple production of a metabolite to growth by comparing the designs resulting from allowing either three or five knockouts. Examining the cases where knockout designs could only be found with the five knockout limit (i.e., a 3 reaction knockout design was not possible), this occurred for the production L-alanine, 2-oxoglutarate, and glycerol. This indicates that these products require a more complex set of interventions to couple production to growth. Interestingly, the production of L-alanine can be coupled to growth at a very high percentage of the theoretical max (99%) given five different reaction eliminations. The other products for which only a five knockout design could be found did not possess a high percentage of the theoretical maximum (less than 50% in all cases). For cases where both a three and five knockout design were found, addition of the two knockouts in the five

knockout designs gave an average increase of 11% and 121% wt% in yield under anaerobic and aerobic conditions, respectively. Increases varied significantly from these means depending on the target, with the median increases being 3% and 51% wt% for anaerobic and aerobic conditions, respectively. This increase in production potential with an increasing knockout number will be further discussed below.

A number of designs calculated from the three and five knockout analysis are predicted to be homofermentors. As stated earlier, these strains are desirable from the aspect of separation of fermentation products as only one product is made. From the analysis, 22 different designs are predicted to be homofermentors, 48% of the total designs. Given the limit of three knockouts, nine substrate / target pairs were predicted to be homofermentatively generated and an additional four pairs were included in this total when five knockouts was the design limit. From these results, four different products could be made homofermentatively, ethanol, D-lactate, L-alanine, and pyruvate. Additional designs calculated using these results as inputs can be examined and compared to the potential yield with higher numbers of designs. With these results as a basis, further computations were performed examining higher numbers of knockouts and additional optimization criteria (**Figure 6.1**) using the OptGene algorithm.

Figure 6.4: The strain designs generated for five different targets from glucose and xylose anaerobically. (facing page) A set of graphs that give the production envelopes for different substrate / target pairs that were calculated during the analysis under anaerobic conditions. The different target production rates (mmol gDW^{-1} hr^{-1}) are shown on the y-axis and the growth rate (hr^{-1}) is given on the x-axis. Shown on each plot (if a solution exists) are the maximum yields, $Y_{p/s}$, for 3 knockouts (yellow, solid line), 5 knockouts (green, solid line), up to 10 knockouts (with a 99.99% deletion penalty, blue, solid line), the maximum substrate-specific productivity (SSP, pink, dashed line), and the maximum strength of growth coupling (SOC, orange, dashed line) design. For example, there are no valid solutions for L-alanine production on xylose given the minimum growth rate of 0.1 hr^{-1}.


ethanol
D-lactate
succinate
2-oxoglutarate
L-alanine
(mmol gDW-1 hr-1)
glucose
(hr-1)
xylose
Max. Yield - 3 KO
Max. Yield - 5 KO
Max. Yield - ≤ 10 KO
Max. SSP - ≤ 10 KO
Max. SOC - ≤ 10 KO

6.3.6 OptGene analysis for maximum yield, substrate-specific productivity, and strength of growth coupling for up to 10 knockouts

The OptGene algorithm was utilized to examine growth-coupled strain designs of *E. coli* with a higher limit of knockouts and additionally, non-linear objectives such as substrate-specific productivity (SSP) and strength of coupling (SOC). OptGene is a genetic algorithm has the advantages of utilization of a non-liner objective and potentially faster run-times to find an optimal solution. One potential drawback is that it does not ensure that an optimal solution is found for the space examined[2]. Nonetheless, it has been shown to calculate strain designs efficiently[2] and the role OptGene played in the analysis is outlined in **Figure 6.3**. As the final and intermediate solutions from the three and five knockout designs from OptKnock were used as an initial population for the algorithm (i.e., inputs), the calculations from this step were both solution filtering steps (e.g., if an intermediate solution found with OptKnock had a higher SSP than the final maximum yield design, it would be chosen) and improvements made to the existing solutions along with finding new solutions.

The results from the examination of the different objective functions using the OptGene algorithm and outlined method (**Figure 6.3**) are given in **Tables 6.4-6.6**. Each table summarizes the results from examining the different objective functions and therefore, each will be discussed separately.

Analysis of maximum yield for up to ten knockout designs: The analysis of maximum yield using OptGene was able to improve the potential yield and find designs for some target / substrate pairs with a higher allowance of knockouts that were not previously identified. For the analysis, OptGene was run using maximization of the yield with a tilted objective function, a deletion penalty of 99.99% (see

Methods), and a maximum allowable number of ten knockouts. Utilizing these parameters for the analysis, it was possible to identify improvements in the maximum predicted yield for 28 substrate / target pairs, with 18 (64%) of these being slight improvements (an increase less than 5%). This comparison was valid for examining increases in production from glucose and xylose as substrates. Additionally, four of these the cases were designs which were earlier not found with the five knockout limit (see below). This total does not include any designs utilizing glycerol as a substrate, as these we not analyzed with the lower knockout limit (see above). In total, there were 6 different targets that could be coupled to growth using glycerol as a substrate in aerobic conditions.

Table 6.4: Strain design properties designed using the OptGene algorithm – maximum yield.

	Glucose	Xylose	Glucose	Xylose	Glycerol
	Anaerobic	Anaerobic	Aerobic	Aerobic	Aerobic
product	P / KO / %TMP	P / KO / %TMP	P / KO / %TMP	P / KO / %TMP	P / KO / %TMP
Ethanol	38.5* / 7 / 100%	31.8* / 9 / 100%	37.9* / 10 / 98%	31.7* / 7 / 100%	18.6* / 7 / 86%
D-Lactate	38.5* / 8 / 100%	31.4* / 10 / 98%	37.7* / 10 / 98%	31.3* / 10 / 98%	18.6* / 10 / 93%
Glycerol	2.1 / 10 / 14%	-	19.4 / 10 / 66%	13.2 / 7 / 55%	N/A
L-Alanine	38.5* / 10 / 100%	-	31.1* / 10 / 81%	22.4* / 5 / 71%	14.3* / 8 / 72%
L-Serine	-	-	-	-	-
Pyruvate	19.8 / 7 / 67%	16.3 / 10 / 78%	38.6* / 10 / 93%	31.5* / 10 / 92%	18.3 / 10 / 86%
Fumarate	0.3 / 3 / 2%	0.2 / 3 / 2%	8.1 / 7 / 23%	11.3* / 10 / 40%	8.3 / 9 / 44%
L-Malate	-	-	-	-	-
Succinate	26.9 / 8 / 93%	19.3 / 6 / 92%	18.9 / 10 / 59%	19.3 / 9 / 74%	8.9* / 10 / 51%
Oxoglutarate	6.8 / 10 / 69%	4.1 / 9 / 62%	15.7* / 10 / 64%	14.3* / 10 / 71%	-
-Glutamate	-	-	-	-	-

P – maximum optimal production rate achievable, KO – the number of knockouts needed to achieve the given production rate, %TMP – percentage of the theoretical maximum achievable optimal production rate, '-' – no design found, N/A – not applicable

* indicates condition where homofermentation of product is possible (< 2% wt% other carbon products, CO_2 exempt)

There were additional substrate / target pair designs that could be identified using the increased number of knockouts. Specifically, the production of 2-oxoglutarate could be coupled to growth aerobically from glucose and xylose as substrates with 64% and 71% of the theoretical maximum potential achievable, respectively. These designs are also homofermenting designs Additionally, there was also a significant increase in the ability to couple the production of 2-oxoglutarate to growth anaerobically on glucose and xylose as well, increases of 38% and 33% in the theoretical maximum potential achievable, respectively. The same results were observed for the production of glycerol from xylose aerobically, where the allowance of more knockouts resulted in finding a previously unidentified growth coupled solution, at 55% the theoretical maximum. Three was also a significant increase, 20% of the theoretical maximum, for production of glycerol from xylose anaerobically and a design could be identified for glycerol production from glucose anaerobically, as well. Examining the products that could be coupled to growth aerobically on glycerol, the production of ethanol, lactate, and pyruvate could all be coupled to growth with over 80% of the theoretical maximum potential predicted. Fumarate, L-alanine, and succinate could also be coupled to growth. These cases represent instances where growth coupling is only achievable (or better achievable) with more complex metabolic inventions to cellular metabolism.

The results from the OptGene analysis sheds some light on the point at which additional knockouts are no longer beneficial to increase yield. Although many of the designs returned have the maximum number of knockouts, ten, there are several cases where the minor deletion penalty deterred additional unbeneficial knockouts. This occurred in 17 out of the 36 cases where solutions were found. There were no target metabolites for which every OptGene design had the maximum number of

knockouts. However, there were four target metabolites where all but one of the optimal solutions had the maximum number of knockouts, D-lactate, pyruvate, 2-oxoglutarate, and glycerol. Another interesting result is that for the anaerobic production of fumarate. The maximum production achievable was with only 3 knockouts. This was the only case where the number of knockouts decreased from the initial design of five indicating that additional knockouts over three are not beneficial to this low producing design. These findings will be further addressed below.

Analysis of maximum substrate-specific productivity for up to ten knockout designs: In addition to optimizing for the maximum yield achievable for a given condition, the additional property of optimizing for maximum substrate-specific productivity was examined. These strains are desirable as they have the potential to maximize the rate of production for a given substrate / target pair as the growth rate is factored into the calculation. **Figure 6.4** and **Table 6.5** present the results from this analysis. By examining the different envelopes on the same plot in **Figure 6.4**, it is possible to visualize the tradeoff between yield and growth rate for maximum SSP designs. **Table 6.5** gives both the SSP for the optimal SSP design and the optimal maximum yield design for comparison.

Table 6.5: Strain design properties designed using the OptGene algorithm – maximum substrate-specific productivity.

	Glucose	Xylose	Glucose	Xylose	Glycerol
	Anaerobic	Anaerobic	Aerobic	Aerobic	Aerobic
product	P / SSP / KO max. SSP max. $Y_{p/s}$	P / SSP / KO max. SSP max. $Y_{p/s}$	P / SSP / KO max. SSP max. $Y_{p/s}$	P / SSP / KO max. SSP max. $Y_{p/s}$	P / SSP / KO max. SSP max. $Y_{p/s}$
Ethanol	33.3* / 13.9 / 2	28.8* / 8.2 / 2	16.6* / 19.4 / 3	13.4* / 12.0 / 3	8.3* / 4.7 / 7
	38.5* / 3.9 / 7	31.8* / 3.2 / 9	37.9* / 4.9 / 10	31.7* / 3.2 / 7	18.6* / 1.9 / 7
D-Lactate	34.7* / 11.5 / 2	30.0* / 6.3 / 2	17.1* / 19.3 / 4	13.3* / 12.1 / 5	8.3* / 4.8 / 6
	38.5* / 3.9 / 8	31.4* / 4.2 / 10	37.7* / 4.2 / 10	31.3* / 3.2 / 10	18.6* / 1.9 / 10
Glycerol	2.1 / 0.5 / 10	-	14.7 / 12.3 / 6	12.6 / 8.8 / 6	N/A
	2.1 / 0.5 / 10	-	16.9 / 2.2 / 10	13.2 / 1.5 / 7	N/A
L-Alanine	37.9* / 5.0 / 6	-	19.1* / 18.9 / 6	14.8* / 12.1 / 5	7.8* / 4.8 / 8
	38.5* / 3.9 / 10	-	31.1* / 4.8 / 10	14.8* / 12.1 / 5	7.8* / 4.8 / 8
Pyruvate	18.8 / 5.8 / 6	15 / 3.1 / 6	22.0* / 23.0 / 3	16.3* / 14.9 / 2	9.9* / 5.8 / 4
	19.8 / 2.0 / 7	16.3 / 1.7 / 10	38.6* / 2.8 / 10	31.5* / 4.4 / 10	18.3 / 2.1 / 10
Fumarate	0.3 / 0.1 / 3	0.2 / 0.1 / 3	8.1 / 3.8 / 7	8.8 / 7.1 / 6	8 / 4.5 / 7
	0.3 / 0.1 / 3	0.2 / 0.1 / 3	8.1 / 3.8 / 7	11.3* / 1.5 / 10	8.3 / 4.1 / 9
Succinate	18.3 / 5.1 / 5	15.4 / 2.6 / 3	14.7 / 14.3 / 9	11.7* / 10.3 / 6	5.6* / 4.1 / 6
	26.9 / 3.7 / 8	19.3 / 1.9 / 6	18.9 / 2.9 / 10	19.3 / 1.9 / 9	8.9* / 4.0 / 10
-Oxoglutarate	5.9 / 1.7 / 9	4.1 / 0.8 / 9	15.7* / 10.8 / 10	11.8* / 8.8 / 8	-
	6.8 / 1.0 / 10	4.1 / 0.8 / 9	15.7* / 10.8 / 10	14.3* / 7.1 / 10	-

P – maximum optimal production rate achievable, SSP – substrate specific productivity, KO – the number of knockouts needed to achieve the given production rate, '-' – no design found, , N/A – not applicable

* indicates condition where homofermentation of product is possible (< 2% wt% other carbon products, CO_2 exempt)

Optimization of the SSP resulted in an overall increased productivity over maximum yield simulations for most of the cases examined. The increase in SSP overall for anaerobic and aerobic conditions was 75% and 221%, respectively. Clearly, the potential for SSP improvement is greatest under aerobic conditions. However, a significantly larger decrease in yield, 31%, has to be accepted (compared to 8% for anaerobic conditions) for the inherent tradeoff when considering maximum SSP under aerobic conditions. Some of the largest increases in SSP were encountered for ethanol (257% and 155%), D-lactate (198% and 50%), and pyruvate (191% and 82%) from glucose and xylose anaerobically, respectively. Under aerobic

conditions, the percent increases are much more drastic and in some cases are over four fold (e.g., pyruvate production from glucose aerobically, 731% increase and glycerol production from xylose aerobically, 484% increase). These trends can be further seen by examining **Figure 6.4** for anaerobic conditions and **Appendix A** for aerobic conditions. Furthermore, the strains that are homofermentors are also marked on the table indicating that although there is a tradeoff for yield, homofermentation is still possible for maximum SSP designs (21 of the designs are homofermenting designs).

Analysis of maximum strength of growth-coupling for up to ten knockout designs: Similar to the maximization of substrate-specific productivity, optimizing for maximum strength of growth-coupling produces desirable production phenotypes. These strain designs show the characteristic of having a stronger likelihood of evolving to the predicted production phenotype with and increasing growth rate. This characteristic is advantageous when using the process of adaptive evolution[35,36] to evolve growth-coupled production strains to their optimal phenotypes. A metric that we used to characterize this level of growth coupling is the growth coupled parameter (GCP) which is the percentage of the growth rate that is growth-coupled (the percent gamma is of growth rate in **Figure 6.1**). The results from an analysis optimizing for maximum strength of growth-coupling is given in **Table 6.6**.

Table 6.6: Strain design properties designed using the OptGene algorithm – maximum strength of growth coupling.

	Glucose	Xylose	Glucose	Xylose	Glycerol
	Anaerobic	Anaerobic	Aerobic	Aerobic	Aerobic
product	P / GCP / KO max. GCP max. $Y_{p/s}$	P / GCP / KO max. GCP max. $Y_{p/s}$	P / GCP / KO max. GCP max. $Y_{p/s}$	P / GCP / KO max. GCP max. $Y_{p/s}$	P / GCP / KO max. GC max. $Y_{p/s}$
Ethanol	34.5* / 100% / 6	28.3* / 100% / 2	31.5 / 100% / 7	28.3* / 100% / 5	17.7* / 100% / 8
	38.5* / 100% / 7	31.8* / 71% / 9	37.9* / 45% / 10	31.7* / 55% / 7	18.6* / 87% / 7
D-Lactate	35.2* / 100% / 6	30.3* / 100% / 5	35.2* / 100% / 9	30.3* / 100% / 8	18.1* / 85% / 6
	38.5* / 100% / 8	31.4* / 100% / 10	37.7* / 89% / 10	31.3* / 83% / 10	18.6* / 85% / 10
Glycerol	2.1 / 3% / 10	-	19.4 / 88% / 10	9.7 / 30% / 7	N/A
	2.1 / 3% / 10	-	19.4 / 88% / 10	13.2 / 21% / 7	N/A
L-Alanine	37.9* / 85% / 7	-	19.3 / 38% / 10	22.4* / 37% / 7	14.3* / 24% / 8
	38.5* / 100% / 10	-	31.1* / 37% / 10	22.4* / 37% / 7	14.3* / 24% / 8
Pyruvate	19.1 / 62% / 8	15.7 / 34% / 6	32.3* / 90% / 10	26.1* / 100% / 10	14.6* / 59% / 10
	19.8 / 100% / 7	16.3 / 35% / 10	38.6* / 100% / 10	31.5* / 100% / 10	18.3 / 22% / 10
Fumarate	0.3 / 6% / 10	0.2 / 6% / 10	7.7 / 13% / 10	8.5 / 22% / 9	6.4 / 14% / 10
	0.3 / 3% / 3	0.2 / 3% / 3	8.1 / 2% / 7	11.3* / 12% / 10	8.3 / 8% / 9
Succinate	17.8 / 100% / 4	15.4 / 100% / 3	6.8 / 100% / 9	11.7* / 100% / 9	3.2 / 100% / 7
	26.9 / 100% / 8	19.3 / 100% / 6	18.9 / 38% / 10	19.3 / 41% / 9	8.9* / 100% / 10
2-Oxoglutarate	5.7 / 100% / 7	3.9 / 100% / 7	13.3* / 100% / 8	11.8* / 100% / 8	-
	6.8 / 100% / 10	4.1 / 100% / 9	15.7* / 3% / 10	14.3* / 4% / 10	-

P – maximum optimal production rate achievable, GCP – growth coupling parameter, percent of growth rate that is coupled with target production (see **Figure 6.1**, the percent gamma is of the growth rate), KO – the number of knockouts needed to achieve the given production rate, '-' – no design found, N/A – not applicable

* indicates condition where homofermentation of product is possible (< 2% wt% other carbon products, CO_2 exempt)

The production of the target metabolites examined revealed that optimization for SOC produces strain designs that have fully coupled production phenotypes (100% GCP) and this property can be increased over maximum yield designs. The optimal SOC design was an increase in the GCP metric for 3 substrate / target pairs aerobically and in 15 cases aerobically. As the GCP metric is not the actual optimization parameter, the value can also decrease with maximum SOC optimization. A decrease was seen in 3 and 1 cases anaerobically and aerobically,

respectively. However, the GCP was chosen as a value to quantify the SOC as it can be easily understood by examining the relationship between gamma and growth rate.

6.3.7 Characterization of the solution space: reactions that contribute to designs and the relationship between number of knockouts and metabolite production

In total, the designs calculated during this analysis contain the reactions that allow the diversion of flux in the *E. coli* network while still generating sufficient energy and biomass precursors. To summarize this set of reactions, all of the optimal and intermediate solutions were complied and all of the reactions participating in this set were identified. In total, 132 reactions, or 77% of possible reaction knockouts, were in this pool and the full list of reactions is given in Supplementary Data[48]. From this pool, some reactions participated more often in growth-coupled designs. Pyruvate formate lyase, pyruvate dehydrogenase, and acetate kinase occurred around eight times more often than the average of 19 solutions per reaction. Lactate dehydrogenase, acetaldehyde dehydrogenase, and phosphoglycerate dehydrogenase all occurred more than four times as often as the average. The uneven distribution of reaction knockout occurrences suggests that certain reactions are critical hubs for diverting carbon flux.

In total, the design space for growth-coupled designs for 11 different targets was analyzed for up to ten reaction knockouts. From this, we characterized the production potential achievable given a number of knockouts for different substrates. **Figure 6.5** gives the relationship between production potential and number of knockouts for anaerobic conditions. The production potential metric given is the percent theoretical maximum achievable as a function of number of knockouts. The

production of each of the targets can be categorized into three different classes: i.) theoretical maximum (or very near) achievable, ii.) moderate production achievable with increasing interventions, and ii.) low production potential regardless of number of interventions. Targets such as ethanol, lactate, succinate, L-alanine belong in the first category of near or at theoretical production achievable (over 80% of TMP) and all of the targets in this category can be coupled with a relatively low number of knockouts (less than or equal to five knockouts). The production of other metabolites, such as 2-oxoglutarate and pyruvate, can be increased in higher amounts with an increasing number of knockouts. However, even with up to ten knockouts, the production never reached the 80% TMP level. The third category, low production potential (less than 20% TMP) regardless of the number of interventions contains the metabolites glycerol and fumarate. These productions cannot be coupled to growth at a high rate given the ten knockout limit used in the analysis. The same type of analysis can be performed for aerobic conditions.

