Title:
Systems biology analysis of Escherichia coli for discovery and metabolic engineering

Author:
Orth, Jeffrey David

Acceptance Date:
2012

Series:
UC San Diego Electronic Theses and Dissertations

Degree:
Ph. D., UC San Diego

Permalink:
http://escholarship.org/uc/item/3h74m1h6

Local Identifier:
b7244721

Abstract:
Systems biology is an emerging field of research that utilizes high-throughput experimental data and computational analysis methods to study biochemical networks. One of the common denominators of systems biology is the genome-scale reconstruction of metabolic networks. These reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledgebases that seek to formally represent the known metabolic activities of an organism. In the first part of this dissertation, the general properties of these reconstructions, along with the use of constraint-based modeling to analyze these networks, is described in detail. Metabolic network reconstructions have many practical uses, including use in discovery of new gene functions and metabolic reactions, and for metabolic engineering. The core E. coli metabolic model, a small-scale model that can be used for in-depth analysis of new constraint-based methods, is presented and analyzed in detail. In the second part of the dissertation, the genome -scale metabolic reconstruction of E. coli was updated and analyzed. This reconstruction was first published in 2000, and has been updated and periodically published since then. The current version of the reconstruction is called iJO1366, and was updated based on new literature and database information. The remaining network gaps were analyzed, and a new gap-filling workflow was developed and used to predict the missing metabolic reactions and genes in iJO1366. Model predicted growth phenotypes were compared to a large experimental dataset of gene knockout strain phenotypes, and model errors were identified. Several in vivo experiments were performed to validate these model-based predictions. In the third part of the dissertation, the metabolic network model of E. coli was used for metabolic engineering. A large-scale computational screen using the algorithms OptKnock and OptGene was performed to design many growth-coupled production strains for various useful compounds. In order to validate the accuracy of these predictions and to determine if adaptive laboratory evolution can be utilized as an effective strain

engineering tool, two strain designs were selected for experimental analysis. These knockout strains were constructed and evolved to optimize their phenotypes. One evolved strain worked as expected, producing a high yield of lactate from glucose. The other failed to produce 1,2-propanediol as predicted. This strain was analyzed by expression profiling, providing evidence for another metabolic gene function

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Systems Biology Analysis of *Escherichia coli* for Discovery and Metabolic Engineering

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Bioengineering

by

Jeffrey David Orth

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Michael J. Heller
Professor Xiaohua Huang
Professor Andrew D. McCulloch
Professor Milton H. Saier

2012

The Dissertation of Jeffrey David Orth is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2012

DEDICATION

For all their support, care, and generosity, this dissertation is dedicated to Desiree, Maeby, and Surely.

EPIGRAPH

The harder I work, the luckier I get.

-Samuel Goldwyn

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Bernhard Palsson, for the expert guidance and support he has provided during my Ph.D. research. I would also like to thank my fellow graduate students and the other researchers of the Systems Biology Research Group. Over the past five years, I have collaborated with and learned a great deal from my outstanding colleagues. I also thank Desiree Nguyen for her help reviewing my publications and this dissertation.

Chapter 1 is, in part, adapted from a chapter that appeared in EcoSal – *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (www.ecosal.org), Chapter 10.2.1, ASM Press, Washington D.C., February 18, 2010. The dissertation author was the primary author of this chapter which was coauthored by Ronan M.T. Fleming and Bernhard Ø. Palsson.

Chapter 1 is also, in part, adapted from a paper that appeared in Nature Biotechnology, Volume 28, Number 3, Pages 245-8, March 2010. The dissertation author was the primary author of this paper which was coauthored by Ines Thiele and Bernhard Ø. Palsson.

Chapter 1 is also, in part, adapted from a paper that appeared in Biotechnology and Bioengineering, Volume 107, Number 3, Pages 403-12, October 15, 2010. The dissertation author was the primary author of this paper which was coauthored by Bernhard Ø. Palsson.

We would like to thank Byung-Kwan Cho, Nathan Lewis, and Karsten Zengler for their helpful comments and insights.

Chapter 2 is, in part, adapted from a chapter that appeared in EcoSal – *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (www.ecosal.org), Chapter 10.2.1, ASM Press, Washington D.C., February 18, 2010. The dissertation author was the primary author of this chapter, which was coauthored by Ronan M.T. Fleming and Bernhard Ø. Palsson.

Chapter 2 is also, in part, adapted from a paper that appeared in Nature Biotechnology, Volume 28, Number 3, Pages 245-8, March 2010. The dissertation author was the primary author of this paper, which was coauthored by Ines Thiele and Bernhard Ø. Palsson.

We would like to thank Byung-Kwan Cho, Neema Jamshidi, Nathan Lewis, and Karsten Zengler for their helpful comments and insights.

Chapter 3 is adapted from a paper that appeared in Molecular Systems Biology, Volume 7, Article Number 535, October 11, 2011. The dissertation author was the primary author of this paper, which was coauthored by Tom M. Conrad, Jessica Na, Joshua A. Lerman, Hojung Nam, Adam M. Feist, and Bernhard Ø. Palsson.

We would like to thank Harish Nagarajan, Vasiliy Portnoy, Jennie Reed, and Ines Thiele for their helpful comments and insights.

Chapter 4 is, in part, adapted from a paper that appeared in Molecular Systems Biology, Volume 7, Article Number 535, October 11, 2011. The dissertation author was the primary author of this paper, which was coauthored by Tom M. Conrad, Jessica Na, Joshua A. Lerman, Hojung Nam, Adam M. Feist, and Bernhard Ø. Palsson.

Chapter 4 is also, in part, adapted from a paper that is being prepared for publication under the title "Gap-filling analysis of the *i*JO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions." The dissertation author was the primary author of this paper, which was coauthored by Bernhard Ø. Palsson.

We would like to thank Pep Charusanti, Tom Conrad, Adam Feist, Harish Nagarajan, Kenji Nakahigashi, Vasiliy Portnoy, Jennie Reed, and Ines Thiele for their helpful comments and insights. I would especially like to thank Professor Martin Robert for hosting me at the Institute for Advanced Biosciences in Tsuruoka, Japan and mentoring me while I performed *in vitro* enzyme assays and LC-MS analysis.

Chapter 5 is, in part, adapted from a paper that appeared in Metabolic Engineering, Volume 12, Number 3, Pages 173-86, May 2010. The dissertation author was a co-author of this paper, which was coauthored by Adam M. Feist, Daniel C. Zielinski, Jan Schellenberger, Markus J. Herrgard, and Bernhard Ø. Palsson.