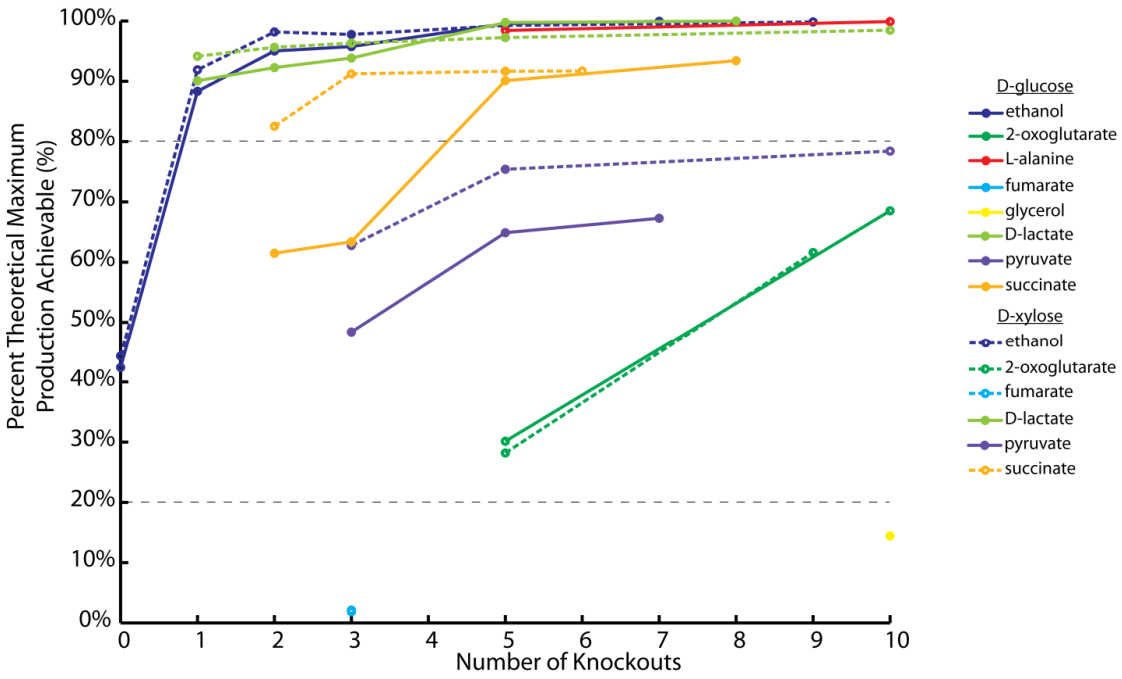


Figure 6.5: Theoretical maximum production achievable for different substrate / target pairs for under anaerobic conditions. A plot of the percent theoretical maximum production achievable for different substrate (top of column) and target (listed) combinations as a function of number of knockouts allowed to the system. Each point is the maximum value for a given number of knockouts. The plot contains the data from examining growth of E. coli under anaerobic conditions. The lowest number of knockouts for a design can be coupled to growth is the leftmost point for each substrate / target pair. Also shown are the cutoffs (20% and 80%) that delineate the three different categories of designs. The plot characterizes the relationship between the number of knockouts necessary to growth couple a product and achievable production. A number of products can be generated from three knockouts very near to the maximum production achievable.

6.4 Discussion

Metabolic engineering of microorganisms is a powerful tool that can be used to generate renewable compounds and products for a growing human population. Systems biology and *in silico* analysis have the potential to accelerate the production of new strains and products through model-driven analyses. This work presents work towards this goal through a systematic analysis of the production potential for native products from the bacterium *E. coli* by examining products that can be coupled to

growth rate. Accordingly, the main results from this study are, i.) sets of specific metabolic interventions (i.e., knockouts) that couple metabolite production at a high yield and high rate to growth rate which can be experimentally implemented, ii.) characterization of the production potential design space for native products for *E. coli* including the key reactions that divert metabolic flux inside the cell, and iii.) an outlined and implementable procedure to examine and characterize the production potential for a given organism. With the increasing adaptation and implementation of the systems biology approach of constraint-based reconstruction and modeling[22,23,28], the procedure can be readily adapted to examine a number of organisms-of-interest. For many production organisms, genome-scale reconstructions and models already exist[22].

Growth-coupling has emerged as a design procedure possessing the potential to integrate model-driven design and strain optimization through adaptive evolution. The adaptive evolution process has been performed and the outcomes characterized[35,36,49] and initial computational analyses examining test cases have been performed[1,33]. Furthermore, initial work demonstrating experimental feasibility of growth-coupling has been demonstrated[33]. However, a rigorous analysis of the design potential was needed to fully examine the range of metabolites that can be produced through coupling to growth. This analysis provides the range and magnitude to which products can be coupled to growth in *E. coli*. Additionally, coupling production of metabolites to growth has advantages over current production strains as they are suitable for continuous culture conditions. Continuous culture has a distinct advantage in overall productivity as it can be run for significantly longer periods of time. Production strains that are not growth coupled will ultimately be overtaken by optimally growing mutants in continuous culture conditions where growth rate is the

selection pressure (a selective pressure commonly applied to continuous culture). However, the strains presented in this work are fundamentally designed to be the top growers under these steady state conditions and therefore hold significant productivity advantages.

The reactions presented in the analysis were designed to be biologically relevant targets that can be eliminated through genetic manipulation (i.e., reaction knockouts) and tested in an *E. coli* strain. This was facilitated by the model preprocessing procedure. The top-down model preprocessing approach is a novel procedure that has not been examined in earlier analyses[1,2,33,37,38], which rely on a bottom-up approach that is more computationally intensive and could inherently miss key reactions. This is due to the fact that they are not rooted in an approach that decides on elimination targets according to function. The knockout genes or reactions selected by these methods could be impractical to implement *in vivo*. In the present analysis, all of the candidate knockout reactions are biologically relevant. Therefore, these additional measures should provide greater confidence in the feasibility of implementing the reported strain designs.

Designs were reported for three different strain design criteria, each with desirable production qualities. The maximum yield optimizes the amount of substrate that can be converted to the target metabolite and thus can result in a strain with the greatest output per unit input. The analysis of SSP quantifies the tradeoff between yield and productivity. If an overall maximum rate is desirable without considering yield, then optimization of SSP is the best objective. Furthermore, maximum SOC designs have implications in success rates of a strain evolving to the optimal phenotype. Even though designs were reported for each of these design types, a

combination of the three properties is probably the best choice. For example, the maximum SOC design for the production of L-alanine from glucose anaerobically in **Figure 6.4** possesses the near optimal for all of the properties.

The production of some metabolites could not be coupled to growth in our analysis. Some of the key parameters relating to this possibility that were used in this analysis are the minimum necessary growth rate, the substrate uptake rate, and the maximum allowable number of knockouts. In regards to the minimum necessary limit that was placed on the growth rate, lowering this value could result in the identification of more designs. However, in doing so, there is the risk that experimental implementation of a strain from this design might result in a lethal combination of knockouts and ultimately a strain that cannot be generated. This result could be possible as there are some modeling assumptions (e.g., maintenance energies) in the modeling approach which might differ from actual requirements *in vivo* and thus, the cell may not be able generate sufficient energy for growth. There is also the opposite case, where the estimates are greater than the actual requirements *in vivo* and lowering the minimum growth rate might allow more compounds to be coupled. In the same respect, increasing the uptake rates of the main substrate might also increase the number of strain designs identified. Furthermore, increase the number of allowable knockouts might also allow the coupling of more products. However, as the plots in **Figure 6.5** indicate, more knockouts do not necessarily correlate with higher production in all cases.

The analysis performed in this study characterizes the solution space for *E. coli* in regards to the potential of the cell to couple production of the examined metabolites to growth rate. Initially, the theoretical maximum potential analysis framed

the scope of the analysis and readily identified native products that can be generated with a high yield given the machinery encoded by *E. coli's* genome. For example, the finding that only ethanol could be made with over a 15% yield from glycerol anaerobically demonstrates that at the biological parameters used for simulations, this substrate anaerobically is not a good production feedstock. The biological parameters that were used in the analysis (e.g., maintenance values, minimum growth rates, etc.) predict that most of the energy contained in glycerol has to be devoted to growth of the cell anaerobically. Aerobically, the advantage that oxygen provides as an electron sink allows the potential for higher substrate yields for *E. coli* growing on glycerol. These yields would change if the actual biological parameters were much different than those used in the analysis (these parameters have been empirically determined from experimental data[3]). However, as the design criteria were set in this analysis, as is done for most engineering approaches, this type of modeling can bring these non-starting issues to light. Similarly, the same case can be made for the input parameters, such as substrate uptake rates. Nonetheless, this analysis frames the production capabilities of *E. coli* and can be readily extended to different medium substrates or supplements, as well as other organisms for which organism-specific models exist[22].

The strain designs presented for the different target and substrate pairs, along with performing the analysis for different numbers of knockouts, allows a characterization of the space which is achievable for growth coupling. This topic can be understood by examining the relationships displayed in **Figure 6.5** for each substrate / target pair. The trend of growth-coupled production relative to the number of interventions necessary provides the tradeoff between making more knockouts in a cell and the resulting potential increase in yield (or SSP, SOC, etc.).Taken together,

all of the combinations characterize the full production potential of *E. coli*. It was demonstrated that there are three different classes of metabolite production; theoretical maximum (or very near) achievable, moderate production achievable with increasing interventions, and low production potential regardless of number of interventions. It is expected that this trend will continue as different substrates or targets are analyzed in a similar fashion. However, some compounds in the moderate production category may be able to reach a value neat the theoretical maximum potential achievable.

A way to understand why some products cannot be coupled to growth is the concept of a hierarchy of metabolites which can be excreted from the system while still making energy. This hierarchy is differentiated by how much energy the cell can obtain from a substrate and subsequently how little it has to excrete in a particular metabolite. To engineer *E. coli* to make one of the products in this hierarchical list, one can think of eliminating pathways to metabolites at the high-energy end of the list and eliminating them one by one until a metabolite of interest is reached. This is of course assuming that it is known how to eliminate the production of metabolites. While performing this elimination of pathways to excrete given metabolites, there will be a point in time where the cell can simply not make the energy it needs to survive by eliminating any more reactions or pathways. This point, or limit, will be defined by the constraints on the system (e.g., the maximum uptake rates of substrates, the minimum growth rate) and some metabolites will lie below this threshold and therefore will not be viable targets in a growth-coupled production scenario. In this analysis, the production of L-alanine could not be produced anaerobically from xylose. However, if the minimum necessary growth rate is lowered (e.g., halved to 0.05 hr-1), a growth-coupled design can be identified.

The workflow and procedure presented in this analysis can be utilized as a platform to perform similar analyses with additional organisms. The combination of using the OptKnock[1] and OptGene[2] algorithms in combinations is a novel approach that has not been previously implemented. It combines the speed and versatility of OptGene with the rigorous search potential of OptKnock to efficiently identify strain designs up to a high knockout number for a variety of conditions. Additionally, it strengthens a potential weakness of OptGene to not find global maxima by feeding into it designs that already have optimal characteristics. However, this does not ensure the identification of the global optimal in all cases, but at the least, provides a strong starting point. Furthermore, the tilting of the objective function is a unique way to efficiently eliminate non-unique (see **Figure 6.1**) solutions from computations which do not ensure the production of the targeted metabolites, even after successful evolution to optimal growth rate. Tilting not only afforded faster computational run time over performing a separate two step optimization, but also allowed a way to implement non-unique design elimination into the mixed-integer linear programming (MILP) framework of OptKnock. With the advancement of the genome sequencing, the established process of metabolic reconstruction and analysis[22,50], and this outlined procedure, it should be possible to identify organisms which intrinsically possess production potential for compounds of interest. Furthermore, this process will expand in scope with the computational analysis of adding content to the cell to expand the range of products and potential of native metabolites.

6.5 Methods

6.5.1 Model

The metabolic reconstruction of *E. coli i*AF1260[3] was utilized as a basis for the model used throughout the work described herein with minor changes to network content (see **Appendix A**). This model has been functionally tested and verified against experimental data to be predictive for computations of growth rates, metabolite excretion rates, and growth phenotypes on a number of substrate and genetic conditions[3]. New additions to the reconstruction and model were added by examining experimental data from published work. For all simulations, the reactions CAT, SPODM, and SPODMpp (oxidative stress reactions) and the FHL reaction were constrained to zero for reasons previously established[3].

6.5.2 Flux balance analysis and strain design computations

Flux balance analysis (FBA) was used for computing optimal phenotypes using *i*AF1260 and the outlined biomass objective function, BOF_{CORE} with the reported maintenance energies, presented with the reconstruction[3]. FBA, performed using an assumption of steady-state metabolite flux, has been described in detail previously[23]. All computations were performed using the MATLAB® (The MathWorks Inc., Natick, MA) and the COBRA Toolbox[51] software packages with TOMLAB (Tomlab Optimization Inc., San Diego, CA) solvers.

OptKnock[1] and OptGene[2] were implemented in the COBRA Toolbox framework as described in their original documentation. OptGene was modified to allow either the deletion of genes or reactions from a simulation to determine genotypes that resulted in desirable production characteristics (see below). To more efficiently determine strain designs that possessed different desirable phenotypes (e.g., maximum yield or maximum substrate-specific productivity), solutions from OptKnock, when available, were used as inputs to OptGene (see **Figure 6.3**).

OptKnock final and intermediate solutions (which were saved whenever OptKnock found a better iterative solution during the course of the simulation) were used to create individuals forming an initial population for each substrate for which a valid OptKnock solution was found. Additionally, individuals created with OptKnock suggested knockouts which were randomly introduced to the population with a 1 / (3 * population size) probability every generation, in order to ensure that these OptKnock solutions continue to be present in the population throughout the simulation. The secondary objective function from which strain selection was based for OptKnock was substrate yield. The secondary objective functions examined for OptGene were substrate yield, substrate-specific productivity, and the degree of growth-coupling (see **Figure 6.1**). Details of the implementation of the OptGene algorithm are given in supplementary text (see **Appendix A**).

The simulations were run to completion for three and five maximum knockout simulations. However, due to time constraints resulting from the extremely large solution space (over 1 quadrillion possible ten knockout combinations from a pool of 150 target reactions) and the breadth of this study, the ten knockout simulations were limited to one week run time, and therefore better solutions could potentially exist.

Consumption rate (or substrate uptake rate) for the main carbon substrate in each simulation was set to 20 mmol gDW^{-1} hr^{-1}. If aerobic conditions were used, an oxygen uptake rate of 20 mmol gDW^{-1} hr^{-1} was also used. These values are close to that observed experimentally for aerobic and anaerobic cultures[45,52].

6.5.3 Tilting of the objective function

Many of the genotypes that the OptKnock function returns after completion result in a ‘non-unique’ phenotype (see **Figure 6.1**), consisting of equivalent optimal

solutions. This is an undesired phenotype as evolution to the predicted optimal behavior does not ensure production of the desired compound (i.e., it could produce the other equivalent product(s)). In order to alleviate this problem, the OptGene and OptKnock algorithms were run with a ‘tilted’ objective function that maximizes growth rate while also slightly minimizing yield, causing the function to return the bottom point of a ‘non-unique’ solution. This point represents the minimum flux expected through the target reaction. Tilting of the objective function in OptGene was accomplished by augmenting the objective vector at the element for the outer membrane transporter for the target production, which is directly coupled to the exchange reaction for the target. In OptKnock, tilting was accomplished by both augmenting the objective vector in the same fashion as with OptGene, as well as setting a constraint minimum production rate on this same reaction equal to the value of augmentation. Tilting the objective has the effect of slightly minimizing the target reaction, since the outer membrane reaction is oriented inwards, resulting in the lower of two previously equivalent solutions to be chosen as optimal. This process was effective at selecting only the minimum production rate for a target reaction for a ‘non-unique’ solution, allowing the algorithm to identify a solution with the highest minimum production rate.

6.5.4 Objective functions used for strain design selection and substrate constraints

Strain designs examined during the project were evaluated by three different production phenotypes (equations 1-3) under the stipulation that all designs had to be growth coupled (see intro). Each equation examines a different desirable production phenotype. The equations examined in this study were product yield, substrate

specific productivity, and strength of growth coupling (see **Figure 6.1**). All of these values are calculated for an optimally performing strain at steady state. The units for each measure are also given in parentheses.

Product yield ($Y_{p/s}$): Maximum amount of product that can be generated per unit of substrate.

$$product\ yield, Y_{p/s} = \frac{production\ rate_{product}}{consumption\ rate_{substrate}} \left(\frac{mmol}{mmol}\right)\left(\frac{gm}{gm}\right) \quad (1)$$

Substrate-specific productivity (SSP): Product yield per unit substrate multiplied by the growth rate

$$substrate\ specific\ productivity =$$

$$product\ yield\ \times growth\ rate \left(\frac{mmol}{mmol \times hr}\right)\left(\frac{gm}{gm \times hr}\right) \quad (2)$$

Strength of coupling: Product yield per unit substrate divided by the slope of the lower edge of the production curve

$$strength\ of\ growth\ coupling = \frac{product\ yield}{slope}\left(\frac{1}{hr}\right) \quad (3)$$

The *slope* in this function is the slope of the line between the point of minimum production rate at maximum growth and the point of maximum growth at zero production on a production envelope plot (**Figure 6.1**). When this slope is high, it is possible for a strain to grow at very close to the maximum growth rate with only a small production rate, which is undesirable. Therefore, optimizing for maximum production rate is the same as optimizing for maximum product yield. Maximizing for substrate specific productivity (also called the Biomass-Product Coupled Yield (BPCY)[2]) introduces a non-linear objective function, which can be handled by OptGene but not OptKnock. Similarly, the strength of coupling is also a non-linear

objective function and can only be handled by OptGene. Additionally, a penalty can be added to the scoring function in OptGene by multiplying the objective function with the following penalty function (equation 4):

$$objective\_new = objective\_original * delPenalty^{numDels} \quad (4)$$

where *objective_new* is the new score of the objective function, *objective_original* is the original objective function (e.g., product yield), *delPenalty* is the deletion penalty, and *numDels* is the number of knockout reactions. This penalty ensures that designs with fewer knockouts will be selected over designs with similar phenotypes, but more knockouts. Fewer knockouts are desirable for ease of strain construction.

6.5.5 Theoretical analysis of the production potential in E. coli

The maximum production potential for *E. coli* was determined by: i.) defining a substrate condition, ii.) setting a minimum growth rate (μ) of 0.1 hr^{-1} (as set by the amount of flux necessary through the BOF_{CORE}) to simulate (at least) a minimal amount of growth, and iii.) maximizing the flux possible (i.e., the production rate) through the exchange reaction that correlated to each product analyzed using FBA. Computational minimal media[3] was used for the simulations with the exception of the main carbon substrates and the presence of oxygen as specified. An uptake value of 20 mmol gDW^{-1} hr^{-1} was used for the substrates and oxygen, when present.

6.5.6 Pre-processing of the model for computation

Before calculation of strain designs, the model was preprocessed under the following procedure after setting the primary carbon substrate and presence/absence of oxygen, in addition to the constraints necessary for computational minimal media

conditions[3]. Preprocessing was condition specific and was done for each culture condition examined. Model pre-processing was six step process (see **Figure 6.2**).