Chapter 5 is also, in part, adapted from a paper that is being prepared for publication under the title "Adaptive laboratory evolution and characterization of computationally designed growth-coupled production strains." The dissertation author

was the primary author of this paper, which was coauthored by Adam M. Feist and Bernhard Ø. Palsson.

We would like to thank Christian Barrett, Tom Conrad, Ronan Fleming, Dae-Hee Lee, Harish Nagarajan, Vasiliy Portnoy, Ines Thiele, and Karsten Zengler for their helpful comments and insights.

Chapter 6 is, in part, adapted from a paper that is being prepared for publication under the title "Adaptive laboratory evolution and characterization of computationally designed growth-coupled production strains." The dissertation author was the primary author of this paper, which was coauthored by Adam M. Feist and Bernhard Ø. Palsson.

We would like to thank Mallory Embree, Dae-Hee Lee, Harish Nagarajan, Vasiliy Portnoy, Douglas Taylor, and Karsten Zengler for their helpful comments, insights, and help with experimental procedures.

VITA

June 2006	Bachelor of Science, Bioengineering University of Washington (Seattle, WA)
March 2012	Doctor of Philosophy, Bioengineering University of California, San Diego (La Jolla, CA)

PUBLICATIONS

Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BØ. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism – 2011. Molecular Systems Biology, 7:535. Oct 11 2011.

Lee SY, Sohn SB, Kim HU, Park JM, Kim TY, **Orth JD**, Palsson BØ. Chapter 1 - Genome-scale network modeling. Systems Metabolic Engineering (book). 2011.

Schellenberger J, Que R, Fleming RMT, Thiele I, **Orth JD**, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox 2.0. Nature Protocols. 6:1290–307. 2011.

Orth JD, Palsson BØ. Systematizing the generation of missing metabolic knowledge. Biotechnology and Bioengineering. 107(3):403-12. Oct 15 2010.

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*. Metabolic Engineering. 12(3):173-86. May 2010.

Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature Biotechnology. 28(3): 245-8. Mar 2010.

Orth JD, Fleming RMT, Palsson BØ. Chapter 10.2.1 - Reconstruction and Use of Microbial Metabolic Networks: the Core *Escherichia coli* Metabolic Model as an Educational Guide. EcoSal – *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. www.ecosal.org. ASM Press, Washington D.C. Feb 18 2010.

ABSTRACT OF THE DISSERTATION

Systems Biology Analysis of *Escherichia coli* for Discovery and Metabolic Engineering

by

Jeffrey David Orth

Doctor or Philosophy in Bioengineering

University of California, San Diego, 2012

Professor Bernhard Ø. Palsson, Chair

Systems biology is an emerging field of research that utilizes high-throughput experimental data and computational analysis methods to study biochemical networks. One of the common denominators of systems biology is the genome-scale reconstruction of metabolic networks. These reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledgebases that seek to formally represent the known metabolic activities of an organism. In the first part of this dissertation, the general properties of these reconstructions, along with the use of constraint-based modeling to analyze these networks, is described in detail. Metabolic network reconstructions have many practical uses, including use in discovery of new gene functions and metabolic reactions, and for metabolic engineering. The core *E. coli* metabolic model, a small-scale model that can be used for in-depth analysis of new constraint-based methods, is presented and analyzed in detail.

In the second part of the dissertation, the genome-scale metabolic reconstruction of *E. coli* was updated and analyzed. This reconstruction was first published in 2000, and has been updated and periodically published since then. The current version of the reconstruction is called *i*JO1366, and was updated based on new literature and database information. The remaining network gaps were analyzed, and a new gap-filling workflow was developed and used to predict the missing metabolic reactions and genes in *i*JO1366. Model predicted growth phenotypes were compared to a large experimental dataset of gene knockout strain phenotypes, and model errors were identified. Several *in vivo* experiments were performed to validate these model-based predictions.

In the third part of the dissertation, the metabolic network model of *E. coli* was used for metabolic engineering. A large-scale computational screen using the algorithms OptKnock and OptGene was performed to design many growth-coupled production strains for various useful compounds. In order to validate the accuracy of these predictions and to determine if adaptive laboratory evolution can be utilized as an effective strain engineering tool, two strain designs were selected for experimental analysis. These knockout strains were constructed and evolved to optimize their phenotypes. One evolved strain worked as expected, producing a high yield of lactate from glucose. The other failed to produce 1,2-propanediol as predicted. This strain was analyzed by expression profiling, providing evidence for another metabolic gene function.

Chapter 1: The use of systems biology methods to study bacterial metabolism

Genome-scale metabolic network reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledge bases that seek to formally represent the known metabolic activities of an organism. These reconstructions are based on the annotated genome of a particular organism, and they are organism specific. The genome-scale reconstruction of the metabolic network of *Escherichia coli* is one of the most complete and well-studied reconstructions available, and has been continuously updated and improved since it was first published in 2000. Microbial metabolic reconstructions have many practical purposes, including use in discovering new gene functions and metabolic reactions, and in guiding metabolic engineering strain design.

1.1 Biochemical network reconstructions and constraint-based modeling

One of the key developments of modern systems biology is the reconstruction of biochemical networks. A network reconstruction can be defined as a knowledgebase that describes all known components and interactions of a specific cellular system [1]. These reconstructions can be assembled for a variety of biological networks, and can be converted to mathematical models allowing for detailed theoretical analysis. Constraint-

based modeling is particularly useful for the study and analysis of metabolic networks and has led to many important insights, as described in this chapter.

1.1.1 Properties of a biochemical network reconstruction

There are currently several different types of biological network reconstructions representing various types of biological networks (metabolic, regulatory, transcription/translation, signaling), although metabolic network reconstructions have been the most extensively used. Early reconstructions of metabolism for model organisms such as the bacterium *Escherichia coli* were small, containing only central metabolic reactions [6]. Today, there are many genome-scale metabolic reconstructions available, based on all known metabolic genes in the annotated genome of an organism along with other data sources [7-10]. Metabolic network reconstructions are biochemically, genetically and genomically (BiGG) structured knowledgebases of biochemical reactions and metabolites. They contain information such as exact reaction stoichiometry, reaction reversibility, and the relationships between genes, proteins, and reactions. Reconstructed networks serve as flexible BiGG knowledge bases [13], storing curated information in a useful format while allowing for content to be updated based on new research. Although many organisms have similar central metabolic networks, there can be important differences even between two closely related organisms. Network reconstructions are therefore organism and strain specific [22]. For most applications of network reconstructions, it is necessary to convert the network into a mathematical model. Metabolic models are usually formulated as a stoichiometric matrix, while regulatory network models are often formulated as Boolean networks.

There have been many practical uses of network reconstructions. Some reconstructions have been used as tools to study bacterial evolution. The effects on metabolism of adding or removing genes from the network can be simulated, enabling studies of horizontal gene transfer [23], adaptation to new environments [24], and evolution to minimal genomes [25]. Reconstructions can also be used for analysis of network properties. In these studies, methods have been developed to determine the interactions between different sets of reactions and compounds, improving our understanding of the organisms under investigation. Some examples include identification of alternate optimal network states [26], identification of sets of coupled reactions [27], and studies of the states of regulatory networks [28, 29]. It has also been determined by simulating thousands of growth conditions that the *E. coli* metabolic network contains a set of common, high-flux backbone reactions [30]. Network reconstructions have been extensively used to study the phenotypic behavior of wild-type and mutant stains under a variety of conditions, linking genotypes with phenotypes. Some of these predictions have been verified by experimental studies [31]. Such phenotypic simulations have allowed for the prediction of growth after genetic manipulations [32, 33], prediction of growth phenotypes after adaptive evolution [34], and prediction of essential genes [35]. Another promising use of reconstructions is in the discovery of unknown biological features. By comparing experimental data such as growth phenotypes [11], metabolic flux measurements [14], or gene essentiality [12] to model based predictions, missing content in reconstructions can be identified. The reconstructions can be updated and new biological knowledge can be elucidated. Finally, network reconstructions have proven to be very useful for metabolic engineering and

synthetic biology [36]. Because of the capacity of models to be used to predict growth and metabolite secretion phenotypes, it is possible to predict the genetic interventions most likely to produce a strain with desired properties [37]. Model based algorithms can even predict non-intuitive designs that couple production of desired metabolites to cell growth [38, 39].

1.1.2 Constraint-based modeling of biochemical networks

Most types of biological models are theory-based, containing a set of equations that define a specific solution from a particular input. These types of models often require a large number of very precisely valued parameters, such as enzyme kinetic parameters. These kinetic parameters can be variable and are often difficult to measure [40, 41], and are thus not available for most biochemical reactions in most organisms. Even when parameters are available, they may suffer from measurement error or may not have been measured under the desired *in vivo* conditions [42-45]. Due to these limitations, no genome-scale theory-based models have been constructed yet [46].

Constraint-based modeling overcomes many of the limitations of theory-based modeling [47, 48]. Instead of defining a system that will give a single solution from a single input, constraint-based models use whatever parameters or limitations are known to define a space of possible solutions. There are many well defined constraints by which different types of biological networks are bound [49]. Physico-chemical constraints are imposed on all biological systems by the laws of physics and chemistry, including conservation of mass and energy, diffusion, and thermodynamics. These are considered “hard” constraints, as they apply to all organisms at all times. Networks also include

topological constraints such as partitioning of enzymes and metabolites into different subcellular compartments or organelles. Metabolites in two different compartments cannot participate in the same reaction. The environment that a cell or organism lives in also affects its capabilities. Environmental constraints will change at different times or under different conditions, and include the availability of nutrients and oxygen. Finally, there are regulatory constraints, in which an organism imposes constraints on itself through various mechanisms to adapt to a particular environment or achieve a particular objective. Regulatory constraints are organism specific and may change over time by evolution.

Although constraints define a range of solutions, it is still possible to identify and analyze single points within the solution space. For example, we may be interested in identifying which point corresponds to the maximum growth rate or maximum ATP production of an organism, given its particular set of constraints. Flux balance analysis (FBA) is one method for identifying such optimal points within a constrained space [50].

1.1.3 Flux balance analysis

FBA is a widely used constraint-based approach for studying biochemical networks, in particular the genome-scale metabolic network reconstructions that have been built in the past decade [4, 7, 8, 51]. FBA calculates the flow of metabolites through this metabolic network, thereby making it possible to predict the growth rate of an organism or the rate of production of a biotechnologically important metabolite. With metabolic models for over 50 organisms already available and high-throughput

technologies enabling the construction of many more each year [13, 52, 53], FBA is an important tool for harnessing the knowledge encoded in these models.

The first step in FBA is to mathematically represent metabolic reactions. The core feature of this representation is a tabulation, in the form of a numerical matrix, of the stoichiometric coefficients of each reaction (**Figure 1.1 a,b**). This is called the stoichiometric (**S**) matrix. These stoichiometries impose constraints on the flow of metabolites through the network. Constraints such as these lie at the heart of FBA, differentiating the approach from theory-based models based on biophysical equations that require many difficult-to-measure kinetic parameters [47, 48].

Constraints are represented in two ways, as equations that balance reaction inputs and outputs and as inequalities that impose bounds on the system. **S** imposes flux (that is, mass) balance constraints on the system, ensuring that the total amount of any compound being produced must be equal to the total amount being consumed at steady state (**Figure 1.1 c**). Every reaction can also be given upper and lower bounds, which define the maximum and minimum allowable fluxes of the reactions. These balances and bounds define the space of allowable flux distributions of a system—that is, the rates at which every metabolite is consumed or produced by each reaction. Other constraints can also be added [49].

The next step in FBA is to define a phenotype in the form of a biological objective that is relevant to the problem being studied (**Figure 1.1 d**). In the case of predicting growth, the objective is biomass production, the rate at which metabolic compounds are converted into biomass constituents such as nucleic acids, proteins, and lipids. Mathematically, an objective is represented by an “objective function” that