The goal of preprocessing was to eliminate certain reactions from the model's total reaction set to obtain a smaller set of selected reactions that could serve as valid targets for gene knockouts. First, all reactions that could not be utilized for a given condition, in other words those reactions that had maximum and minimum fluxes equal to zero, were removed. Furthermore, the upper and lower bounds were set to values that were potentially obtainable given the input conditions (instead of arbitrarily high and low values, respectively). This step generated the 'reduced' model. Next, to further narrow down the list of reactions to consider for removal, all reactions that had been experimentally found to be essential for growth were removed from consideration[53-55]. Also removed were reactions that were found to be computationally essential, such that when the reactions were knocked out, the growth was reduced to less than 5% of the wild type. Non-gene associated reactions as well as spontaneous and diffusion reactions were excluded due to the fact that biological knockouts of these reactions are impossible. Reactions from certain subsystems that were determined to be excluded were also removed, including cell envelope biosynthesis, glycerophospholipid metabolism, inorganic ion transport and metabolism, lipopolysaccharide biosynthesis and recycling, membrane lipid metabolism, murein biosynthesis, murein recycling, inner membrane transport, outer membrane transport, outer membrane porin transport, and tRNA charging. Reactions that act on molecules containing more than a certain number of carbons (7 carbons) were removed from the pool as they are unlikely to carry high flux in the production of the metabolite targets examined from the given substrates. Lastly, for coupled reactions, only one reaction per set was included, since knockouts are equivalent. Other reactions that were

manually removed included those that dealt with glycogen production, as glycogen is a generalized molecule in the network and those that contained thioredoxin and flavodoxin, as these compounds role are poorly characterized. By removing these reactions from the full model, the solution space that the OptKnock and OptGene algorithms search was significantly decreased, effectively reducing computational time.

Acknowledgements

Chapter 6, in full, is adapted from Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli* that is in preparation. The dissertation author was the primary author of this paper, which was co-authored by Daniel C. Zielinski, Jeff D. Orth, Jan Schellenberger, Dr. Markus J. Herrgård, and Dr. Bernhard Ø. Palsson.

References

1. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003;84:647-57.

2. Patil KR, Rocha I, Forster J, Nielsen J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 2005;6:308.

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

4. Atsumi S, Liao JC. Metabolic engineering for advanced biofuels production from Escherichia coli. *Curr Opin Biotechnol* 2008.

5. Chartrain M, Salmon PM, Robinson DK, Buckland BC. Metabolic engineering and directed evolution for the production of pharmaceuticals. *Current Opinion in Biotechnology* 2000;11:209-14.

6. Keasling JD, Chou H. Metabolic engineering delivers next-generation biofuels. *Nat Biotechnol* 2008;26:298-9.

7. Khosla C, Keasling JD. Metabolic engineering for drug discovery and development. *Nat Rev Drug Discov* 2003;2:1019-25.

8. Nakamura CE, Whited GM. Metabolic engineering for the microbial production of 1,3-propanediol. *Curr Opin Biotechnol* 2003;14:454-9.

9. Bailey JE. Toward a science of metabolic engineering. *Science* 1991;252:1668-75.

10. Park JH, Lee KH, Kim TY, Lee SY. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci U S A* 2007;104:7797-802.

11. Trinh CT, Carlson R, Wlaschin A, Srienc F. Design, construction and performance of the most efficient biomass producing E. coli bacterium. *Metab Eng* 2006;8:628-38.

12. Trinh CT, Unrean P, Srienc F. Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Appl Environ Microbiol* 2008;74:3634-43.

13. Lee KH, Park JH, Kim TY, Kim HU, Lee SY. Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* 2007;3:149.

14. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY. Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. *Appl Environ Microbiol* 2005;71:7880-7.

15. Altaras NE, Cameron DC. Metabolic engineering of a 1,2-propanediol pathway in Escherichia coli. *Appl Environ Microbiol* 1999;65:1180-5.

16. Zhang X, Jantama K, Moore JC, Shanmugam KT, Ingram LO. Production of L -alanine by metabolically engineered Escherichia coli. *Appl Microbiol Biotechnol* 2007;77:355-66.

17. Jaluria P, Chu C, Betenbaugh M, Shiloach J. Cells by design: a mini-review of targeting cell engineering using DNA microarrays. *Mol Biotechnol* 2008;39:105-11.

18. Santos CN, Stephanopoulos G. Combinatorial engineering of microbes for optimizing cellular phenotype. *Curr Opin Chem Biol* 2008;12:168-76.

19. Tyo KE, Alper HS, Stephanopoulos GN. Expanding the metabolic engineering toolbox: more options to engineer cells. *Trends Biotechnol* 2007;25:132-7.

20. Barrett CL, Kim TY, Kim HU, Palsson BO, Lee SY. Systems biology as a foundation for genome-scale synthetic biology. *Curr Opin Biotechnol* 2006;17:488-492.

21. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* 2007;104:1777-82.

22. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microbial organisms. *Nat Rev Microbiol* 2008;Accepted.

23. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

24. Reed JL, Palsson BO. Thirteen Years of Building Constraint-Based In Silico Models of *Escherichia coli*. *J Bacteriol* 2003;185:2692-9.

25. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology* 2003;4:R54.1-R54.12.

26. Zhang Y. Structural genomics of the Thermotoga maritima sets the stage for the molecular level analysis of its central metabolism. *In preparation* 2008.

27. Price ND, Papin JA, Schilling CH, Palsson B. Genome-scale microbial in silico models: the constraints-based approach. *Trends in Biotechnology* 2003;21:162-169.

28. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 2008;26:659-667.

29. Kim HU, Kim TY, Lee SY. Metabolic flux analysis and metabolic engineering of microorganisms. *Molecular BioSystems* 2008;4:113-120.

30. Kim TY, Sohn SB, Kim HU, Lee SY. Strategies for systems-level metabolic engineering. *Biotechnol J* 2008;3:612-23.

31. Lee SY, Lee DY, Kim TY. Systems biotechnology for strain improvement. *Trends Biotechnol* 2005;23:349-58.

32. Park JH, Lee SY, Kim TY, Kim HU. Application of systems biology for bioprocess development. *Trends Biotechnol* 2008;26:404-12.

33. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91:643-8.

34. Fong SS, Nanchen A, Palsson BO, Sauer U. Latent pathway activation and increased pathway capacity enable Escherichia Coli adaptation to loss of key metabolic enzymes. *J Biol Chem* 2005.

35. Fong SS, Palsson BO. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat Genet* 2004;36:1056-58.

36. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186-9.

37. Pharkya P, Burgard AP, Maranas CD. Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol Bioeng* 2003;84:887-99.

38. Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res* 2004;14:2367-76.

39. Perlack RD, Wright LL, Graham RL, Stokes BJ, Erbach DC. Biomass as a feedstock for a bioenergy and bioproducts industry: The technical feasibility of a billion-ton annual supply. In: Energy USDo, ed.: Oak Ridge National Laboratory, Oak Ridge, TN, 2005.

40. Ma F, Hanna MA. Biodiesel production: a review. *Bioresource Technology* 1999;70:1-15.

41. Paster M, Pellegrino JL, Carole TM. Industrial Bioproducts: Today and Tomorrow: U.S. Department of Energy, 2003.

42. Top Value Added Chemicals from Biomass. In: Werpy T, Petersen G, eds.: U.S. Department of Energy, 2004.

43. Sauer M, Porro D, Mattanovich D, Branduardi P. Microbial production of organic acids: expanding the markets. *Trends Biotechnol* 2008;26:100-8.

44. Leuchtenberger W, Huthmacher K, Drauz K. Biotechnological production of amino acids and derivatives: current status and prospects. *Appl Microbiol Biotechnol* 2005;69:1-8.

45. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol* 1993;59:2465-73.

46. Dharmadi Y, Murarka A, Gonzalez R. Anaerobic fermentation of glycerol by Escherichia coli: a new platform for metabolic engineering. *Biotechnol Bioeng* 2006;94:821-9.

47. Murarka A, Dharmadi Y, Yazdani SS, Gonzalez R. Fermentative utilization of glycerol by Escherichia coli and its implications for the production of fuels and chemicals. *Appl Environ Microbiol* 2008;74:1124-35.

48. Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgard MJ, Palsson BO. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *In preparation* 2008.

49. Fong SS, Marciniak JY, Palsson BØ. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 Using a Genome-scale in silico Metabolic Model. *Journal of Bacteriology* 2003;185:6400-8.

50. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7:130-41.

51. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

52. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* 1994;60:3724-3731.

53. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2006;2:2006.0008.

54. Joyce AR, Palsson BO. Predicting gene essentiality using genome-scale in silico models. *Methods Mol Biol* 2008;416:433-57.

55. Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, Mori H, Lesely SA, Palsson BO, Agarwalla S. Experimental and Computational Assessment of Conditionally Essential Genes in *Escherichia coli*. *J Bacteriol* 2006;188:8259-8271.

Chapter 7

Model-driven metabolic engineering, Part 2: Construction and evolution of *E. coli* production strains designed through model-driven metabolic engineering

7.1 Abstract

Metabolic engineering is a growing field for which new methods are being developed to generate new and existing products more rapidly and efficiently. Two approaches that can aid this process are model-driven analysis based on constraint-based modeling and growth-coupled production, and adaptive evolution for strain optimization. We combined these two approaches by selecting strains identified from a model-driven analysis of the production of native *E. coli* metabolites[1], evolving them using adaptive evolution, and characterizing their production capabilities. Two out of the three strains selected during this analysis resulted in their predicted production phenotypes with substrate yields of 0.98 g lactate g^{-1} glucose and 0.84 g lactate g^{-1} xylose. The third strain failed to grow after construction. The strains displayed volumetric productivities of 1.7 g lactate L^{-1} hr^{-1} and 0.13 g lactate L^{-1} hr^{-1} at low cellular densities on glucose and xylose, respectively. The optimization process of adaptive evolution resulted in a significant increase in growth rate for both strains and a significant increase in substrate consumption rate for the glucose consuming strain.

These results demonstrate that a model-driven analysis can be used to design successful productions strains and adaptive evolution can be used as a design principle.

7.2 Introduction

Through the process of metabolic engineering, microbial organisms have been engineered for the production of various desirable compounds. Such engineered microbial strains and their products will become increasingly important as they can produce chemicals from renewable feedstocks rather than from nonrenewable petroleum. Metabolic engineering has been practiced for many years, and the traditional approaches typically rely on strategies such as random mutagenesis with selection for an over-producer or rational design of over-expression of genes either directly responsible for secondary metabolite production or genes indirectly involved in increasing metabolite production[2-4]. Some products that have been derived from metabolically engineered cells that are currently of interest include: succinate, acetate, citrate, 1,3-propanediol, 1,2-propanediol, lactic acid, 3-hydroxypropionic acid (3HP), ethanol, butanol, and more complex pharmaceutically relevant compounds[5-8]. Advances in the field of systems biology are now being used in new metabolic engineering strategies[9]. Traditional metabolic engineering is challenged by the necessity of a high level of organism familiarity and biological intuition required to make successful production strains. However, with systems biology methods and genome-scale metabolic models, it is possible to reliably and systematically predict the phenotype of microbial organisms. This allows for strain designs that are often non-intuitive and non-obvious. Successful cases of systems biology driven strain designs for *E. coli* include production of the metabolites

lycopene[10,11], lactic acid[12], succinic acid[13-15], and amino acids[16,17]. These success stories demonstrated in *E. coli* as well as those for additional organisms have been recently reviewed[9,18] and these cases represent the earliest examples of metabolic engineering driven by systems biology. Furthermore, it is apparent that to fully demonstrate the utility of these approaches, experimental validation of computationally designed strains is necessary.

An exceptionally promising systems biology approach to metabolic engineering is the concept of growth coupled design, in which the production of a metabolite by a microbial strain increases as growth rate increases[19]. Conventional strain designs rely on genetic manipulations that alter the metabolism in a way that typically redirects metabolic flux from producing biomass to producing a specific desired product. Normally, such strains are highly unstable, and if left unsupported, mutations that increase growth rate at the cost of production will occur and the strain will lose productivity over time. It is unusual for the increased secretion of a metabolic by-product to increase growth rate, so growth-coupled strains are non-intuitive. However, genome-scale metabolic models and constraint-based analysis methods are able to predict genetic manipulations that couple production objectives to a selection pressure (i.e., growth rate). These models can be used to predict which genetic modifications lead to growth coupling of metabolites and the key implementation and strain optimization procedure of laboratory adaptive evolution can then be used as a tool to optimize *in vivo* strain designs. The first systems biology algorithm to design growth coupled strains was OptKnock, a bi-level mixed-integer optimization algorithm that systematically searches for sets of gene deletions that lead to growth coupled production of a desired metabolite[19]. This algorithm is guaranteed to find an optimal solution, but it may take significant time to do so.

Another design algorithm that can be utilized for designing growth-coupled strains is OptGene, which uses a genetic algorithm to find sets of knockouts[20]. OptGene is capable of utilizing nonlinear objective functions, but it may not identify globally optimal solutions. Several OptKnock designs for the production of lactic acid were constructed and adaptively evolved *in vivo*, and it was found that the experimental results closely agreed with the computational predictions[12]. More recently, a thorough screen of the growth-coupled design potential of *E. coli* was conducted utilizing the genome-scale metabolic reconstruction and model *i*AF1260 and the aforementioned design algorithms[1]. In this study, a number of central metabolism compounds were selected, and growth coupled designs for these compounds were identified using both OptKnock and OptGene for different desirable production characteristics. In the present study, we have validated the computationally driven growth-coupled strain design process through the construction of several of these strain designs *in vivo*, their optimization through the adaptive evolution process, and characterization of their production phenotype.

Strain designs for the production of lactic acid and L-alanine from the sugar feedstocks glucose and xylose have been selected to demonstrate the approach of growth coupled design followed by adaptive evolution. These designs were chosen based on computationally predicted properties such as high predicted product yields, ability to produce and secrete only one compound (homofermentation), and use of characterized metabolic pathways to produce the targeted products. Lactic acid and L-alanine are also industrially relevant chemicals with many practical uses. Lactic acid and many of its derivatives are used extensively in the food and beverage industry as an acidulant or as a preservative, and polylactic acid (PLA) is a biodegradable plastic. Other important products derived from lactic acid are ethyl lactic acid, acrylic acid,

and propylene glycol. Most lactic acid is produced by industrial fermentation from glucose and carbohydrate sources[8,21]. As an example, over 150,000 tons of polylactic acid are produced every year at the NatureWorks PLA manufacturing plant in Blair, NE[22]. As the demand for biodegradable polymers for packaging increases, the production of lactic acid for PLA is expected to increase to over 4 million tons by 2020[8]. Several studies have demonstrated the feasibility of using *E. coli* to produce lactic acid[12,23-26]. A strain that produces L-alanine by fermentation also has potential for industrial use, and several recent studies demonstrate that it is feasible in *E. coli*[27-29]. L-alanine and other amino acids are commonly used as food additives and in pharmaceuticals. Amino acids have the fastest growing market volume of all fermentation products, with a value of about $3.5 billion in 2004 that is expected to increase to $5 billion in 2009[30].

Herein, we describe the experimental implementation of strains designed computationally and validated to make useful production strains. We describe the selection criteria for the constructed strains, a sensitivity analysis used to minimize genetic interventions (i.e., knockouts), and construction of the strains. An adaptive evolution to remove strain auxotrophy is described along with the evolution to a production phenotype. Lastly, final endpoint strain characterization is presented along with integration and comparison to modeling. This work demonstrates how production growth-coupled production strains can be computationally designed and implemented.

7.3 Results

7.3.1 Selection of strain designs

The strains constructed in this analysis were selected from a computational analysis that identified designs for which metabolite production could be coupled to growth in *E coli*[1]. The process of selecting strain designs from the design pool of the computational screen for experimental implementation is given in **Figure 7.1**. From the pool of computationally designed strains, three different designs were chosen for growth on two different substrates for construction *in vivo* (**Table 7.1**). These three strains were designed to produce, i) D-lactic acid from D-glucose, ii) D-lactic acid from D-xylose, and iii) L-alanine from D-glucose. Herein, these metabolites will also be referred to as lactate, glucose, xylose, and L-alanine. The three strains designs were chosen because each were homofermentative designs for which the target product was the only predicted product (i.e., no other predicted carbon containing products greater than 2% wt% of the yield). Additionally, these designs also possessed high substrate yields (see **Table 7.1**). Finally, when the predicted pathways which were necessary to produce this high-production phenotype were compared to wild type simulations under similar conditions, the pathways were analyzed and predicted to be active therefore providing confidence in the ability of the strain to reach the production phenotype. The production envelopes for each strain are given in **Figure 7.2**; these envelopes detail the computationally predicted production rates of the strains as a function of growth rate. The selected strain designs are detailed herein.

Table 7.1**: Properties of the strain designs constructed.**[a]

Strain	Substrate	Production Target	No. of operon deletions	No. of genes deleted	No. of reactions eliminated	Aerobicity	Predicted Yield	Production Rate	Predicted growth rate	Byproduct	Byproduct Rate	Byproduct Yield
							wt %	mmol gDW-1 hr-1	hr-1		mmol gDW-1 hr-1	wt %
BOP338	Glucose	D-Lactic Acid	3	7	2	Anaerobic	88.7%	35.46	0.308	succinate	0.103	0.3%
										CO2	0.603	0.9%
BOP374	Xylose	D-Lactic Acid	6	16	5	Anaerobic	91.4%	30.47	0.194	succinate	0.065	0.2%
										CO2	0.38	0.6%
BOP360	Glucose	L-Alanine	5	9	4	Anaerobic	92.8%	37.54	0.122	succinate	0.127	0.4%
										pyruvate	0.671	1.6%
BOP368	Glucose	L-Alanine	7	11	5	Anaerobic	92.9%	37.59	0.120	succinate	0.125	0.4%
										pyruvate	0.657	1.6%

[a] simulations ran with 20 mmol gDW^{-1} hr^{-1} uptake rate of substrate and minimal medium conditions

7.3.2 Properties of selected strain designs

Each strain design selected was analyzed prior to construction and experimental implementation given the process outlined in **Figure 7.1**. Each strain design process possessed different properties and the resulting approach to each was therefore tailored to each strain. The full set of reaction abbreviations for the model are contained in the *i*AF1260 reconstruction[31] and those used here are, ALCD2x (alcohol dehydrogenase (ethanol)), FBA (fructose-bisphosphate aldolase), FUM (fumarase), GLUDy (glutamate dehydrogenase (NADP)), LDH_D (D-lactate dehydrogenase), MDH (malate dehydrogenase), PDH (pyruvate dehydrogenase), PFK (phosphofructokinase), and PFL (pyruvate formate lyase).

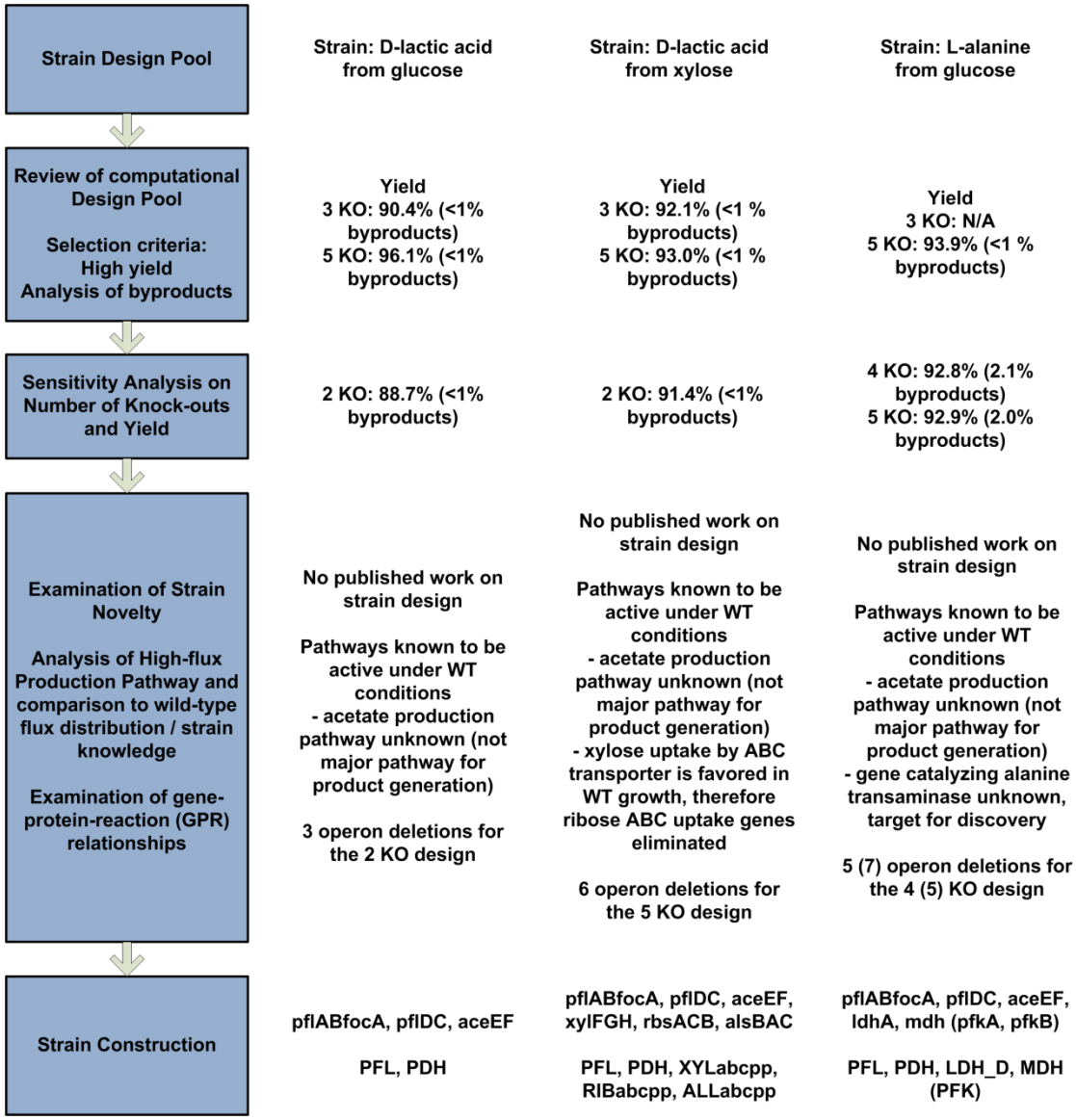


Figure 7.1: Process used to select strains for construction and experimental implementation.