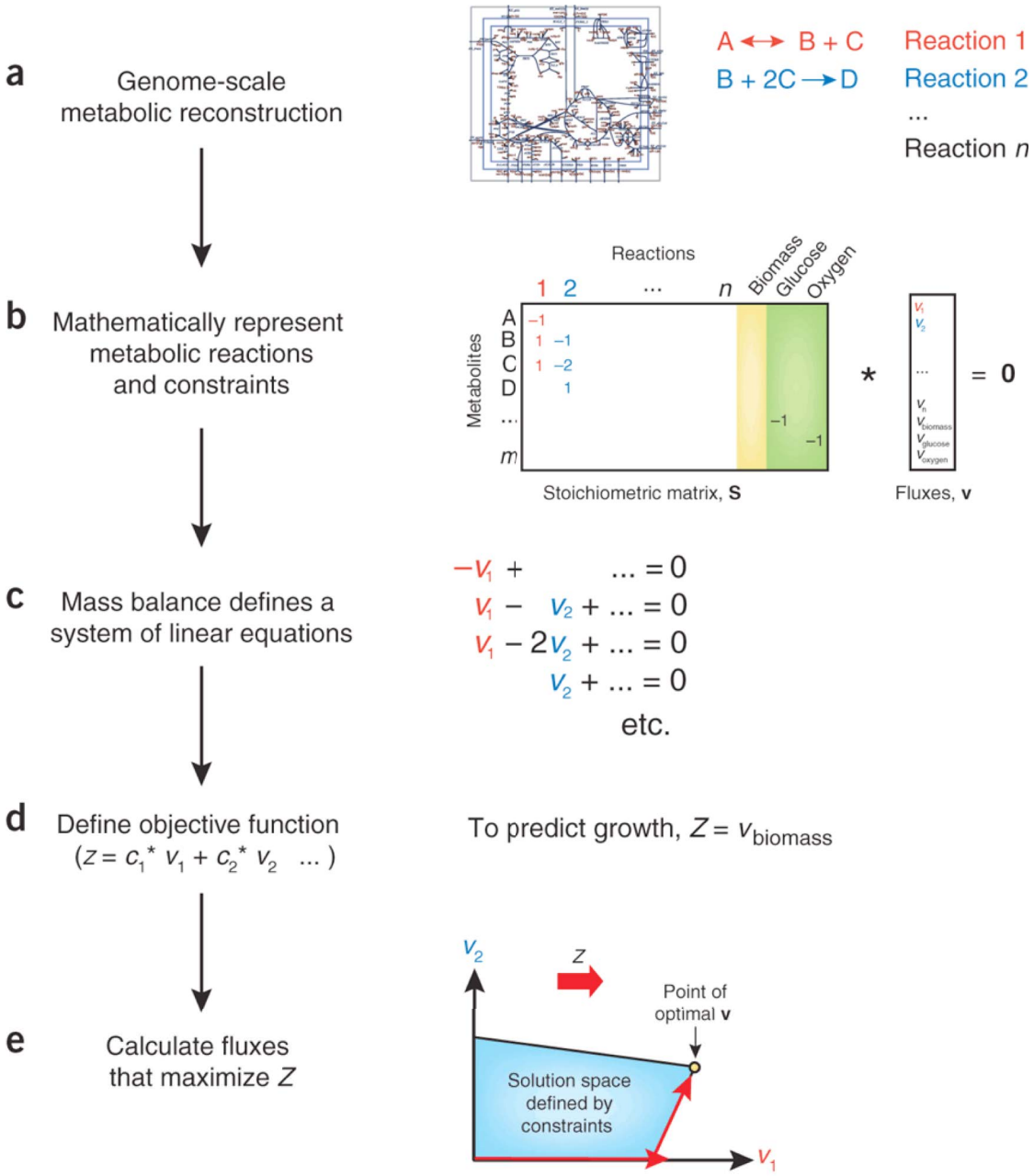

Figure 1.1 Formulation of an FBA problem. (**a**) First, a metabolic network reconstruction is built, consisting of a list of stoichiometrically balanced biochemical reactions. (**b**) Next, this reconstruction is converted into a mathematical model by forming a matrix (labeled **S**) in which each row represents a metabolite and each column represents a reaction. (**c**) At steady state, the flux through each reaction is given by the equation $\mathbf{Sv} = 0$. (**d**) An objective function is defined as $Z = \mathbf{c}^{T}\mathbf{v}$, where **c** is a vector of weights (indicating how much each reaction contributes to the objective function. (**e**) Finally, linear programming can be used to identify a particular flux distribution that maximizes or minimizes this objective function while observing the constraints imposed by the mass balance equations and reaction bounds.

indicates how much each reaction contributes to the phenotype. A "biomass reaction" that drains precursor metabolites from the system at their relative stoichiometries to simulate biomass production is selected as the objective function in order to predict growth rates. The biomass reaction is based on experimental measurements of biomass components. This reaction is scaled so that the flux through it is equal to the exponential growth rate (μ) of the organism.

Taken together, the mathematical representations of the metabolic reactions and the objective define a system of linear equations. In FBA, these equations are solved using linear programming (**Figure 1.1 e**). Many computational linear programming algorithms exist, and they can very quickly identify optimal solutions to large systems of equations. The COBRA Toolbox [54] is a freely available Matlab toolbox for performing these calculations.

Although growth is easy to experimentally measure, computational approaches such as FBA shine in simulations to predict metabolic reaction fluxes and simulations of growth on different substrates or with genetic manipulations. FBA does not require kinetic parameters and can be computed very quickly even for large networks, so it can be applied in studies that characterize many different perturbations. FBA has limitations, however. Because it does not use kinetic parameters, it cannot predict metabolite concentrations. It is also only suitable for determining fluxes at steady state. Except in some modified forms, FBA does not account for regulatory effects such as activation of enzymes by protein kinases or regulation of gene expression, so its predictions may not always be accurate.

1.2 The metabolic network reconstruction process

The construction of a genome-scale metabolic network is a long-term process (from many months to several years, depending on the size of the network) consisting of four major steps, each requiring the use of different types of biological data (**Figure 1.2**). In the first step, an organism's annotated genome is used to generate a draft reconstruction. Second, the draft reconstruction is curated in a long process that involves the study of many highly specific data sources. In the third step, the reconstruction is converted to a mathematical model, and model based simulations can be compared to phenotypic data. In the fourth and final step, high-throughput data can be integrated with the model, allowing for biological discovery and iterative model refinement.

1.2.1 Initial reconstruction based on an annotated genome

The first step in building any new genome-scale reconstruction, whether of metabolic or regulatory networks, is to use the genome annotation for the desired organism to generate an initial list of functions. Genomes have been annotated to some degree for over 1000 microbial organisms, and these annotations provide several types of valuable information. First, they contain the genome sequence of an organism, from which open reading frames (ORFs) can be identified. The function of each ORF can then be determined through a variety of methods [11]. The strongest evidence for the function of a particular ORF usually comes from direct biochemical analysis, such as isolation and characterization of the function of an enzyme. *E. coli* is a very well-studied model organism and many of its ORFs have been experimentally characterized [55]. Unfortunately, for many other organisms, little biochemical data is available. To identify