7.3.3 D-lactate production strain from glucose

From the computational pool of designs[1], the production of D-lactate from glucose was identified as a production target as it was predicted to be a homofermentation product with a high yield. Additionally, it was a good initial target as lactic acid production from glucose has been demonstrated in *E. coli* previously[12,23-26].

The original designs generated from the computational analysis gave two designs of three and five reaction knockouts that could result in the high yield phenotype (**Table 7.2**). A sensitivity analysis was performed on each of the designs to determine if the high yield could be sustained or improved with less or more metabolic interventions (i.e., knockouts). This analysis was performed using the 'analyzeGCdesign' algorithm with a penalty for knockouts of 90% (see Methods). A plot of number of knockouts and the resulting acceptance in objective function drop is given in **Appendix B** for this analysis. The sensitivity analysis returned a two reaction design (reactions PFL and PDH) as the optimal design for the given knockout penalty examining maximum yield starting from the three knockout design, and a single reaction removal (ALCD2x) starting from the five knockout design. It should be noted that the sensitivity analysis results will vary when the knockout penalty and design objective[1] is changed, therefore multiple solutions should be considered.

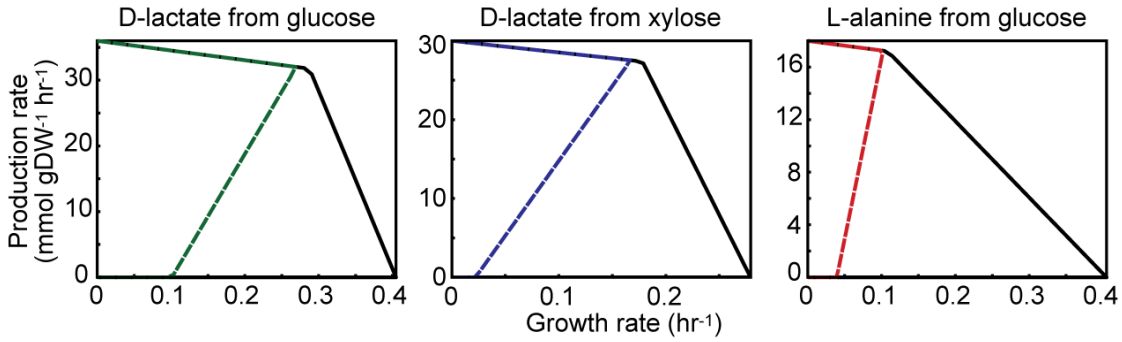


Figure 7.2: The production envelopes predicted for the constructed strains. The predicted production envelopes for the three strains constructed for this study. The target product and substrate are listed for each. These production envelopes were determined to be superior designs out of a pool of designs that was computationally generated. The substrate uptake rate in each of the plots is 18 mmol gDW^{-1} hr^{-1} (a typical wild-type anaerobic uptake rate) and minimal medium conditions were used.

Table 7.2: Results of computational analysis of strain designs.

Substrate	Production Target	3KO design	3KO production rate	3KO yield	3KO byproducts	3KO Byproduct Rate	3KO Byproduct Yield	3KO growth rate	Max Yield - 90% KO penalty - 3KO design
			mmol gDW-1 hr-1	wt %		mmol gDW-1 hr-1	wt %	hr-1	
		OptKnock	OptKnock	OptKnock	OptKnock	OptKnock	OptKnock	OptKnock	analyzeGCdesign
Glucose	D-Lactic Acid	PDH, PFL, PGI	36.158	90.4%	succinate	0.087	0.3%	0.261	PFL, PDH
					CO2	0.51	0.6%		
Xylose	D-Lactic Acid	PDH, PFL, GLUDy	30.690	92.1%	succinate	0.06	0.2%	0.179	PFL, PDH
					CO2	0.351	0.5%		
Substrate	**Production Target**	**5KO design**	**5KO production rate**	**5KO yield**	**5KO byproducts**	**5KO Byproduct Rate**	**5KO Byproduct Yield**	**5KO growth rate**	**Max Yield - 90% KO penalty - 5KO design**
Glucose	D-Lactic Acid	ALCD2x, ATPS4rpp, G6PDH2r, GHMT2r, PGI	38.429	96.1%	acetate	0.138	0.2%	0.100	ALCD2x
					succinate	0.033	0.1%		
					CO2	0.196	0.2%		
Xylose	D-Lactic Acid	GLUDy, PDH, PFL, PGCD, THD2pp	30.984	93.0%	succinate	5.3%	0.2%	0.159	PFL, PDH
					CO2	0.312	0.5%		
Glucose	L-Alanine	ACALD, ACKr, ALCD2x, LDH_D, FBA	37.924	93.8%	acetate	0.0754	0.1%	0.130	ACALD, ACKr, ALCD2x, LDH_D, FBA
					formate	0.6261	0.8%		
					succinate	0.0433	0.1%		

KO – knockout

The strain designs returned from the sensitivity analysis and design algorithms were analyzed to examine strain novelty, to determine the active high-flux pathways necessary to produce the predicted production phenotype, and to determine the gene knockouts that were necessary to remove these reactions from *E. coli*. The PFL and PDH design for the production of lactic acid from glucose was found to be a novel design. Although similar designs exist[25,26]. Particularly, the pyruvate dehydrogenase (PDH) knockout has not been a target in purely anaerobic production strains as it is thought to be inactive anaerobically despite evidence that suggests otherwise[32]. Even though this design will be grown anaerobically where levels of PDH activity are presumably lower, the use of adaptive evolution might increase this activity, therefore, its encoding genes were removed for this design. For the analysis of predicted high-flux pathways (i.e., pathways that carried 10% or more of the input flux value, see Methods), there were no pathways which caused suspicion about their potential of being able to be carry high-flux under glucose anaerobic conditions. The high-flux reactions were essentially glucose uptake through the PTS system, glycolysis, and the D-lactate dehydrogenase reaction. Aside from these characterized anaerobic

pathways, an interesting aspect of this strain is that the pathway to generate acetate (and acetyl-coenzyme A) in such a genotype is unknown, and the model predicts potential pathways for its generation that can be further investigate in the end-point production strain. Acetyl-coenzyme A (acetyl-CoA) is necessary for growth as it is an important building block of cellular lipids. Lastly, three operons were necessary for removal to eliminate the activity of the reactions in the design. Two segments for the PFL reaction; *pflABfocA*, the main isozyme and additionally the transporter that allows passage of the reaction byproduct formate, and *pflDC*, the minor isozyme. One segment was necessary to remove the PDH reaction, *aceEF*, encoding the core of the pyruvate dehydrogenase catalyzing enzyme (see **Figure 7.1** and **Table 7.3**). These removals were defined in the gene to protein to reaction associations in *i*AF1260[31]. This strain was labeled BOP338 (see **Figure 7.1**, **Table 7.1**, and **Table 7.3**).

7.3.4 D-lactic acid production strain from xylose

Similar to production with glucose, production of D-lactate from xylose could be achieved homofermentatively at a yield of over 92% for a three or greater knockout design. However, different from production on glucose, demonstration of lactic acid production in *E. coli* from xylose is less established[24] with a majority of production attributed to other organisms[33].The three and five knockout design yields are given in **Table 7.2**. Furthermore, a sensitivity analysis using a 90% penalty on knockouts resulted in the same set of two reactions to be removed when considering the optimal three and five reaction knockout designs as inputs; the PFL and PDH reactions. Additionally, this design was also a novel design to produce lactic acid from

xylose. Given these findings, this design was chosen for construction for production of lactic acid from xylose.

Different from the analysis of the production phenotype on glucose, the high-flux pathways for production of D-lactate from xylose indicated that an optimal phenotype required a pathway different from that observed in wild-type growth. This specific difference was that the high-production phenotype required uptake of xylose through the xylose major facilitator superfamily transporter (encoded by the *xylE* gene, a proton symporter[34]) and not through the ATP-utilizing ABC transport system (encoded by the *xylFGH* genes[35]) in *E. coli*. Thus, this transport difference results in a significant difference in energy required for transport (i.e., ATP) where the symport requirement allows for a higher production phenotype and growth rate as less energy is diverted to xylose transport. Despite this, it has been suggested that transport of through the ABC system is the major means of transport for xylose into the cell, both aerobically and anaerobically, and that the *plf* genes are essential for growth on xylose[36]. However, as the model predicts growth and production with removal of the ABC transport method and the PFL and PDH genes, they were targeted for elimination. Because of the transport requirements, the *xylFGH* along with the other ribose ABC transporters (encoded by the *rbsACB* and *alsBAC* genes) were removed from the genome, as transporters often transport multiple similar metabolites (e.g., the xylose transporter has been shown to transport ribose[37]). The resulting high-flux pathways in this phenotype include the xylose degradation, pentose phosphate, and glycolytic pathways along with the D-lactate dehydrogenase reaction. Similar to the glucose design, acetate production is also uncharacterized in this strain. The genes that were removed from this strain, labeled BOP374, are given in **Figure 7.1**, **Table 7.1**, and **Table 7.3**.

7.3.5 L-alanine production strain from glucose

Several different L-alanine production strains were identified from the pool of computational designs which possessed desirable production characteristics. The optimal five knockout design is given in **Table 7.2**. Despite the high yield five knockout design, there were no three knockout growth-coupled designs. Multiple knockout designs were identified through a sensitivity analysis using the five knockout strain given in **Table 7.2** (from OptKnock) utilizing several different objective functions. By maximizing yield and substrate specific productivity with a 90% deletion penalty, the knockout reactions PFL, PDH, LDH_D, FUM and PFL, PDH, LDH_D, FBA were identified as growth coupled strains, respectively. Further computational analysis was performed, replacing each of these knockouts with other reactions individually to find equivalent knockout sets. Thought this, it was found that the reaction MDH could be knocked out instead of FUM or FBA in the design, resulting in a strain with equivalent production characteristics. The MDH reaction was a superior choice for *in vivo* knockout because only the gene, *mdh,* needed to be knocked out to eliminate this reaction. Whereas, in *E. coli,* there are three isozymes for the FUM reaction, *fumA*, *fumB*, *fumC* and also three for the FBA reaction, *fbaA, fbaB, chbG*; all three of the isozymes would need to be knocked out to deactivate either reaction.

Further analysis was performed to analyze an L-alanine production strain. Flux balance analysis was used to predict the high flux pathways of this strain and found that most glucose is converted to pyruvate by glycolysis and then to L-alanine using L-alanine transaminase. The L-alanine transaminase reaction has been shown to be present in *E. coli* in cell free extract analysis, however the catalyzing gene is unknown[38,39], and therefore provides an opportunity for characterization if upregulated and active in the final strain design. The reaction GLUDy (glutamate dehydrogenase)

also carries a very high predicted flux, converting α-ketoglutarate to L-glutamate for production of L-alanine by transamination of pyruvate. This strain also secretes small amounts of pyruvate and succinate optimally (approximately 2% of the total yield). A total of five operons were knocked out to produce this strain *in vivo*, including the *pflABfocA*, *aceEF*, and *pflDC* operons that were knocked out in the lactate producing strains. The gene *ldhA* was knocked out to eliminate the LDH_D reaction. There is a known LDH_D isozyme, *dld*, but it was not knocked out because is predicted to be active only during respiration[40]. The gene *mdh* was also knocked out, eliminating the MDH reaction. This strain was labeled BOP360. After deciding on the construction of the BOP360 strain, further analysis was performed to determine if additional reactions could be eliminated from strain BOP360 to strengthen the extent of growth coupling[1] (see also Methods). This analysis identified the reaction PFK for elimination, which also was predicted to have a slightly higher yield than BOP360. Therefore the design PFL, PDH, LDH_D, MDH, and PFK, incorporating the PFK knockout was generated. To construct this strain, the genes *pfkA* and *pfkB*, which code for PFK isozymes, were knocked out of BOP360. This new strain was named BOP368 (see **Figure 7.1**, **Table 7.1**, and **Table 7.3**).

7.3.6 Strain construction

The strains that were generated during the course of the project are listed in **Table 7.3** and **Figure 7.3**. Each strain was given a 'BOP' tag and number under the protocol used in our lab. The starting strain for each design was wild-type *E. coli* strain K12 MG1655 (ATCC 700926). We have previously characterized this strain extensively physiologically[12,41-43], resequenced its genotype[44], and the computational model *i*AF1260 is based on the K-12 MG1655 genome. Gene disruptions were

performed using homologous recombination of PCR-amplified linear fragments[45] (see Methods).

Table 7.3: Strains that were constructed for this project.

Strain	parent	evolution	genotype
BOP27	N/A		MG1655 ATCC#47076
BOP328	BOP27		*ldhA kan+*
BOP330	BOP27		*pflABfocA kan+*
BOP332	BOP328		*ldhA pflABfocA kan+*
BOP334	BOP330		*pflABfocA ldhA kan+*
BOP336	BOP330		*pflABfocA pflDC kan+*
BOP338	BOP336		*pflABfocA pflDC aceEF kan+*
BOP340	BOP334		*pflABfocA ldhA aceEF kan+*
BOP352	BOP338		*pflABfocA pflDC aceEF mdh kan+*
BOP354	BOP340		*pflABfocA ldhA aceEF mdh kan+*
BOP356	BOP340		*pflABfocA ldhA aceEF pflDC kan+*
BOP360	BOP354		*pflABfocA ldhA aceEF mdh pflDC kan+*
BOP364	BOP360		*pflABfocA ldhA aceEF mdh pflDC pfkA kan+*
BOP368	BOP364		*pflABfocA ldhA aceEF mdh pflDC pfkA pfkB kan+*
BOP370	BOP338		*pflABfocA pflDC aceEF xylFGH kan+*
BOP372	BOP370		*pflABfocA pflDC aceEF xylFGH rbsACB kan+*
BOP374	BOP372		*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC kan+*
BOP374e	BOP374	SS, 16 days	*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC kan+*
BOP376	BOP360		*pflABfocA ldhA aceEF mdh pflDC mgsA kan+*
BOP376e	BOP376	SS, 12 days	*pflABfocA ldhA aceEF mdh pflDC mgsA kan+*
BOP384	BOP374e		*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC mutS kan+*
BOP386	BOP376e		*pflABfocA ldhA aceEF mdh pflDC mgsA mutS kan+*
BOP392	BOP368		*pflABfocA ldhA aceEF mdh pflDC pfkA pfkB mutS kan+*
BOP384eG1	BOP384	SPE, 8.5 days	*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC mutS kan+*
BOP384eG2	BOP384	SPE, 8.5 days	*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC mutS kan+*
BOP384eX1	BOP384	SPE, 8.5 days	*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC mutS kan+*
BOP384eX2	BOP384	SPE, 8.5 days	*pflABfocA pflDC aceEF xylFGH rbsACB alsBAC mutS kan+*

SS – steady-state evolution, SPE – serial passage exponential

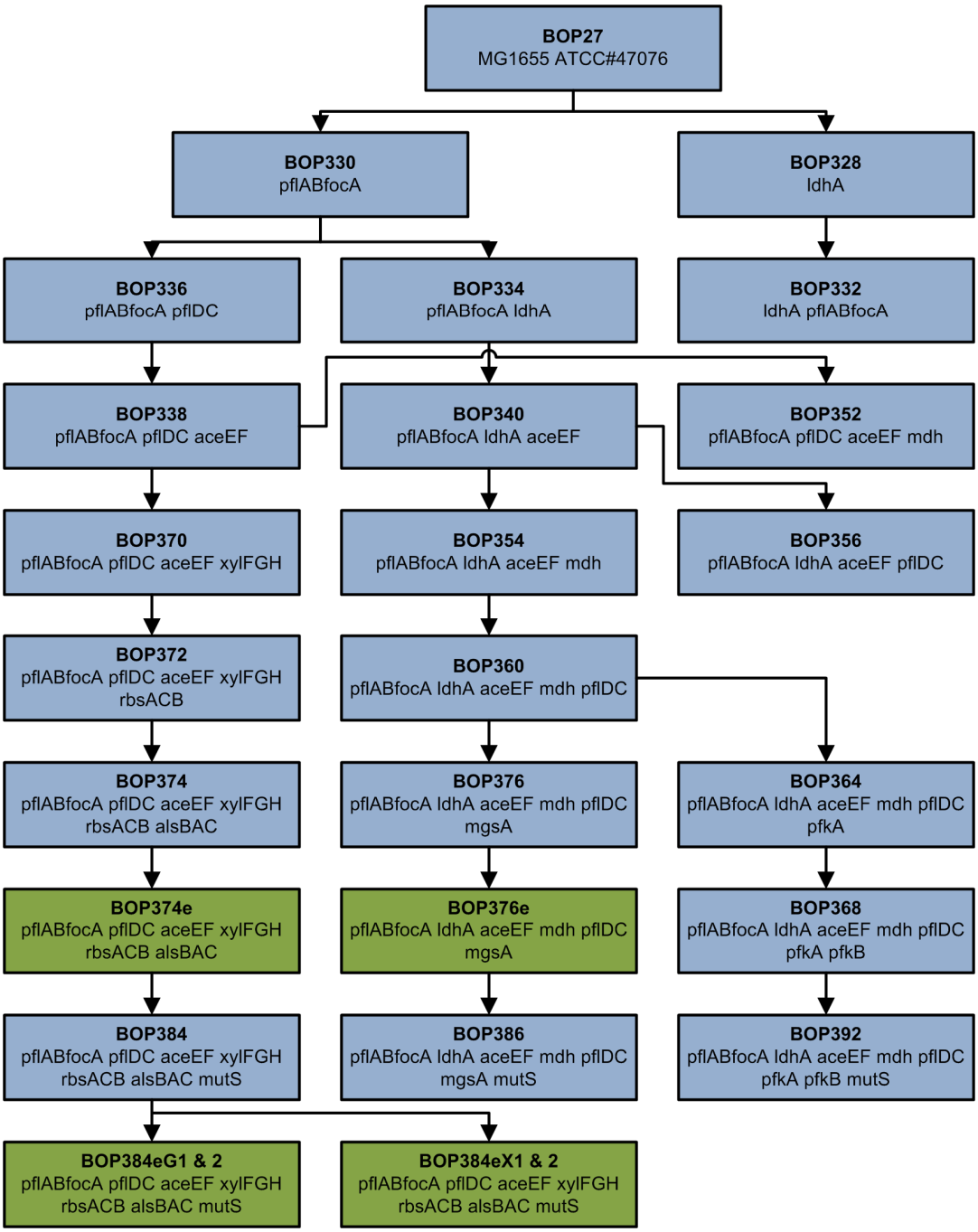


Figure 7.3: *E. coli* strains constructed. The *E. coli* strains generated for the project and the lineage for the construction of each. A blue box indicates a strain that has been generated using a gene deletion procedure[45], a green box indicates a strain generated from an adaptive evolution procedure.