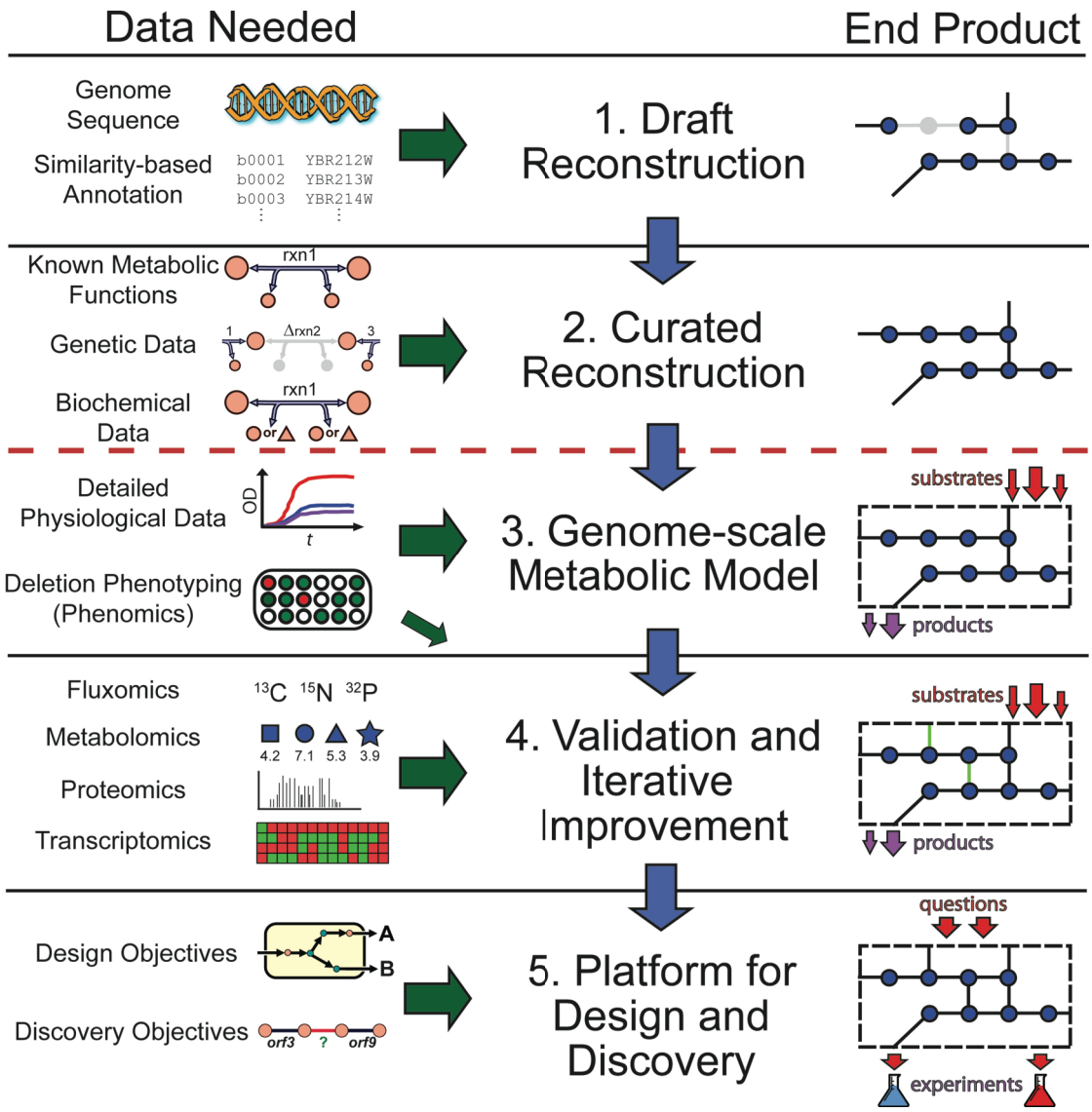

Figure 1.2 The phases and data utilized in generating a metabolic reconstruction. The genome-scale metabolic reconstruction process can be summarized in four major phases, each of the latter phases building off the previous. The fifth phase is use of the complete reconstruction for practical purposes. Characteristic of the reconstruction process is the iterative refinement of reconstruction content that is driven by experimental data and occurs in phases 2-4. For each phase, specific data types are necessary and these range from high-throughput data types (e.g., metabolomics) 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 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. Figure adapted from [4].

the ORFs in these organisms (and also to identify uncharacterized ORFs in well studied organisms), their genomic sequences are compared to the genomes of other organisms to identify homologous genes. *In silico* methods can also be used for annotation, including methods that identify genes based on protein-protein interactions, transcriptomics, phylogenetic profiles, protein fusions, and operon clustering [56]. These methods typically allow for 40-70% of the genes in a new genome to be annotated. When a high quality genome annotation is not available for a particular organism, it becomes more challenging to build a reconstruction of that organism on a genome-scale.

There are organism specific databases for some genome annotations, including EcoCyc [55] for *E. coli* and *Saccharomyces* Genome Database (SGD) [57] and Comprehensive Yeast Genome Database (CYGD) [58] for yeast. For many other microbial organisms [59], Comprehensive Microbial Resource (CMR) [60], Genome Reviews [61], and Integrated Microbial Genomes (IMG) [62] contain useful genomic information. All of the metabolic genes identified in the genome annotation for the desired organism can be assembled into an initial parts list. From this initial list of genes, an initial list of enzymes and the reactions they catalyze can be constructed by mapping each gene to one or more reactions according to information from a database. The data used for this mapping can be included in the genome annotations or it can be obtained from metabolic databases such as KEGG [63], BRENDA [64], ENZYME [65], MetaCyc [66], the SEED [67], or TransportDB [68]. Most databases include EC (Enzyme Commission) numbers that can be used to easily identify the enzymes and reactions associated with a particular gene. With the appropriate databases, the reactions known to be associated with each gene can be identified. Some of these databases will be more

useful for certain organisms than others. The process of building an initial reconstruction from a genome annotation and reaction information from databases can be performed manually, or it can be partially or fully automated. Tools available for building draft reconstructions include SimPheny (Genomatica, Inc., San Diego, CA), PathoLogic [69], and PRIAM [70]. SimPheny is a commercial software platform for building and analyzing metabolic constraint-based models. It can download the annotated genome of an organism and provides a framework for manually associating metabolic genes with reactions. SimPheny is also useful in the other stages of building reconstructions, as it contains tools for manual curation and model quality control and quality assurance, as well as tools for performing simulations and analyzing experimental data. PathoLogic, part of the Pathway Tools software system, is tool for mapping genes to reactions in an automated manner. It requires a fully annotated genome, and uses EC numbers, Gene Ontology terms [71], or the annotated gene names to identify which reactions are associated with a particular gene. It can then predict which pathways are present in an organism by comparing the predicted set of reactions to a reference database such as MetaCyc. PRIAM is another automated method that identifies enzymes in any genome sequence. This program uses all of the known sequences for any individual enzyme in the ENZYME database to identify the characteristic sequence modules of that enzyme. Specific rules that can identify an enzyme based on which modules are present in a sequence are then formulated. PRIAM forms these modules and rules for every enzyme in the database, and then uses scoring matrices to identify modules in the genome of interest. It then uses the rules to predict which enzyme is associated with every gene in the genome. This algorithm can be very useful because it does not require a fully

annotated genome. The result of the initial mapping process is a draft reconstruction that lists most of the metabolic genes and reactions in an organism with reasonable accuracy.

1.2.2 Curation of the initial reconstruction

The next step in the reconstruction process is to manually curate the initial reconstruction. Any reconstruction resulting from a fully automated procedure will be incomplete and in some cases incorrect. Some reactions will be missing because it is not known which genes encode their enzymes, leaving gaps in pathways. Other reactions may be mistakenly included due to incorrect genome annotation or nonspecific information in databases. Reactions in the reconstruction may have incorrect or unbalanced stoichiometries or cofactor usage, as these attributes are often unique to enzymes in specific organisms [22]. Gene-protein-reaction associations (GPRs) must also be included to formally connect reactions to one or more functional proteins, and every protein to one or more known genes [72]. To correct any mistakes and improve the reconstruction, a researcher must manually curate the list of reactions using data from many different sources. Organism specific textbooks and databases are very useful for this purpose. The genome-scale *E. coli* reconstruction *i*AF1260 [8] relied heavily on the textbook Escherichia coli *and* Salmonella*: Cellular and Molecular Biology* [73]. Unfortunately, such texts are usually not available for less well-studied organisms. Literature data from both primary and review articles is also extremely useful. These articles can contain useful and specific information about reaction stoichiometry and directionality and can indicate the presence of many reactions with unknown genes. There are many different types of studies that can be useful, including enzyme assays,

gene knockout studies, metabolomic studies including flux measurements, and protein localization studies. Often, the information from these sources cannot be found in databases. The manual curation process is extremely labor intensive, usually requiring the study of hundreds of literature sources over a period of several months or years.

1.2.3 Conversion of the reconstruction to a computational model

After a high quality reconstruction has been assembled, it must be converted into a genome-scale constraint-based model to be further analyzed [1]. A reconstruction is a BiGG knowledge base, a list of stoichiometrically balanced reactions and their associated genes and proteins. A model is a network in a mathematical format with defined system boundaries and constraints on the reactions [6]. While a reconstruction is unique to an organism, many different models (i.e., condition specific) can be derived from a reconstruction. Metabolic network models are usually encoded in an **S** matrix, in which each unique metabolite is represented by a row in the matrix and each reaction is represented by a column. The entries in each column are the stoichiometric coefficients of the metabolites in a reaction, with negative coefficients for consumed metabolites and positive coefficients for produced metabolites. The properties of this matrix can be investigated through various constraint-based analysis methods, including FBA [6, 50, 74-76]. In order to simulate growth with a genome-scale metabolic model, a biomass reaction is needed [6, 75]. To construct a biomass reaction, the relative amounts of nucleic acids, lipids, proteins, and other macromolecules of an organism must be known. These macromolecules can then be broken down into building blocks such as nucleotides or amino acids. The relative amounts of each of these building blocks give the

stoichiometric coefficients of the biomass reaction. Experimentally determined growth data must also be used to determine the amount of energy needed for growth and for non-growth associated maintenance functions, representing the energy demands of the cell [8]. Once the biomass reaction has been constructed, FBA can be used to predict optimal growth rates under many different conditions [76]. A completed genome-scale metabolic model can be used to assess the quality of the reconstruction. There must be continuous pathways to every biomass precursor or the *in silico* cell cannot grow. Other gaps in the network can be identified by unreachable reactions or metabolites [77]. Another common test of a new model is to compare growth simulations under different conditions to experimental growth phenotype data [31].

1.2.4 Model validation and iterative improvement

Genome-scale metabolic models can be used to map many types of experimental data to a biological network, allowing for the integration of different data types. This data can be compared to model predictions, and the discrepancies can lead to discovery of new reactions and pathways. High throughput screens for growth on different media conditions can be used to reveal previously unknown substrate uptake pathways. *In vivo* knockout screens and synthetic lethal screens can be compared to *in silico* predicted knockout effects (**Figure 1.3**), with discrepancies indicating the presence of alternative metabolic enzymes or pathways [11]. Metabolomic data can predict the presence of metabolites not accounted for in a reconstruction, necessitating the addition of new production and utilization pathways. Proteomic and transcriptomic data can be used to suggest the genes and proteins that fill gaps in a reconstruction [78]. As new biological

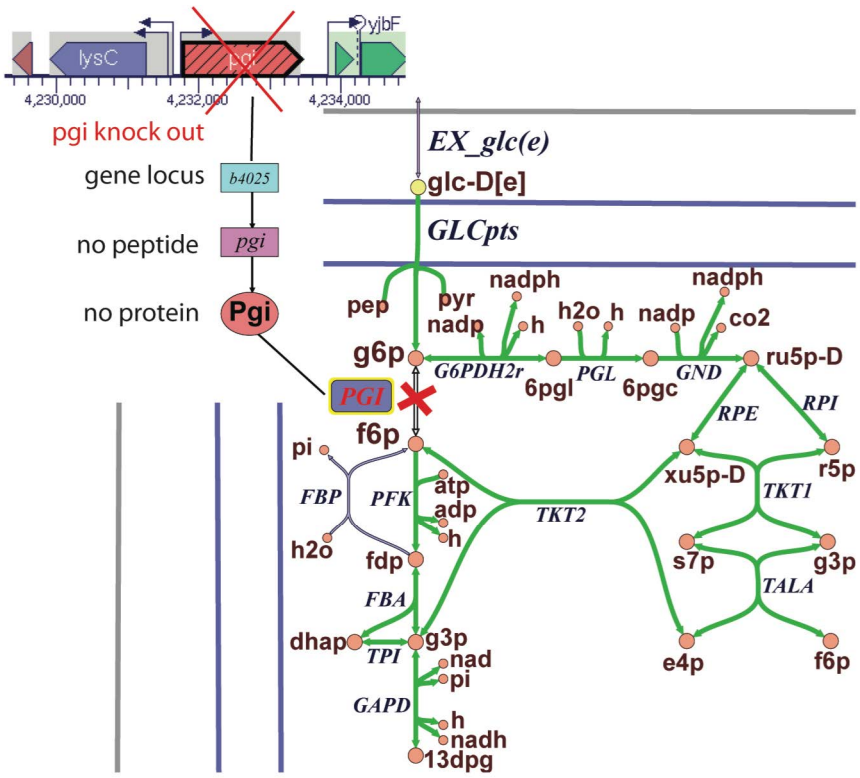

Figure 1.3 An *in silico* knockout of glucose-6-phosphate isomerase (*PGI*), which catalyzes the conversion of D-glucose-6-phosphate (g6p) to D-fructose-6-phosphate (f6p). The Δ*pgi in silico* model predicts that loss of glucose-6-phosphate isomerase results in use of the pentose phosphate pathway as the primary route of glucose catabolism, as has been observed experimentally [3].

features and capabilities are discovered, the model can be improved by incorporating this new data. The updated model can then be used to probe different pathways and features, leading to an iterative cycle of discovery and model improvement.