7.3.7 Initial Strain Characterization

After construction of the production strains, each was analyzed for growth properties under the minimal medium conditions for which they were designed to be evolved and optimized in (see Methods and **Appendix B** for medium compositions). This was done in batch mode and the results are shown in **Table 7.4**. This screen revealed that the lactate and L-alanine productions strains constructed were auxotrophic for acetate and it was necessary to supply it to the medium for growth. This indicates that the strains were not able to make the necessary acetyl-CoA required for generation of lipids. Additionally, with the stains sharing similar PFL and PDH knockouts (the characterized pathways to generate acetate from the glycolytic pathway), this result was not entirely unexpected. The growth rates shown in **Table 7.4** are for strains that have not been subjected to adaptive evolution. These rates could potentially increase with adaptation to these conditions, as was observed with the L-alanine production strain on MES1 medium with acetate supplementation when the growth rate approximately doubled after several doublings (data not shown). Furthermore, supplementation with yeast extract, a widely used culture supplement (see below) would also increase the growth rate in a few instances examined, but could not support growth solely as a supplement with glucose or xylose without acetate (data not shown).

The requirement of necessary acetate supplementation for the L-alanine and lactate production strains prompted an adaptive evolution process to remove their auxotrophy (see below). For the L-alanine production strain from glucose, removal of acetate was necessary as the predicted optimal growth phenotype with acetate supplemented medium displayed a different phenotype than the designed envelope (**Figure 7.2**). Namely, the production of ethanol was predicted to occur with an

acetate supplemented medium. Although, the optimal phenotype was not affected for the lactate production strains with acetate addition to medium, it was determined desirable to alleviate acetate supplementation. This was due to determine if the modeling predictions of not requiring acetate in culture medium were accurate, to prevent the possible necessity for separating excess acetate from the homofermented product, and because previous work demonstrated that excess acetate can lead to byproduct formation, namely succinate[26]. Additionally, an undesirable byproduct could be formed from the additional supplementation, such as ethanol as identified examining the L-alanine production strain.

Table 7.4: Initial growth characteristics of production strains.

Strain	Culture Conditions	Supplement	Aerobicity	μ (hr-1)
BOP338	4 g/L glucose M9	none	aerobic	NG
BOP338	4 g/L glucose M9	1 g/L acetate	aerobic	0.37
BOP374	4 g/L glucose M9	none	aerobic	NG
BOP374	4 g/L glucose M9	1 g/L acetate	aerobic	0.43
BOP374	4 g/L glucose M9	none	anaerobic	NG
BOP374	4 g/L glucose M9	1 g/L acetate	anaerobic	0.16
BOP374	4 g/L xylose M9	none	anaerobic	NG
BOP374	4 g/L xylose M9	1 g/L acetate	anaerobic	0.01
BOP360	4 g/L glucose MES1	none	anaerobic	NG
BOP360	4 g/L glucose MES1	1 g/L acetate	anaerobic	0.05
BOP376	4 g/L glucose MES1	none	anaerobic	NG
BOP376	4 g/L glucose MES1	1 g/L acetate	anaerobic	0.02

NG – no growth

7.3.8 Removal of the mgsA gene

Given that the L-alanine production strain did not grow to a high density and consume all of the substrate available without a major change in medium composition (such as pH) before entering the stationary and death phase in batch, it was suspected that a toxic intermediate was accumulating. It was suspected that a toxic accumulation of methylglyoxal could be causing the premature entry into stationary

phase[46]. To alleviate this possibility, the *mgsA* gene (encoding methylglyoxal synthase) was removed in strain BOP360 to generate BOP376. BOP376 was then analyzed during anaerobic growth. BOP376 behaved similarly to BOP360 and it was concluded that methylglyoxal accumulation was not the cause of the low final density of the cells. However, as this strain performed similarly and given that the removal of the methylglyoxal synthase could be beneficial to avoid potential complications with this toxicity later in the project, BOP376 was used for weaning off the strain from acetate auxotrophy.

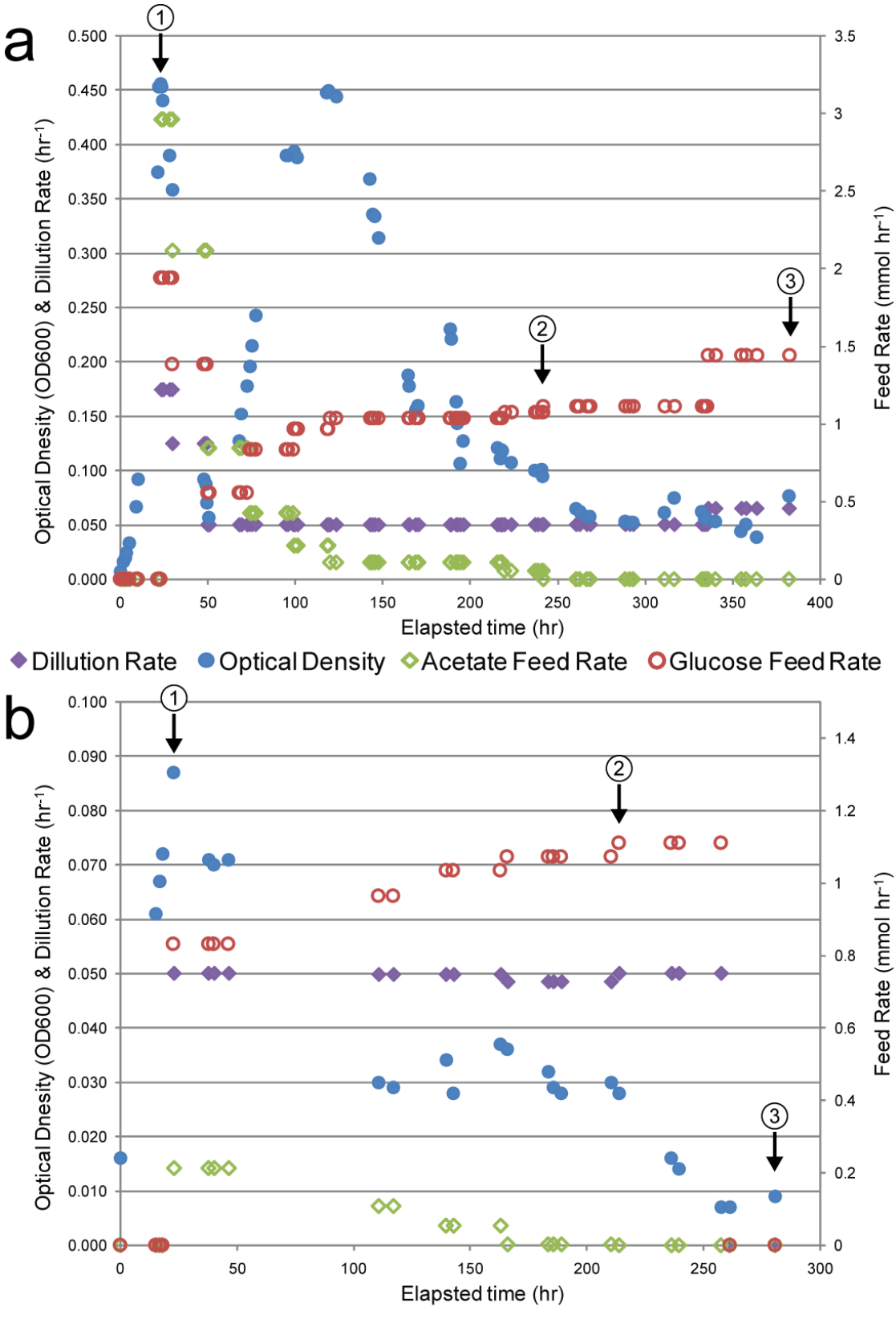
7.3.9 Chemostat evolution to remove auxotrophy

As determined from the initial strain characterization, the lactate and L-alanine production strains required acetate for growth. Therefore, it was decided that evolution would be used to attempt to alleviate this auxotrophy. To minimize the number of strains that needed to be evolved to remove acetate auxotrophy, one strain, BOP374, was chosen going forward as the design to be used for lactate production. This strain would be used for production from both glucose and xylose as the designs for each were very similar. Specifically, the additional gene deletions that were incorporated into the xylose utilizing strain were determined to have no effect on the growth rate from glucose, **Table 7.4**. BOP374 (the strain with additional xylose uptake gene knockouts) actually even possessed a higher growth rate on glucose than BOP338 when examined. Additionally, glucose was chosen as the substrate for evolution as the growth rate for this strain on this substrate was higher (**Table 7.4**) and thus, the evolution time would be shorter.

To remove the auxotrophy of acetate, BOP374 and BOP376 were evolved in 1.0 L chemostats anaerobically with a stepwise decreasing acetate feed rate (**Figure**

7.4). To do this, each strain was initially inoculated into the fermentor in a batch mode with a glucose and acetate mixture, after the culture grew up to an appreciable density (point 1 on the plots), the culture was run in continuous culture mode during which the acetate feed rate was sequentially lowered until the strain was growing solely on glucose minimal medium (point 2 on the plots). At the end of the evolution, a colony was isolated from the fermentor and was designated and preserved with a new strain number (point 3 on the plots). Continuous culture was chosen for the evolution as it allowed for automation and monitoring of the process along with a straightforward procedure to drop auxotrophy feed rates. It could also be used since selection of the fastest growing strain was not necessary for this step as a strain that could grow without acetate was the objective.

Figure 7.4: Adaptive evolution of production strains to remove auxotrophy. (facing page) The continuous culture adaptive evolution of strains BOP374 and BOP376 to remove acetate auxotrophy. Plot **a** shows the evolution of the lactate production strain, BOP374. Plot **b** shows the evolution of the L-alanine production strain, BOP376. Three different time points are denoted, (1) the time at which continuous culture was initiated, (2) the point at which acetate feed to the culture was ended, and (3) the time at which a strain was collected. The dilution rate (hr-1), optical density (OD600), and feed rates for glucose and acetate (mmol hr-1) are given. For both strains, as the acetate feed was decreased, the cellular density decreased. At point 3, both evolutions resulted in isolation of a strain that could grow solely on glucose.


a
1
2
3
Optical Dnesity (OD600) & Dillution Rate (hr-1)
Feed Rate (mmol hr-1)
Elapsted time (hr)
Dillution Rate
Optical Density
Acetate Feed Rate
Glucose Feed Rate
b
1
2
3
Optical Dnesity (OD600) & Dillution Rate (hr-1)
Feed Rate (mmol hr-1)
Elapsted time (hr)

The data from the evolutions indicate that the strains grew better in the presence of acetate, but at the end of evolution, the strains could survive and grow continually for a number of generations without acetate supplementation. For strain BOP374, **Figure 7.4a**, an initial batch phase of growth (from the beginning until point 1) was ran until the culture reached a relatively high density. This could be due to the presence of acetate in the medium, initially 1.0 g/L, or potentially a small amount of oxygen or rich medium (from inoculum) being present initially. However, when continuous culture was initiated, the density dropped sharply. Although, this drop could be a result of cell washout, at the time of the drop, the anaerobic gas (95% N2, 5% CO2, see Methods) supplied to the fermentors to maintain anaerobicity was depleted and before reinitiating of sparging causing an inconsistent gassing rate to the fermentor. This could have affected the growth rate of the culture. The same issue was again encountered around 180 hours as the gas feed was depleted and was inconsistent for a short period of time (approximately 5 hours). Nonetheless, the overall dilution rate was dropped at the elapsed time point of approximately 50 hours in the evolution to prevent cell washout. At point 2 on the plot (just over 245 hours into the evolution), the acetate feed rate was changed to zero and the strain maintained a steady density until the evolution was ended and a colony was isolated from a sample (point 3 on the plot). The total number of doublings that the culture underwent during continuous culture when just supplied with glucose as a feed was 11 and this corresponds to 1.1×10^{12} duplication events factoring in the cell concentration. At this point in time, it was determined that the strain had evolved and was present that could grow solely on glucose. HPLC analysis confirmed that there was no acetate in the culture medium shortly after the acetate feed was stopped. The hydraulic retention time at the given dilution rate is 20 hours and the length of time

the culture was sustained with no acetate feed was over 140 hours. The total number of doublings for the entire evolution was 30 and the total number of duplication events was 8.6×10^{12}. The evolved strain harvested at the end of the evolution was confirmed to have the same genotype as the starting strain (in terms of gene knockouts) and was designated BOP374e.

To alleviate acetate auxotrophy from the L-alanine production strain, BOP376 was also evolved anaerobically in a continuous culture evolution. Data from the evolution is shown in **Figure 7.4b**. Similarly, the culture was initially grown in batch mode to reach an appreciable density until continuous culture was initiated (point 1 on the plot). For this evolution, the dilution rate was kept constant at approximately 0.5 hr-1 for the entire experiment. At the time point of approximately 50 hours elapsed time, there was a failure in the probe that controlled the level of the culture in the vessel. This caused the culture volume to drop from the desired level of 1.0 L to approximately 100 mL. Because of this, the feed rates at this time were both increased proportionally until the level was brought up to 1.0 L and the culture was operating in batch model until approximately 100 hours elapsed time when continuous culture was restarted. The cellular density during this time dropped to approximately half of the level before the temporary probe failure. When continuous culture was again established, the acetate feed rate was dropped until it was no linger fed into the culture (point 2 on the plot). At this point, a drop in cell density was observed due to cellular washout. Despite this, the density leveled off at an optical density of approximately 0.01. At elapsed time of just over 260 hours, the feed of glucose was stopped and a strain was isolated from the culture (point 3 on the plot). This strain was designated BOP376e. After collection of this sample and isolating a single colony from it, the fermentor was continued and operated in batch model for

approximately 200 hours to examine stain viability. The density steadily increased to approximate OD of 0.06 until it was ended after 200 hours (data not shown). In total, the culture underwent 17 doublings and 1.8 x 10^{12} duplication events while in continuous culture, with 5 doublings and 1.6 x 10^{11} division events growing solely on glucose. The isolated colony was shown to have the game genotype as the starting strain by confirming the knockouts by PCR and was labeled BOP376e.

7.3.10 Increasing the rate of mutation: deletion of the mutS gene

In order to increase the rate of mutation and subsequently reduce the time necessary to evolve strains to an optimal phenotype, the *mutS* gene was removed from strains BOP374e and BOP376e to generate mutator strains BOP384 and BOP386, respectively (see **Figure 7.3** and **Table 7.3**). The *mutS* gene is involved in DNA mismatch repair and is also conserved across species[47,48], thus its removal will allow an increase in mutation rate. This increase in adaptation mutation rate has been previously demonstrated to be significant in *mutS* mutants of *E. coli*[49,50].

7.3.11 Adaptive evolution to optimal production phenotypes

Utilizing the strains that were able to grow on minimal medium, strains were subject to the adaptive evolution process after initial characterization in the desired anoxic conditions for strain optimization. This procedure was performed in anaerobic serial-passage in 100 mL batch cultures following procedures we have developed in our lab[12,41,42,44,51]. These procedures are optimized so that the cells never reach stationary phase where sufficiently large transfer volumes are desired to reduce the chance of fixation of hitchhiker mutations.

7.3.12 Characterization of strain subject to adaptive evolution

Before evolution, strains were initially characterized to examine their growth properties in anoxic conditions. Both the lactate and L-alanine production strains were analyzed separately.

7.3.13 Lactate production strains

Initial characterization of the lactate production strain revealed that supplementation of the medium was necessary to sustain anaerobic growth during the adaptive evolution process. To initiate the evolution process, the lactate production strain BOP384 was inoculated from frozen stock under kanamycin selection in aerobic conditions in Luria-Bertani (LB) broth and inoculated into 4 g/L glucose M9 medium and underwent several doublings aerobically before use for inoculation into flasks in anoxic conditions in an anaerobic chamber (see Methods). This was the general method used for inoculating strains prior to adaptive evolution. **Figure 7.5** shows the growth curves of strain BOP384 in anoxic conditions on both glucose and xylose, along with different levels of supplementation. The supplement identified to sustain growth was yeast extract as it is widely used in bacterial culture to increase cell densities[16,17,29].

Figure 7.5: Growth curve for the characterization of the lactate production strain on glucose and xylose. (facing page) BOP384, the lactate production strain was examined for its growth on (**a**) glucose and (**b**) xylose minimal medium with and without supplementation of yeast extract (YE) at different levels. For growth on glucose, all levels of supplementation had a similar growth profile and final optical density, both greater than those for no supplementation. For growth on xylose, the supplementation of YE had a direct effect on growth rate and final density. Similar growth rates were obtained for the highest supplemented cultures, and successively slower rates for the 1.0 g/L and unsupplemented cultures. Additionally, the final density was proportional to the amount of YE supplied. The arrows in plot **b** indicate where cultures were used to inoculate other flasks. See **Table 7.5** for additional growth data from this analysis.

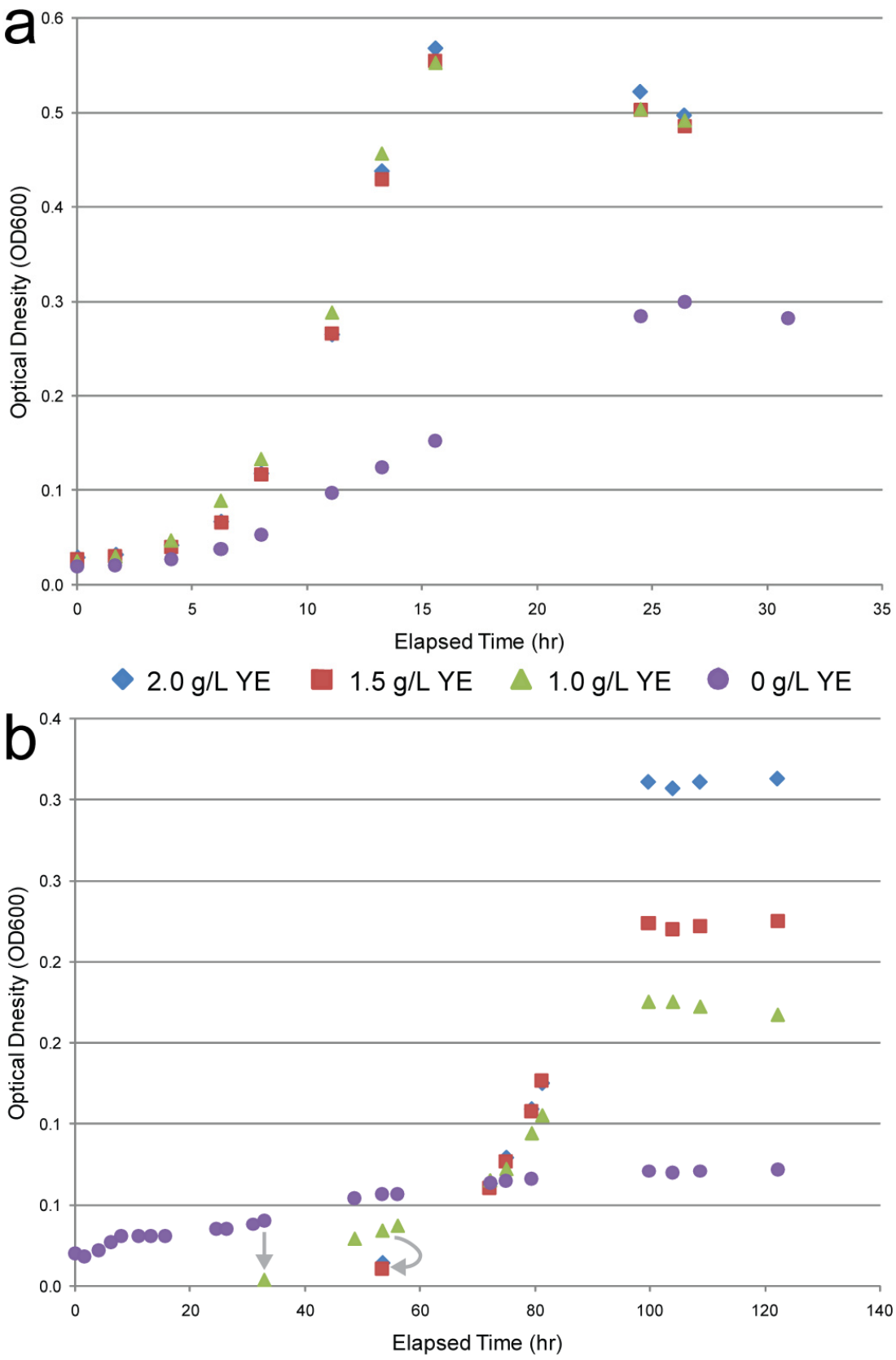

a
Optical Dnesity (OD600)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30
35
Elapsed Time (hr)
2.0 g/L YE
1.5 g/L YE
1.0 g/L YE
0 g/L YE
b
Optical Dnesity (OD600)
0.0
0.1
0.1
0.2
0.2
0.3
0.3
0.4
0
20
40
60
80
100
120
140
Elapsed Time (hr)

Table 7.5: Characterization of the lactate production strain BOP384 prior to evolution.

Culture Conditions	Supp.	μ	Final pH	Product / Substrate	Production / Consumption Rate	% $Y_{p/s}$ Steady-state	% $Y_{p/s}$
		hr-1			mmol gDW^{-1} hr^{-1}	wt%	wt%
4 g/L glucose M9	1, 1.5, 2 g/L YE	0.26 ± 0.01	5.48 ± 0.01	glucose	20.7 ± 0.7		
				lactate	41.1 ± 4.2	99.7 ± 0.1%	101.3 ± 6.4%
				succinate	0.6 ± 0.1	1.8 ± 0.4%	3.7 ± 0.5%
4 g/L glucose M9	none	0.20	5.28	glucose	18.4		
				lactate	27.7	75.3%	97.3%
				succinate	0	0%	3.9%
4 g/L xylose M9	1.5, 2 g/L YE	0.08 ± 0.01	5.84 ± 0.14	xylose	NC		
				acetate	NC	NC	4.3 ± 0.5%
				lactate	NC	NC	80.4 ± 8.6%
				succinate	NC	NC	8.9 ± 0.1%
4 g/L xylose M9	1, 0 g/L YE	0.05, 0.02	6.02, 6.30	xylose	NC		
				lactate	NC	NC	111%, 151%
				succinate	NC	NC	14.8%, 16.3%

μ - growth rate, % $Y_{p/s}$ – percent production yield, Supp. – supplement, YE – yeast extract, NC – not calculated

For growth on glucose, strains growing with supplementation all possessed approximately the same growth profile and production characteristics when compared to the unsupplemented culture. For the supplemented cultures, the maximum grow rate was 0.26 ± 0.01 hr-1. The culture without supplementation had a lower growth rate at a max of 0.20 hr-1 during exponential phase. The overall batch product yields, $Y_{p/s}$, are given in **Table 7.5**. The products generated from fermentation during exponential growth for the yeast extract supplemented cultures were almost entirely lactate at a rate of 41.1 ± 4.2 mmol gDW^{-1} hr^{-1} and a minor amount of succinate production 0.6 ± 0.1 mmol gDW^{-1} hr^{-1}. This corresponds to a 99.7± 0.1% product yield at steady-state for the three supplemented cultures. For the unsupplemented culture, the uptake rate of glucose was slightly less, but the production rate for lactate was significantly lower, at a product yield of 75.3% at steady-state. This indicates that the yeast extract supplementation provided additional nutrients to generate cellular

biomass during steady-state. The final cellular concentrations and overall product yields indicate that it was likely that some of the supplements were converted to lactate. At the end of the growth phase, all of the cultures containing yeast extract consumed all of the glucose whereas the culture with no supplementation consumed nearly all of the glucose, approximately 93%.

For growth on xylose, supplementation had a greater effect on the growth rate when compared to results for growth on glucose. Different from the growth study on glucose, the strains were grown initially with no supplementation and then passed to xylose medium with supplementation. For supplementation with 1.5 and 2.0 g/L of yeast extract, the growth rate was 0.08 ± 0.01 hr-1 and for supplementation of 1.0 g/L and no supplementation of yeast extract, the growth rate was 0.05 hr-1 and 0.02 hr-1, respectively. The final cellular density of the cultures before entering stationary phase varied significantly with the higher level of supplementation ending up with higher final cell densities (see **Figure 7.5**). The product yields at the end of fermentation of xylose were lactate at a product yield of 80.4 ± 8.6%, succinate at 8.9 ± 0.1%, and acetate as 4.3 ± 0.5% for the two highest supplemented cultures. For the 1.0 g/L and unsupplemented cultures, the product yields of lactate were over 100%. This is most likely due to the low overall conversion of substrate and the passage of initial glucose and other supplements in the initial inoculums. Additionally, none of the cultures fully consumed all of the xylose available; with the amount of yeast extract supplemented contributing directly to the total xylose consumed (85%, 78%, 64%, and 15% for the 2.0, 1.5, 1.0, and 0 g/L supplemented cultures, respectively). These results indicate that growth of unevolved BOP384 on xylose was at an extremely low rate and was unsustainable (as further described below).

After the initial characterization of the strains anaerobically, medium conditions were chosen for the adaptive evolution process. This medium was M9 minimal medium with supplementation of 1.0 g/L of yeast extract. Supplementation at 1.0 g/L was chosen as it allowed for a faster growth rate (which would allow for earlier detection of positive results) and it allowed for a significant increase in consumption of the substrate (thus providing a longer growth period) in the case of xylose as the substrate. The smallest level of supplementation was chosen as the exact composition of yeast extract is unknown, thus complicating modeling of experimental findings. From this result, it could be concluded that the lack of YE in the chemostat evolution could have led to the low cell densities observed during the process (**Figure 7.4**). Further justifying the choice of medium supplementation with yeast extract, after BOP384 growing with no supplementation was passed to fresh medium (with the same initial composition, no supplementation) under anaerobic conditions (both glucose and xylose), growth ceased in numerous attempts after approximately two to three culture doublings (results not shown).

7.3.14 L-alanine production strain

Initial characterization of the BOP386 L-alanine production strain in anaerobic conditions revealed that the strain was not viable under anaerobic growth conditions. Growth of the L-alanine strain was possible in MES1 base medium with a concentration of 4 g/L glucose aerobically, but after passage to anaerobic conditions, growth ceased after approximately one population doubling. This result was independent of supplementation with different medium additives (e.g., yeast extract) and was repeated several times. Due to the lack of sustainable growth anaerobically, work with the L-alanine production strains (BOP386 and BOP392) was halted.

7.3.15 Evolution for optimization of the lactate production strain on glucose

Evolution of the lactate producing strain resulted in a significant increase in growth rate over the initial starting strain. BOP384 was evolved in duplicate on glucose using the established batch serial passage adaptive evolution process to keep the population in exponential growth phase (see Methods). The starting strain was unevolved BOP384 supplemented with 1.0 g/L YE (characterized in **Table 7.5**). Passages were performed at a target optical density of 0.2 to keep the cell growing during exponential growth phase. Both of the duplicated evolutions resulted in endpoint strains with very similar growth and production profiles. Therefore, they will be described as biological replicates. The percent increase over the initial 0.26 ± 0.01 hr-1 growth rate during the course of evolution (200 hours) was roughly 500%, as shown in **Figure 7.6**. There was a rapid increase in the growth rate during the adaptation with the cultures reaching essentially their final growth rates in 2.5 days. This adaptation period of 2.5 days is much less than that of wild-type *E. coli* evolving aerobically on glycerol in which strains reached their final growth rate in 10 - 30 days[41]. Furthermore, a similar trend was seen (evolution of 10 – 30 days) for single deletion knockout strains of *E. coli* when evolved on different substrates[51] and previous lactate growth-coupled designs [12]. However, a few strains from the single deletion strain analysis had an adaptation period on the same order[51]. The final growth rate anaerobically was much higher than observed in earlier studies, which will be discussed later. In total, the evolution process was carried out for slightly over 8 days until the growth rate stopped increasing. At the end of the evolution, the cultures underwent 292 ± 1 doublings and 6.1 ± 0.2 x 10^{11} cellular division events. At day 2.5, the cultures had undergone 71 ± 0 doublings and 2.2 ± 0.1 x 10^{11} cellular division events. Single colonies were isolated from the final cultures of each evolution, their

knockout genotypes were confirmed by PCR, and were designated as BOP384eG1 and BOP384eG2.

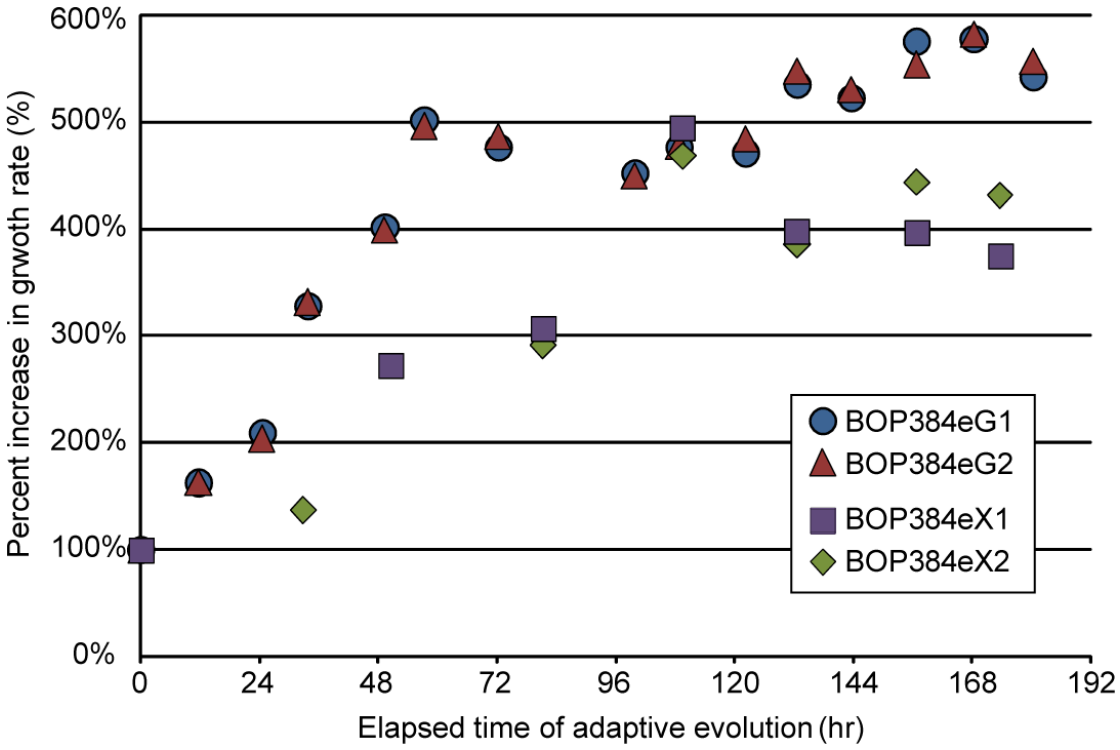


Figure 7.6: Increase in growth rate during adaptive evolution of lactate production strains. The percent increase of growth rate over the initial unevolved lactate production strain, BOP384, over time when evolved on glucose and xylose. Each strain was evolved in duplicate in parallel evolutions. The initial growth rates were determined each on glucose and xylose and the same starting strain was used for all of the evolutions. Both duplicates of the glucose (BOP384eG1 and BOP384eG2) and xylose (BOP384eX1 and BOP384eX2) evolutions possessed a similar growth rate path to the end phenotype and similar final endpoint growth rates.

The evolved lactate production strains displayed a biphasic growth during the adaptive evolution process. Strains BOP384eG1 and BOP384eG2 were cultured from frozen stock and examined for their production capabilities. When reintroduced into anaerobic conditions, both cultures were allowed to undergo several culture doublings (approximately 10) to ensure that there was no oxygen present in the culture and to re-establish the observed growth rate. A characteristic of both strains was that during the evolution process, the strains would display a biphasic exponential growth rate for different ranges of cellular density. When growing from a typical starting population of roughly 10^3 - 10^6 cells per liter to the passage density (an optical density of 0.2), the cells would display a higher growth rate (as depicted and characterized in **Figure 7.6**, a maximal growth rate of approximately 1.2 -1.4 hr^{-1}). This growth rate is much higher than previous anaerobic measurements of growth rate in our lab[12,52]. When growing from a cellular density above this value, the growth rate would again be exponential, but at a value of roughly 2/3 of the growth rate observed when starting from low initial densities (i.e., low density being ≤ 10^6 cells). This density region where metabolite concentration changes could be detected was characterized (**Table 7.6**). This biphasic is likely due to the consumption of one or multiple substances initially present in the culture medium present from initial yeast extract supplementation and the disappearance of this substance at the time of the switch in growth rate. These potential metabolites were attempted to be identified through HPLC with no success. Furthermore, when it was attempted to grow cells without yeast extract supplementation after evolution, or to wean them off of this supplement, cultures ceased to grow after 2 - 4 doublings (data not shown). Despite this, the endpoint lactate production strains could be characterized growing at a steady-state, which is now described.

Table 7.6: Characterization of the evolved lactate production strains BOP384eG1, G2, X1, X2.

Culture Conditions	Supp	μ	Product / Substrate	Production / Consumption Rate	% $Y_{p/s}$ Steady-state	% $Y_{p/s}$
	g/L	hr-1		mmol gDW^{-1} hr^{-1}	wt%	wt%
4 g/L glucose M9	1 g/L YE	0.86 ± 0.00	glucose	43.1 ± 1.3		
			lactate	84.4 ± 1.5	97.9 ± 1.2%	*98.4 ± 3.4%
			succinate	4.3 ± 0.3	6.5 ± 0.3%	*3.4 ± 2.8%
4 g/L xylose M9	1 g/L YE	0.13 ± 0.02	xylose	9.5 ± 0.8		
			acetate	1.0 ± 1.4	4.0 ± 5.7%	2.3 ± 3.2%
			lactate	13.0 ± 0.4	82.3 ± 4.1%	83.7 ± 3.0%
			succinate	2.6 ± 0.8	21.4 ± 4.7%	18.7 ± 2.5%

μ - growth rate, % $Y_{p/s}$ – percent production yield, Supp. – supplement, YE – yeast extract

*Overall yield monitored throughout evolution and was consistent after 2.5 days.

Characterization of the final lactate producing strains indicated evolution to a production phenotype in agreement with computational predictions. **Table 7.6** contains data from the characterization of the endpoint strains. The final production rate of lactate was 84.4 ± 1.5 mmol gDW^{-1} hr^{-1} when considering both strains and additionally succinate was made a rate of 4.3 ± 0.3 mmol gDW^{-1} hr^{-1}. This correlated to a 97.9 ± 1.2% and 6.5 ± 0.3% wt% product yield at steady-state during the exponential growth phase for lactate and succinate, respectively. Overall percent product yields for lactate and succinate were 98.4 ± 3.4% and 3.4 ± 2.8% wt%, respectively. The overall yield was monitored throughout the evolution and was consistent at these values after 2.5 days into the evolution. The steady-state production rate of lactate increased over 2 fold for the endpoint strain over the unevolved strain. The increase in the glucose consumption rate also has a similar 2 fold increase, however the production rate of succinate in molar production increased 7 fold. The steady-state wt% yield for lactate was approximately the same and that of succinate increased approximately 3.5 fold. The overall wt% yields were

approximately the same for both lactate and succinate. The summation of the wt% yields over 100% indicate that some of the supplemented yeast extract was contributing to the lactate and/or succinate production. Overall, 93.8% of the total products generated during fermentation at steady-state was lactate, close to homofermentative criteria set for the strain of less than 2% byproducts in the computational selection of strain designs. These endpoints display superior performance in terms of production rate, growth rate, and byproduct formation over previously generated lactate production strains[12]. In comparison to additional lactate production studies on an industrial scale, the steady-state and overall yields generated in this study of 0.98 g g^{-1} are at the same level or above previously reported values of 0.9 g g^{-1} (see ref.[23]), 0.93 g g^{-1} (see ref.[24]), 1.0 g g^{-1} (see ref.[25]), and 0.86 g g^{-1} (see ref.[26]). Furthermore, even at the relatively low cellular densities used for this process, a volumetric productivity of 1.7 g L^{-1} hr^{-1} was achieved and again compares well with previous studies where cell densities were driven roughly an order of magnitude greater and productivity values of 0.7 - 3.5 g L^{-1} hr^{-1} were reported[23-26].

Figure 7.7a presents the data from the unevolved and evolved strains on glucose integrated with modeling predictions. Shown are the production envelopes for the initial unevolved strains with (red) and without (green) supplementation along with the endpoint strains (blue). These envelopes are predicted using the *i*AF1260 model, the experimentally measured glucose uptake rates (solid lines, averages; dashed lines, considering standard deviation; see **Table 7.6**), input medium conditions, and the maintenance parameters determined using glucose as a substrate (see Methods). Also shown are the exponentially measured data points with standard deviations (error bars, unevolved unsupplemented was a single measurement).

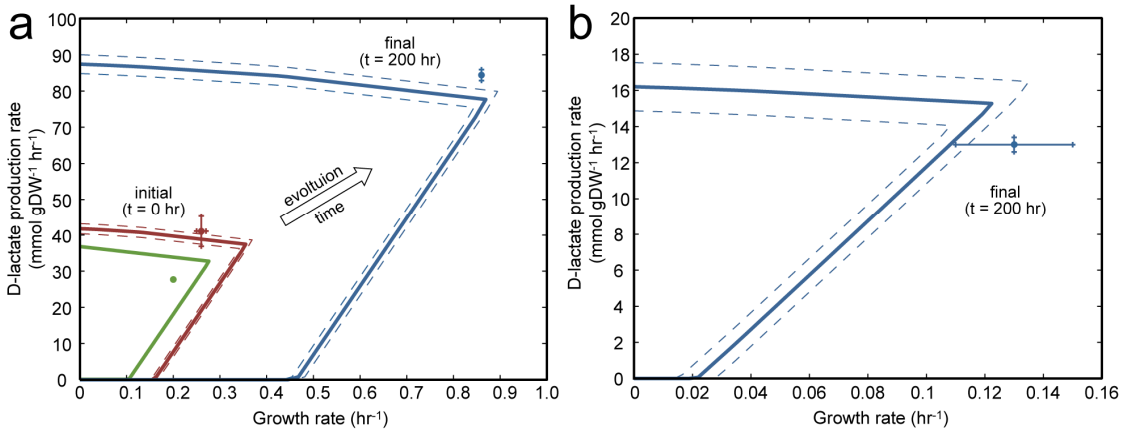


Figure 7.7: Predicted production envelopes and experimental measurements for evolved lactate production strains. The predicted production envelopes and experimental production measurements for unevolved and end point lactate production strains from (**a**) glucose (BOP384, BOP384eG1, and BOP384eG2) and (**b**) xylose (BOP384eX1 and BOP384eX2) evolutions. For each plot, the production envelopes are given based on experimentally measured substrate uptake rates (mmol gDW^{-1} hr^{-1}), solid lines (averages) and dashed lines (considering standard deviation). Also plotted are the experimentally measured values for the lactate production rates (mmol gDW^{-1} hr^{-1}) and growth rates (hr^{-1}). Experimental values are given in Table 5 (unevolved strains) and Table 6 (evolved endpoint strains). For the glucose evolution, the initial unevolved production envelopes and experimental data points are plotted. For each, the outermost blue lines and points are for the endpoint strains (with error bars), and for the glucose utilizing strain, green is the unevolved strain without supplementation (single measurement) and red is the unevolved strain with yeast extract supplementation. For each of the endpoints, the optimal growth rate values for the envelopes lie near the experimentally determined endpoint growth rates and lactate production rates in each of the cases. The glucose evolution resulted in a significant increase in production and growth rates.

The computational predictions and experimental evolved endpoint measurements display good agreement. For the unevolved unsupplemented measurement of growth on glucose, the initial phenotype is suboptimal with the production rate and growth rate less than the optimally predicted point. Coincidentally, the glucose uptake rate of 18.4 is almost identical to that observed for anaerobic growth of *E. coli* in an earlier modeling and experimental evaluation of growth[52] and gives confidence in the single measurement. Supplementation with yeast extract for unevolved glucose growth displays near optimal behavior as

predicted with the model at the measured uptake rate. This demonstrates that yeast extract allows more incomings carbon (glucose and yeast extract content) to be used for lactate production and for biomass generation. Potentially more for lactate production than for growth as the point lies above the production envelope on the production rate axis. The endpoint predictions and experimental measurements are in good agreement, with the experimentally measured growth rate and lactate production rate contributing to a point very near the optimal growth rate. The increased rates are due to the over 2 fold increase in the glucose uptake rate. This characterization of an adaptively evolved production strain demonstrates that a computationally designed strain can result in an experimentally verified production phenotype in good agreement with modeling predictions.