1.3 Building genome-scale regulatory network reconstructions

The process of reconstructing transcriptional regulatory networks is not as well developed as the process for building metabolic reconstructions. To date, there are only a

few examples of genome-scale regulatory reconstructions [79-81]. The oldest of these is *i*MC1010, which contains the regulators of *E. coli* metabolism [82]. This model was built by a process similar to that used for metabolic reconstructions, relying on a variety of data types including the genome annotation and literature sources. A gene expression study was then conducted in which gene expression under aerobic and anaerobic conditions was compared in several different transcription factor knockout strains. By analyzing the discrepancies between the experimental results and the model predictions, the model was iteratively updated and improved.

Several automated methods for inferring transcriptional regulatory networks have been developed recently [79, 83-85]. Progress in high throughput experimental methods is allowing for transcriptional regulatory reconstructions to be assembled in a top-down, automated manner [86-88]. The connectivity of an organism's transcriptional regulatory network can be determined by performing ChIP-chip experiments (chromatin immunoprecipitation followed by microarray hybridization). In these studies, all of the DNA binding sites of a particular transcription factor on an entire genome under a particular set of conditions can be identified *in vivo*. First, proteins are fixed to genomic DNA in living cells and then the DNA is extracted and sheared into small fragments. Next, the fragments with a particular transcription factor bound are extracted using antibodies, and the fragments are identified by hybridization to a DNA microarray. When ChIP-chip experiments are repeated under a variety of conditions, all of the binding sites of a transcription factor can be found, identifying all of the genes regulated by that transcription factor.

A set of ChIP-chip experiments must be run for every transcription factor to elucidate an entire regulatory network. As useful as these experiments are, they do not reveal what the effect of a transcription factor on its targets is, or how different transcription factors interact. The direct and indirect effects of transcription factors can be determined by expression profiling strains with those transcription factor genes knocked out [89]. Performing so many high throughput experiments is a very expensive and time consuming process, but improvements in parallel sequencing technologies should improve the effectiveness of these approaches. ChIP-seq experiments (chromatin immunoprecipitation followed by DNA sequencing) are an example of this trend [86, 87, 90-92].

As with metabolic reconstructions, regulatory reconstructions must be converted to computational models to utilize their predictive potential. There are several different model types that can be used. The most common modeling framework is the Boolean model, which represents the connections between genes or other variables as logical rules. Boolean models can qualitatively describe the functions of a regulatory network and make accurate predictions of behavior [93]. An equivalent structure to the Boolean model is the “regulation matrix” [94]. By reformulating a Boolean model as a matrix, more advanced mathematical analysis is possible, and all of the possible expression states of the model can be sampled [95]. Newer regulatory network models are being formulated in structures similar to the stoichiometric matrix, allowing them to be interrogated with constraint-based analysis methods [96].

1.4 The history of the *Escherichia coli* metabolic network reconstruction

To date, the most complete network reconstruction of any organism is the one for the metabolic network of *E. coli*. Just as this bacterium is a widely used model organism in *in vivo* experiments, its reconstruction has been used as a model for many new types of constraint-based analysis [4]. The first metabolic reconstructions were assembled about 20 years ago and have been updated and improved many times since then. Reconstructions for other cellular systems of *E. coli* (transcription and translation [96], the transcriptional regulatory network, and two-component signaling) have also recently been assembled.

1.4.1 Pre-genome-scale reconstructions

The first constraint-based model of parts of *E. coli* metabolism was constructed in 1990 by Majewski and Domach [97]. This simple model contained parts of glycolysis and the TCA cycle, and was used to study aerobic acetate production. A more complete model of the central metabolism of *E. coli*, including glycolysis, the pentose phosphate pathway, the TCA cycle, oxidative phosphorylation, and a biomass reaction for simulating growth, was constructed by Varma and Palsson in 1993 [6, 75], and was later expanded to include the synthesis pathways for amino acids and nucleotides [98]. These models were characterized extensively by FBA and other constraint-based analysis methods. Network properties such as maximal yields of precursors and cofactors were analyzed [6], and growth and by-product secretion phenotypes under varying oxygenation states were predicted [76]. In 1997, an even more extensive model was constructed by Pramanik and Keasling, accounting for cofactor biosynthesis and a

growth-dependent biomass objective function [99]. This model was used to predict internal fluxes for growth on different combinations of glucose and acetate.

1.4.2 Genome-scale reconstructions

Once the genome of *E. coli* was completely sequenced and annotated in 1997, a genome-scale metabolic network reconstruction could finally be assembled to account for the reactions catalyzed by all known *E. coli* metabolic enzymes. The first such reconstruction was *i*JE660, published in 2000 [74]. This model includes the products of 660 genes with 627 reactions and 438 metabolites, and was constructed through extensive searches of literature and databases to ensure correct stoichiometry and cofactor usage. Many phenotypic studies were conducted on this model, including studies of gene essentiality, robustness analysis, and phenotypic phase planes [31]. Many of these studies were compared to experimental data, confirming that the model predictions were correct. In 2003, an updated version of this model, *i*JR904, was published [100]. This model had an expanded scope, including pathways for the consumption of alternate carbon sources and more specific quinone usage in the electron transport system. Genes and proteins were explicitly connected to metabolic reactions with Boolean rules. Extensive quality control/quality assurance checks were performed to ensure that all of the reactions in the model were elementally and charge balanced. Gaps in the model were identified and filled when possible. This model has 931 reactions, 625 metabolites, and 904 genes. A transcriptional regulatory network in the form of Boolean rules was added to this metabolic model to form the combined metabolic and regulatory model *i*MC1010, published in 2004 [82]. This combined model was found to be more accurate than the

metabolic model alone at predicting phenotypes with different gene knockouts and environmental conditions.