7.3.16 Evolution for optimization of the lactate production strain on xylose

Evolution of the lactate producing strain on xylose similarly resulted in a significant increase in growth rate over the initial starting strain. Similar to the evolution on glucose, BOP384 was evolved in duplicate with 1.0 g/L YE, passages were performed at a target optical density of 0.2, and both of the separate evolutions resulted in endpoint strains with very similar growth and production profiles. The percent increase over the initial 0.05 hr-1 growth rate during the course of evolution (200 hours) was roughly 400%, as shown in **Figure 7.6**. In contract to the glucose evolution, the increase in the growth rate during the adaptation was more gradual with the culture reaching essentially its final growth rate in 5.5 days. In total, the evolution process was carried out for slightly over 8 days until the growth rate stopped increasing. At the end of the evolution, the cultures underwent 41 ± 0 doublings and 1.3 ± 0.0 x 10^{11} cellular division events. At day 5.5, the cultures had

undergone 29 ± 1 doublings and 8.7 ± 0.9 x 10^{10} cellular division events. Single colonies were isolated from the final cultures of each evolution, their knockout genotypes were confirmed by PCR, and were designated as BOP384eX1 and BOP384eX2. Similarly to the glucose production strains, the xylose utilizing strains also displayed biphasic growth most likely due to yeast extract.

Characterization of the final lactate producing strains on xylose indicated evolution to a production phenotype in agreement with computational predictions. **Table 7.6** contains data from the characterization of the endpoint strains. The final production rate of lactate was 13.0 ± 0.4 mmol gDW^{-1} hr^{-1} when considering both strains and additionally succinate was made a rate of 2.6 ± 0.8 mmol gDW^{-1} hr^{-1} and acetate at a rate of 1.0 ± 1.4 mmol gDW^{-1} hr^{-1}. This correlated to a 82.3 ± 4.1%, 21.4 ± 4.7%, and 4.0 ± 5.7% wt% product yield at steady-state during the exponential growth phase for lactate, succinate and acetate, respectively. Overall product wt% yields for lactate, succinate, and acetate were 83.7 ± 3.0%, 18.7 ± 2.5%, and 2.3 ± 3.2% wt%, respectively. The overall yield was calculated for the endpoint strains. The summation of the wt% yields above 100% for steady-state calculations and near to 100% for overall product yields indicated that some of the supplemented yeast extract was contributing to the lactate, succinate, and/or acetate production. Overall, 76.4% of the product generated during fermentation at steady-state was lactate, a value less than that for growth on glucose. The overall production yield for the evolved xylose strains increased for lactate and succinate, whereas that for acetate decreased when compared to the two highest supplemented unevolved strains. In comparison to previously published work generating lactate from xylose in *E. coli*, a yield of 0.78 g g^{-1} (see ref.[24]) was achieved and the 0.82 g g^{-1} steady-state and 0.84 g g^{-1} overall product yield of the endpoint strains generated here were higher. A volumetric

productivity of 0.13 g L^{-1} hr^{-1} was achieved for the endpoint strains on xylose. The comparison study was again at industrial densities and substrate concentrations with a plasmid based production system[24].

Figure 7.7b presents the data from evolved endpoint strains on xylose integrated with computational predictions. Shown is the production envelope for the endpoint strains (blue). Again, the experimentally measured xylose uptake rates (solid lines, averages; dashed lines, considering standard deviation; see **Table 7.6**), input medium conditions, and modeling maintenance parameters were used to calculate the envelopes. For the xylose simulations, the non-growth associated maintenance energy requirement was lowered within a previously examined range[31] (see Methods). Also shown is the exponentially measured data point with standard deviations (error bars).

The computational predictions and experimental evolved endpoint measurements were again in good agreement for xylose endpoint strains. The experimental point for the endpoint measurements lies very near the point of optimal growth on the production envelope. However, different form the glucose evolved strains; the experimental data point lies to the higher growth rate side of the envelope. This is most likely due to a larger effect on growth rate due to supplementation. Taken together, characterization of the endpoint phenotypes resulted in validation that the methods utilized were able result in final strains that produced the intentioned products, both in a high percentage product yield.

7.4 Conclusions and Discussion

Metabolic engineering will become increasingly important to generate products for a growing worldwide population. Computational approaches to designs

strains and interpret data will play a major role in this effort. Although a number of computational analyses are being developed, few have been experimentally validated. Here, we present work on the construction and characterization of growth-coupled production strains designed using constrain-based optimization[1] and the process of adaptive evolution. The major results from this study are accordingly, i.) validation of two out of three strains constructed using this method demonstrating the predicted production phenotypes and their agreement with modeling predictions, ii.) development of new and improved methods for production strain optimization under anaerobic adaptations for both the removal of auxotrophy and production strain optimization, including the *mutS*$^-$ mutator strains, and iii.) generation of endpoint production strains suitable for genetic analysis and further development as industrial strains. Each topic will be discussed further.

The outlined process resulted in generation of computationally predicted production strains. This occurred for two out of the three cases examined, with the third case not growing anaerobically after the predicted gene deletions were implemented *in vivo*. However, the L-alanine production stain did grow aerobically and thus, an adaptive evolutionary process to transition from aerobic growth to increasingly anoxic growth conditions may allow for the strain to evolve to an anaerobic growth phenotype. This method shows promise as a strain was isolated during the evolution to alleviate acetate auxotrophy under anoxic conditions, even though that strain later was unculturable under anaerobic conditions in minimal medium batch culture. For the two success cases, the predicted production of lactate was successfully demonstrated. This finding provides confidence that further designs identified from such an analysis can have similar success, such as others identified from the screen of native compounds used in this study[1].

The methods developed here can be readily applied to generate further *E. coli* strains or strain designs for additional organisms. These methods demonstrate that success is possible when implementing computational designs which growth couple metabolite production to growth rate. Greatest immediate potential is for other production organisms for which metabolic reconstructions already exist (for a list, see[53]), such as *Bacillus subtilis*[54]. The use of steady-state adaptation is a novel approach from our lab that has not been previously utilized. Furthermore, evolution under steady-state conditions might be expanded to evolve strains to higher growth rates by incrementally driving up dilution rate to select for higher growing mutants. Direct expansion of the method should include computational evaluation of new reactions that can be implemented *in vivo* through homologous gene expression.

The use of supplementation of media components with yeast extract had an effect on the endpoint strains and interpretation of results with modeling. Due to the presented results and outcomes, evolution without supplementation is recommended as, i.) it would provide a consistent evolution condition that would be limited by the main substrate, a desired production condition, ii.) supplementation might cause formation of byproducts, as seen in the increase of succinate and acetate yields for evolutions on glucose and xylose when compared to unsupplemented unevolved strains, and iii.) it would allow more direct computational interpretation and integration of experimental data (e.g., uptake, production, and growth rates) as medium composition would be fully defined. Furthermore, cultures without supplementation might allow (or might allow more) genetic adaptations to accumulate as it is clear from the unevolved study that supplementation alters the interpretable phenotype compared to no supplementation and it might not let strains with advantageous mutations outperform (i.e., outgrow) others as the benefits from supplementation

might overcome any such mutations. After evolved strains are generated with no supplementation, yeast extract could always be added later to take advantage of its beneficial attribute of generating higher culture densities. Additionally, as seen for both evolutions to remove auxotrophy, the density of the cultures did not maintain a high value during the evolution. This could be due to the lack of supplements in the medium, which were verified to increase cell densities in the later experiments.

The strains generated from this analysis can be used as reagents for the study of causal (and non-causal) genetic changes encountered during adaptive evolution. The deletion of the *mutS* gene is a new approach that has not been studied in the context of generating production strains and could be a means to shorted evolution times, thus accelerating the strain design process. The *mutS* deletion for the lactate evolved strains could have played a role in the short adaptation period seen the significant increase in growth rate on glucose. This could be directly determined by evolving the parent strain to the final evolved strain that carries the wild-type *mutS* gene. In line with this, due to the short adaptation time and the repeatability of the adaptation (both in the duplicate endpoints presented here and in other attempts of a similar evolution with supplements not shown here), the increase in growth rate might not be due to genetic changes. Nonetheless, identification of beneficial genetic changes in production strains, such as those presented here, can allow for more focused design of future production strains through incorporation of such advantageous mutations.

The endpoint production strains show promise for further development as industrial strains. The production phenotype for the glucose endpoint strain demonstrate a high production rate, more than 2 fold for the best strain isolated from

a previous related study[12]. Additionally, the overall product yield is very close to theoretical maximum for a growing strain[1], and is even above the theoretical value due to supplementation. Although similar lactate production strains have been constructed[23-26], the strains generated here are advantageous as, i.) the strains were generated without any recombinant DNA, thus simplifying the construction and fermentation process (e.g., no induction necessary), ii.) do not require complex fermentations, such as two-stage aerobic and anaerobic fermentations, and iii.) have the potential for continuous processing. Furthermore, the production characteristics (product yields and volumetric productivities) of the strains demonstrated at low densities were in the range of those reported which were examined at a more industrially relevant scale [23-26]. Model-driven design also had an advantage in that it offered a means to predict the effect of additional knockouts and supplementation of medium, which was mostly speculative and performed by trial and error in other studies. One specific finding of modeling was that acetyl-CoA could be generated in the *pflAB, pflDC, aceEF* knockout strains from additional pathways detailed in the model that was speculated not to be possible[26]. The weaning of the strains off of acetate supplementation performed in this study provides evidence that this is indeed possible; however the necessity of yeast extract for continued growth indicates that further testing is needed. Modeling predicted that the lactate produced here is optically pure D-lactate, as the actively predicted enzyme is catalyzed by the D-lactate specific *ldhA* gene. Optically pure lactate is preferred for polymer generation[33] and can be directly assayed in these strains. The next step for the evaluation of adaptively evolved strains generated here using serial passage is evolution under continuous processing conditions and higher substrate conditions. Strains that are growth coupled are, in theory, suitable for continuous culture as they will not be

outperformed by mutants that exhibit faster growth under such conditions. This continuous processing potential has significant implications to impact the field of metabolic engineering.

7.5 Methods

7.5.1 Model

The metabolic reconstruction of *E. coli i*AF1260[31] was utilized as a basis for the model used throughout with the minor changes described[1].This model has been functionally tested and verified against experimental data to be predictive for computations of growth rates, metabolite excretion rates, and growth phenotypes on a number of substrate and genetic conditions[31]. For all simulations, the reactions CAT, SPODM, and SPODMpp (oxidative stress reactions) and the FHL reaction were constrained to zero for reasons previously established[31]. Flux balance analysis (FBA) was used for computing optimal phenotypes using iAF1260 and the outlined biomass objective function, BOF_{CORE} with the reported maintenance energies, presented with the reconstruction[31]. FBA performed using a steady-state assumption was been described in detail previously[55]. All computations were performed using the MATLAB® (The MathWorks Inc., Natick, MA) and the COBRA Toolbox[56] software packages with TOMLAB (Tomlab Optimization Inc., San Diego, CA) solvers.

Medium conditions for minimal medium were set to computational minimal medium as previously defined[31]. Minimal medium with yeast extract supplementation (for experimental comparison) was simulated by allowing amino acid and nucleotide base uptake rates for simulations in amounts proportional to that required for supporting a given experimentally determined growth rate computationally. In

modeling terms, a given uptake rate for an amino acid or nucleotide base was equal to the stoichiometric coefficient of that component in the biomass objective function multiplied by the experimental growth rate. All phosphorylated compounds in the biomass objective function were substituted with their unphosphoylated base to match the probable biological composition of yeast extract. For growth on glucose, half of the requirements were allowed and a full contribution of was allowed for growth on xylose. The growth and non-growth maintenance modeling parameters identified in model development for growth on glucose were used for all design calculations[31]. Furthermore, these same parameters were used for growth on glucose for comparison to experimental data, whereas the non-growth maintenance parameter was reduced to 50% of the 8.39 mmol ATP gDW^{-1} hr^{-1} value for some xylose computations, a value previously described in a sensitivity analysis[31].

7.5.2 Computational Analyses: Selection of and sensitivity analysis on produced strains

In order to improve computationally identified growth coupled knockout strains and reduce the number of knockouts necessary, a COBRA Toolbox function[56] called analyzeGCdesign was created. This function uses a simple algorithm and objective function to find a better growth coupled solution, given an OptKnock or OptGene solution (or any set of knockouts) as an input, as calculated previously for a number of metabolites[1]. It iterates through the set of knockout reactions and replaces each reaction, one at a time, with every reaction in a predetermined set of selected target reactions[1]. AnalyzeGCdesign also deletes each reaction from the knockout set one at a time, and adds every selected reaction to the full knockout set one at a time. After each change is made, the value of the objective function is calculated. The single

change that produces the highest objective function is then recursively passed back to the analyzeGCdesign function, which then uses the same algorithm to try to improve this new set of knockouts. The function continues replacing, adding, and removing reactions from the set of knockouts until no single change can further increase the value of the objective function. Any of the following eight objective functions can be used:

1. $objective = maxProd$ (yield)
2. $objective = maxProd * growthRate$ (substrate specific productivity, SSP)
3. $objective = maxProd * delPenalty^{numDels}$ (yield with knockout penalty)
4. $objective = maxProd * growthRate * delPenalty^{numDels}$ (SSP with knockout penalty)
5. $objective = \frac{maxProd}{slope}$ (growth coupled yield)
6. $objective = \frac{maxProd * growthRate}{slope}$ (growth coupled SSP)
7. $objective = \frac{maxProd * delPenalty^{numDels}}{slope}$ (growth coupled yield with knockout penalty)
8. $objective = \frac{maxProd * growthRate * delPenalty^{numDels}}{slope}$ (growth coupled SSP with knockout penalty)

where *growthRate* is the maximum possible growth rate, *maxProd* is the minimum production rate of the specified product at the highest growth rate, *slope* is the slope of the lower edge of the production curve (indicating the degree of growth coupling), *delPenalty* is the deletion penalty, and *numDels* is the number of knockout

reactions. These functions can be used to increase growth and production rates while decreasing the slope and number of deletions.

Product yield, $Y_{p/s}$, calculated computationally through modeling was calculated from coinciding uptake and secretion rates of a strain growing optimally (i.e., at its fastest rate). This yield correlates with the steady-state yield calculated from a cell growing at a functionally constant rate during exponential growth and is designated as $Y_{p/s}$ at steady-state. Product yield that was calculated at the end of a batch fermentation, over the entire cycle of a strain, was also calculated by:

$$Y_{p/s} = \frac{P_f - P_i}{S_f - S_i}$$

where *P* and *S* are the concentration of product and substrate, respectively, at time final (*f*) and initial (*i*). Volumetric productivities were calculated for endpoint evolved strains at an optical density of 0.5 for glucose and 0.25 for xylose, values achieved in the exponential growth phase.

7.5.3 Determination of High-flux pathways

Simulations were run to predict the pathways that carried high levels of flux under an optimal growth production phenotype. Pathways (and reactions) that carried a value of equal or greater than 10% of the input flux value for the mail substrate were classified as ‘high-flux’. These pathways were compared to wild-type simulations under given similar conditions. This analysis is similar to the characterization of a high-flux backbone where the overall activity of metabolism is dominated by several reactions with very high fluxes[57].

7.5.4 Analysis of genetic lethality

The removal of reactions was chosen as a design variable instead of genes as it results in a smaller solution space and speeds up computation. A downside of this is that genes which catalyze multiple reactions can be targets for elimination and ultimately result in computational lethality. Therefore, an analysis to determine lethality was performed by manually analyzing reaction gene to protein to reaction associations outlined in the *i*AF1260 reconstruction[31]. FBA simulations were performed by knocking out targeted genes and examining resulting growth rates and optimal phenotypes.

7.5.5 Strain construction

The starting strain was wild-type *E. coli* strain K12 MG1655 (ATCC 700926). We have previously characterized this strain extensively physiologically[12,41-43], resequenced its genotype[44], and the computational model *i*AF1260 is based on the K-12 MG1655 genome. Gene disruptions were performed using homologous recombination of PCR-amplified linear fragments[45]. During the gene deletion process, strains were grown aerobically in LB liquid medium and on 1.0% agar plates and the antibiotics kanamycin, chlormaphenocal, and ampicillin were used for section. Plasmids pKD46, pKD13, and pCP20 were also used in the process. Strains are preserved at -80 C and were given a 'BOP' tag and number under the protocol used in our lab (for example, BOP27).

7.5.6 Medium Selection

Minimal M9 medium was selected for growth of the lactic acid producing strains as it has been demonstrated that lactic acid can be made with the given defined minimal nutrients[12]. For production of L-alanine, a MES buffered medium was

chosen as a pH of 6.0 has been shown to be optimal for the production of a similar amino acid, valine[16]. This MES medium was based from M9 composition and with the buffer compound changed. MES is a Good buffer compound for the given pH range[58]. Also added to the L-alanine medium was a higher amount of NH_4 by means of ammonium sulphate as a high-producing strain of L-alanine would require sufficient nitrogen. Yeast extract (Sigma, Catalog # 8013-01-2) and sodium acetate was used as supplementation where specified.

7.5.7 Continuous Culture Evolutions

Continuous culture was performed in 1.0 L New Brunswick BioFlo fermentors. Either M9 or MES1 medium was used (see **Appendix B**) in the fermentors as specified. The agitation rate was 500 and was constant for the run. Temperature was maintained and controlled at 37 C. Culture pH was maintained with 5% NaOH and was maintained at 7.0 for M9 medium cultures and 6.0 for MES1 medium cultures. Dissolved oxygen was also monitored for the evolution and was maintained at zero. The volume of medium in the fermentor was maintained at 1.0 L unless noted otherwise. To maintain anoxic conditions in the fermentor, 5% CO_2 balance N_2 was supplied to the fermentor at 1 VVM. For the evolution of strains and weaning off of acetate, the glucose was 4.0 g/L and the acetate was at 2.0 g/L. Feed rates are presented in the results section as a function of time. Samples were removed aseptically and optical density (OD) measurements (A600). Samples were also analyzed by HPLC.

7.5.8 HPLC analysis

Products were identified and quantified by HPLC using an Aminex 87-H ion exchange column at 65º C. The mobile phase was 5 mM H_2SO_4 at an isocratic flow was 0.5 ml/minute. Sample injection volume was 10 µL. Products were identified by retention time using utilizing ultraviolet detection at 210 nm and refractive index detection at 30º C internal temperature and 45º C external temperature and quantified by relating peak area to those of standards.

7.5.9 Calculation of growth rates, culture doublings, and division events

Cell concentration in cultures was determined by measuring the optical density at 600 nm (OD600) using a Biomate 3 spectrophotometer (Thermo Scientific, USA). A value of $1.55*10^{12}$ cells L^{-1} $OD600^{-1}$ was used to calculate cell numbers with a dry cell weight of $2.9*10^{-13}$ gDW $cell^{-1}$, total biomass can be calculated as 0.45 gDW L^{-1} $OD600^{-1}$ (see ref.[59]).

Growth rates of batch cultures were determined to be the maximum growth rate during exponential growth and were determined using at least three samples examining cell and metabolite concentration data points. Regression was used to fit a line to the natural logarithm of the data points, with the slope of this line equal to the exponential growth rate. The mutations that accumulate during adaptive evolution occur randomly during cell division, so it was useful to calculate the total number of cell doublings that have occurred at any time. The formula used for this was:

$$D = (2^G - 1) * I$$

where *D* is the total number of cellular division events, *I* is the initial number of cells, and *G* is the number of cell divisions per initial cell (the number of generations). One doubling occurred for every new cell in the culture.

7.5.10 Adaptive Evolution

Adaptive evolution was conducted in 100 mL flasks with either M9 or MES1 medium supplemented with 4.0 g/L glucose or xylose. Other supplements are stated in the results. Cultures were maintained at 37 C in an anaerobic chamber with the atmospheric gas being a mixture of 7.5 % H_2 / 10% CO_2 with balance N_2. Experiments were designed and calculated to keep cells growing in exponential growth phase. To do this, the inoculum was changed throughout the course of evolution for each passage. Cultures were frozen and stored at -80 C at regular intervals throughout adaptive evolution, approximately every other day.

Acknowledgements

We would like to thank Alex Azuma for help with HPLC analysis and medium preparation, Karsten Zengler for his help with anoxic cultures, and additionally Vasiliy Portnoy, Kenyon Applebee, and Dae-hee Lee for their invaluable insight in various project aspects.

Chapter 7, in full, is adapted from construction and evolution of *E. coli* production strains designed through model-driven metabolic engineering that is in preparation. The dissertation author was the primary author of this paper, which was co-authored by Jeff D. Orth, Daniel C. Zielinski, and Dr. Bernhard Ø. Palsson.

References

1. Feist AM, Zielinski DC, Orth JD, Schellenberger J, Herrgard MJ, Palsson BO. Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *In preparation* 2008.

2. Bailey JE. Toward a science of metabolic engineering. *Science* 1991;252:1668-75.

3. Lee SY, Papoutsakis ET. Metabolic Engineering. CRC Press, 1999.

4. Stephanopoulos G, Nielsen J, Aristidou A. Metabolic Engineering. San Diego: Academic Press, 1998.

5. Keasling JD, Chou H. Metabolic engineering delivers next-generation biofuels. *Nat Biotechnol* 2008;26:298-9.

6. Top Value Added Chemicals from Biomass. In: Werpy T, Petersen G, eds.: U.S. Department of Energy, 2004.

7. Chang MC, Keasling JD. Production of isoprenoid pharmaceuticals by engineered microbes. *Nat Chem Biol* 2006;2:674-81.

8. Paster M, Pellegrino JL, Carole TM. Industrial Bioproducts: Today and Tomorrow: U.S. Department of Energy, 2003.

9. Park JH, Lee SY, Kim TY, Kim HU. Application of systems biology for bioprocess development. *Trends Biotechnol* 2008;26:404-12.

10. Alper H, Jin YS, Moxley JF, Stephanopoulos G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in *Escherichia coli*. *Metab Eng* 2005;7:155-64.

11. Alper H, Miyaoku K, Stephanopoulos G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. *Nat Biotechnol* 2005;23:612-6.

12. Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91:643-8.

13. Lee SJ, Lee DY, Kim TY, Kim BH, Lee J, Lee SY. Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. *Appl Environ Microbiol* 2005;71:7880-7.

14. Wang Q, Chen X, Yang Y, Zhao X. Genome-scale in silico aided metabolic analysis and flux comparisons of *Escherichia coli* to improve succinate production. *Appl Microbiol Biotechnol* 2006;V73:887-894.

15. Lee SY, Kim JM, Song H, Lee JW, Kim TY, Jang YS. From genome sequence to integrated bioprocess for succinic acid production by Mannheimia succiniciproducens. *Appl Microbiol Biotechnol* 2008;79:11-22.

16. Park JH, Lee KH, Kim TY, Lee SY. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc Natl Acad Sci U S A* 2007;104:7797-802.

17. Lee KH, Park JH, Kim TY, Kim HU, Lee SY. Systems metabolic engineering of *Escherichia coli* for L-threonine production. *Mol Syst Biol* 2007;3:149.

18. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech* 2008;26:659-667.

19. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003;84:647-57.

20. Patil KR, Rocha I, Forster J, Nielsen J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* 2005;6:308.

21. Datta R, Henry M. Lactic acid: recent advances in products, processes and technologies - a review. *Journal of Chemical Technology & Biotechnology* 2006;81:1119-1129.

22. Sauer M, Porro D, Mattanovich D, Branduardi P. Microbial production of organic acids: expanding the markets. *Trends Biotechnol* 2008;26:100-8.

23. Chang DE, Jung HC, Rhee JS, Pan JG. Homofermentative production of D- or L-lactate in metabolically engineered Escherichia coli RR1. *Appl Environ Microbiol* 1999;65:1384-9.

24. Dien BS, Nichols NN, Bothast RJ. Recombinant Escherichia coli engineered for production of L-lactic acid from hexose and pentose sugars. *J Ind Microbiol Biotechnol* 2001;27:259-64.

25. Zhou S, Causey TB, Hasona A, Shanmugam KT, Ingram LO. Production of Optically Pure D-Lactic Acid in Mineral Salts Medium by Metabolically Engineered Escherichia coli W3110. *Appl. Environ. Microbiol.* 2003;69:399-407.

26. Zhu Y, Eiteman MA, DeWitt K, Altman E. Homolactate fermentation by metabolically engineered Escherichia coli strains. *Appl Environ Microbiol* 2007;73:456-64.

27. Lee M, Smith GM, Eiteman MA, Altman E. Aerobic production of alanine by Escherichia coli aceF ldhA mutants expressing the Bacillus sphaericus alaD gene. *Appl Microbiol Biotechnol* 2004;65:56-60.

28. Zhang X, Jantama K, Moore JC, Shanmugam KT, Ingram LO. Production of L - alanine by metabolically engineered Escherichia coli. *Appl Microbiol Biotechnol* 2007;77:355-66.

29. Smith GM, Lee SA, Reilly KC, Eiteman MA, Altman E. Fed-batch two-phase production of alanine by a metabolically engineered Escherichia coli. *Biotechnol Lett* 2006;28:1695-700.

30. Marz U. World Markets for Fermentation Ingredients BCC Research, 2005:117.

31. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. A genome-scale metabolic reconstruction for

Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007;3.

32. de Graef MR, Alexeeva S, Snoep JL, Teixeira de Mattos MJ. The steady-state internal redox state (NADH/NAD) reflects the external redox state and is correlated with catabolic adaptation in Escherichia coli. *J Bacteriol* 1999;181:2351-7.

33. Hofvendahl K, Hahn-Hägerdal B. Factors affecting the fermentative lactic acid production from renewable resources1. *Enzyme and Microbial Technology* 2000;26:87-107.

34. Davis EO, Henderson PJ. The cloning and DNA sequence of the gene xylE for xylose-proton symport in Escherichia coli K12. *J Biol Chem* 1987;262:13928-32.

35. Sumiya M, Davis EO, Packman LC, McDonald TP, Henderson PJ. Molecular genetics of a receptor protein for D-xylose, encoded by the gene xylF, in Escherichia coli. *Receptors Channels* 1995;3:117-28.

36. Hasona A, Kim Y, Healy FG, Ingram LO, Shanmugam KT. Pyruvate formate lyase and acetate kinase are essential for anaerobic growth of Escherichia coli on xylose. *J Bacteriol* 2004;186:7593-600.

37. Song S, Park C. Utilization of D-ribose through D-xylose transporter. *FEMS Microbiol Lett* 1998;163:255-61.

38. Raunio RP, Jenkins WT. D-alanine oxidase form Escherichia coli: localization and induction by L-alanine. *J Bacteriol* 1973;115:560-6.

39. Wang MD, Buckley L, Berg CM. Cloning of genes that suppress an Escherichia coli K-12 alanine auxotroph when present in multicopy plasmids. *J Bacteriol* 1987;169:5610-4.

40. Dym O, Pratt EA, Ho C, Eisenberg D. The crystal structure of D-lactate dehydrogenase, a peripheral membrane respiratory enzyme. *Proc Natl Acad Sci U S A* 2000;97:9413-8.

41. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186-9.

42. Fong SS, Marciniak JY, Palsson BØ. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 Using a Genome-scale in silico Metabolic Model. *Journal of Bacteriology* 2003;185:6400-8.

43. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 2006;103:17480-4.

44. Herring CD, Raghunathan A, Honisch C, Patel T, Applebee MK, Joyce AR, Albert TJ, Blattner FR, van den Boom D, Cantor CR, Palsson BO. Comparative genome

sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat Genet* 2006;38:1406-1412.

45. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A.* 2000;97:6640-5.

46. Kalapos MP. Methylglyoxal in living organisms: chemistry, biochemistry, toxicology and biological implications. *Toxicol Lett* 1999;110:145-75.

47. Acharya S, Foster PL, Brooks P, Fishel R. The coordinated functions of the E. coli MutS and MutL proteins in mismatch repair. *Mol Cell* 2003;12:233-46.

48. Schlensog V, Bock A. The Escherichia coli fdv gene probably encodes mutS and is located at minute 58.8 adjacent to the hyc-hyp gene cluster. *J Bacteriol* 1991;173:7414-5.

49. Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, Taddei F. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science* 2001;291:2606-8.

50. Shaver AC, Dombrowski PG, Sweeney JY, Treis T, Zappala RM, Sniegowski PD. Fitness evolution and the rise of mutator alleles in experimental Escherichia coli populations. *Genetics* 2002;162:557-66.

51. Fong SS, Palsson BO. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat Genet* 2004;36:1056-58.

52. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Applied and Environmental Microbiology* 1994;60:3724-3731.

53. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microbial organisms. *Nat Rev Microbiol* 2008;Accepted.

54. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 2007.

55. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886-897.

56. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox. *Nat. Protocols* 2007;2:727-738.

57. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* 2004;427:839-843.

58. Good NE, Winget GD, Winter W, Connolly TN, Izawa S, Singh RM. Hydrogen ion buffers for biological research. *Biochemistry* 1966;5:467-77.

59. Neidhardt FC. *Escherichia coli* and *Salmonella*: cellular and molecular biology. Washington, D.C.: ASM Press, 1996:2 v. (xx, 2822 , lxxvii).

Appendix A

Additional figure, tables, and text from the computational evaluation of the production potential of *E. coli*

Methods

Description of OptGene procedure

OptGene was implemented as previously described, with the following modifications, the genotype of the population was changed from genes to reactions. This was done to reduce the number of total targets needed as the number of gene associated with the set of target reactions was larger than the number of reactions (effectively decreasing search time) and to make OptGene input have the same format as OptKnock, the mutation function was modified as described below, some parameters were modified as described below, this was performed in a conservative fashion to increase the chance of finding an optimal solution, and solutions from OptKnock which had a positive production were used as an initial population for OptGene simulation, in order to increase the efficiency of the algorithm by inputting reasonably good initial guesses for further refinement. OptGene was implemented in the Matlab software framework using the Genetic Algorithms and Directed Search toolbox.

Specific Change: *Mutation Function:* It was noted that having an independent mutation function where at each generation, each reaction is toggled independently with probability 1/nrxns will tend to increase the number of mutations until half the reactions are knocked out. As this is undesirable, a small modification was added where with 50% knockouts would be randomly removed until the child genome has fewer KO's than the parent. This heuristic makes the probability of increasing the number of KOs roughly equal to the probability of removing one.

Specific Change: Hashing of genotypes: A speed improvement made was the hashing of previously explored genotypes. All genotypes for which a fitness has been computed are put into a hash table (the key being a concatenation of the genotype). When a new genotype needs to be evaluated, the hash table is queried first to see if it has been previously computed. This results in a significant speedup in performance without affecting the results of the algorithm. The hashtable is periodically purged to prevent memory overflow. Note that this improvement has no effect on the outcome of the algorithm, only the speed.

Table A.1: Theoretical Yields - Molar Yields.

	Substrate	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol
	Aerobicity	Anaerobic	Anaerobic	Anaerobic	Aerobic	Aerobic	Aerobic
product	no. of carbons	molar yield	molar yield	molar yield	molar yield	molar yield	molar yield
Ethanol	2	193%	159%	98%	193%	159%	108%
D-Lactate	3	193%	159%	13%	193%	159%	101%
Glycerol	3	73%	45%		148%	121%	
L-Alanine	3	193%	127%	13%	193%	157%	99%
L-Serine	3	81%	50%	5%	198%	163%	102%
Pyruvate	3	147%	104%	7%	208%	171%	106%
Fumarate	4	86%	52%	4%	174%	142%	94%
L-Malate	4	86%	52%	4%	174%	142%	94%
Succinate	4	144%	105%	10%	161%	131%	88%
2-Oxoglutarate	5	50%	34%	2%	122%	100%	64%
L-Glutamate	5	55%	37%	2%	113%	93%	61%

Table A.2: Changes from iAF1260 to make iAF1260b.

Reaction added (abbreviation)	**Reason for addition**
DHORTfum	Added to no longer make fumarate reductase reaction essential
MALt3pp	Means of excretion
ALAt2rpp	Means of excretion
GLYt2rpp	Means of excretion
CITt3pp	Means of excretion
ASPt2rpp	Means of excretion

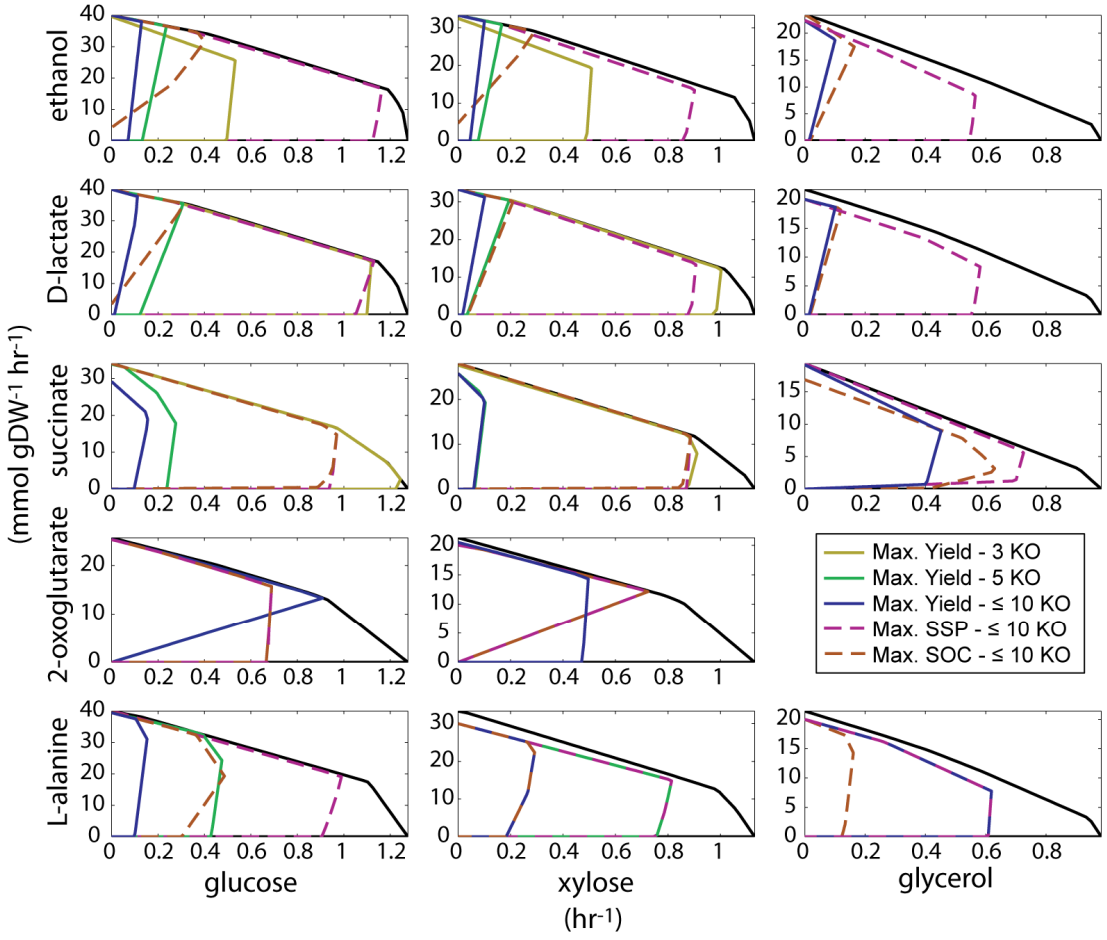


Figure A.1: The strain designs generated for five different targets from glucose, xylose, and glycerol aerobically. A set of graphs that give the production envelopes for different substrate / target pairs that were calculated during the analysis under anaerobic conditions. The different target production rates (mmol gDW^{-1} hr^{-1}) are shown on the y-axis and the growth rate (hr^{-1}) is given on the x-axis. Shown on each plot (if a solution exists) are the maximum yields for 3 knockouts (yellow, solid line), 5 knockouts (green, solid line), up to 10 knockouts (with a 99.99% deletion penalty, blue, solid line), the maximum substrate-specific productivity (SSP, pink, dashed line), and the maximum strength of growth coupling (SOC, orange, dashed line) design. For example, there are no valid solutions for 2-oxoglutarate production on glycerol given the minimum growth rate of 0.1 hr^{-1}. Some of the solutions are identical for multiple objectives.

Appendix B

Additional figure and table from the construction and evolution of *E. coli* production strains

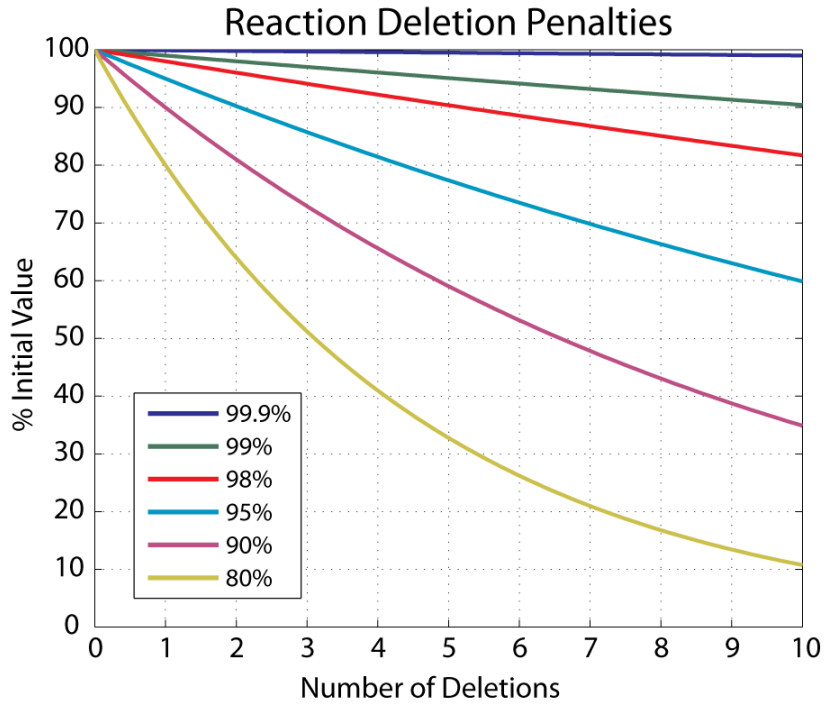


Figure B.1: The effect of knockout penalties on the objective function. A plot of number of knockouts versus the percent drops in the initial value of the objective function that result for a range of different penalty values. For more severe penalties, the drop in initial value can be significant and can result in an environment where lower knockout designs are heavily leveraged.

Table B.1: Culture medium.

M9 Minimal Medium		
Preparation and sterilization	**Substrate**	**Concentration (g/L)**
Sterilized and added individually	Glucose (or Xylose)	4
	$MgSO_4$	0.24
	$CaCl_2$	0.0111
M9 Salts – made in 10x concentration stock and sterilized by autoclave	Na_2HPO_4	0.68
	KH_2PO_4	0.3
	NaCl	0.05
	NH_4Cl	0.1
Trace Elements – made in 4000x concentration stock and sterilized by autoclave	$FeCl_3.6H_2O$	0.1
	$ZnSO_4.7H_2O$	0.02
	$CuCl_2.2H_2O$	0.004
	$MnSO_4.H_2O$	0.01
	$CoCl_2.6H_2O$	0.006
	$Na_2EDTA.2H_2O$	0.006
Purchased or made in entirety and sterilized by filtration	Wolfe's Vitamin Solution	(1x concentration)
MES1 Minimal Medium		
Preparation and sterilization	**Substrate**	**Concentration (g/L)**
Sterilized and added individually	Glucose	4
	$MgSO_4$	0.24
	$CaCl_2$	0.0111
MES Salts – made in 10x concentration stock and sterilized by autoclave	$(NH_4)_2SO_4$	12.5
	Betaine	0.2
	2-(N-morpholino)ethanesulfonic acid	15
	KH_2PO_4	1
	NaCl	0.5
Trace Elements – made in 4000x concentration stock and sterilized by autoclave	$FeCl_3.6H_2O$	0.1
	$ZnSO_4.7H_2O$	0.02
	$CuCl_2.2H_2O$	0.004
	$MnSO_4.H_2O$	0.01
	$CoCl_2.6H_2O$	0.006
	$Na_2EDTA.2H_2O$	0.006
Purchased or made in entirety and sterilized by filtration	Wolfe's Vitamin Solution	(1x concentration)