In 2007, the *E. coli* metabolic reconstruction was updated again, this time called *i*AF1260 [8]. The scope of the model was expanded again, now including many reactions for the synthesis of cell wall components. Lumped reactions were separated into distinct enzymatic reactions, and the periplasm was accounted for as a distinct compartment. The thermodynamic properties of each reaction were calculated using the group contribution method [101], and this was used to set lower bounded on predicted irreversible reactions. *i*AF1260 contains 2077 reactions, 1039 metabolites, and 1260 genes. This model and its predecessors have been used in many different applications, divided into five categories: metabolic engineering, biological discovery, studies of phenotypic behavior, network analysis, and studies of bacterial evolution [4]. A modified subset of *i*AF1260 consisting of glycolysis, the pentose phosphate pathway, the TCA cycle, nitrogen metabolism, fermentation, and a simplified electron transport system and oxidative phosphorylation was constructed and published in early 2010 (see **Chapter 2: The core *Escherichia coli* metabolic network reconstruction**). This core *E. coli* metabolic model is useful for educational purposes and for testing new constraint-based tools. The genome-scale metabolic model of *E. coli* was expanded again in 2011 to *i*JO1366 (see **Chapter 3: Updating the genome-scale metabolic network reconstruction of *Escherichia coli*, *i*JO1366**). Unlike previous updates, this one did not significantly expand the scope of the model; it was mainly a refinement of *i*AF1260. New genes and reactions characterized since 2007 were added, and the GapFind algorithm was used to identify all gaps and blocked pathways in the model. These gaps were then manually curated to separate the

true knowledge gaps from the scope gaps, and literature and database searches were conducted to identify potential gap-filling reactions. Low confidence reactions in *i*AF1260 were also investigated and updated when possible. A total of 254 new reactions, 150 new metabolites, and 107 new genes were added to *i*JO1366, which is now the most complete version of the *E. coli* metabolic network reconstruction.

1.4.3 Reconstructions beyond metabolism

The entire transcriptional and translational (tr/tr) process of *E. coli* was reconstructed and represented in the form of the Expression matrix (E-matrix) [96]. This reconstruction contains the essential components and processes of the tr/tr network, and is the largest reconstructed network to date, with 11,991 components and 13,694 reactions. The reconstruction was converted into a mathematical model and used to quantitatively integrate various high-throughput data sets such as substrate uptake rates and gene expression measurements, and to compute functional network states of the data-constrained network and compare them to independent observations. Computed ribosome production rates were consistent with observed rates [96]. *In silico* simulations of rRNA operon deletions predict that a high RNA polymerase density on the remaining rRNA operons is needed in order to reproduce the reported experimental ribosome numbers [96]. This first comprehensive reconstruction of the transcriptional and translational machinery of *E. coli* will promote further understanding of the genotype-phenotype relationship and is expected to have a similar impact on systems biology to the impact already made by the metabolic reconstructions

The model of the transcriptional regulatory network of *E. coli* was originally constructed as a set of Boolean rules that were connected to the metabolic reactions of the *i*JR904 model [82, 100]. This network description was converted from a Boolean description to an equivalent matrix format, called the Regulation matrix (R-matrix) [94]. The network was also updated based on recent literature, and currently contains 147 external stimuli, 125 transcription factors, and 503 regulated downstream metabolic network genes [95]. By constructing the TRN as a matrix, many of the matrix analysis methods originally developed for constraint-based metabolic networks could be used. Random sampling was used to characterize all possible states of this network, and results were compared to gene expression data [95].

A reconstruction of the two-component signaling systems (2CSSs) of *E. coli* has recently been built. These relatively simple signaling systems allow *E. coli* to respond to new environments and stresses by activating various transcription factors, although the exact inputs to some 2CSSs are unknown at this time [102-104]. This network contains the sensor kinases and response regulators of the 22 2CSSs of *E. coli*, represented by 128 components and 125 reactions. This reconstruction has not been published yet, but it can be integrated with the TRN reconstruction to provide input signals to many of the transcription factors in that network.

1.5 Gap-filling of metabolic networks

Even the most complete models are not perfect; they all contain gaps, or missing information. There are two types of missing information in metabolic network

reconstructions. The first type is gaps in a network, places where a reaction that consumes or produces a metabolite is missing, creating a dead-end. The second type is orphan reactions. These are reactions that are known to exist, but it is not yet known which gene or genes code for their enzymes. Both types of gaps are the result of our incomplete knowledge of the metabolism of an organism. In recent years, methods have been developed to predict which genes or reactions should be associated with the gaps in metabolic network reconstructions. These gap-filling methods have the potential to be very useful because they can improve the predictive capabilities of models by making them more realistic while simultaneously improving our knowledge of an organism by characterizing a previously unknown gene. Gap-filling is therefore a discovery tool as well as a model refinement tool. Some of the various computational methods that are used to fill gaps in metabolic networks are summarized in **Table 1.1** and **Figure 1.4**.

1.5.1 Types of gaps and orphans

There are two fundamentally different types of missing information in metabolic network reconstructions (**Figure 1.5**). First, there are gaps; holes in the network where a reaction that should occur in the organism is absent. These gaps are manifest in network reconstructions as dead-end metabolites, which have a producing reaction but no consuming reaction. It is also possible for a metabolite to be consumed by a reaction but not produced by any reactions. These two types of dead-ends have been termed *root no-consumption metabolites* and *root no-production metabolites* [5]. When simulating steady-state reaction fluxes in these networks (as in FBA), the reactions producing or consuming dead-end metabolites can never carry flux under any conditions. These

Table 1.1 Summary of *in silico* gap-filling and orphan-filling methods.

Method	Gaps or orphans?	Data needed (other than the original metabolic network)	References
BNICE	gaps	a set of generalized enzyme reactions	[2]
GapFill	gaps	database of potential reactions (e.g., MetaCyc)	[5]
SMILEY	gaps	growth phenotype data (e.g., Biolog), database of potential reactions (e.g., KEGG)	[11]
GrowMatch	gaps	gene essentiality data, database of potential reactions (e.g., MetaCyc)	[12]
OMNI	gaps	metabolic flux data, database of potential reactions (e.g., KEGG)	[14]
PHFiller-GC	orphans	annotated gene/protein sequences from other organisms (e.g., Swiss-Prot, PIR), genomic-context data	[15, 16]
SEED	orphans	annotated genome sequences from other organisms (the SEED)	[17, 18]
ADOMETA	orphans	Co-expression data (e.g., SMD), annotated genome sequences from other organisms (e.g., GenBank), sets of gene clusters (e.g., KEGG)	[19-21]

reactions are said to be "blocked," and any reactions upstream or downstream from them will also be unable to carry flux under steady-state conditions. An extreme case of a dead-end metabolite is one that has no producing or consuming reactions, and thus forms an isolated island [105].

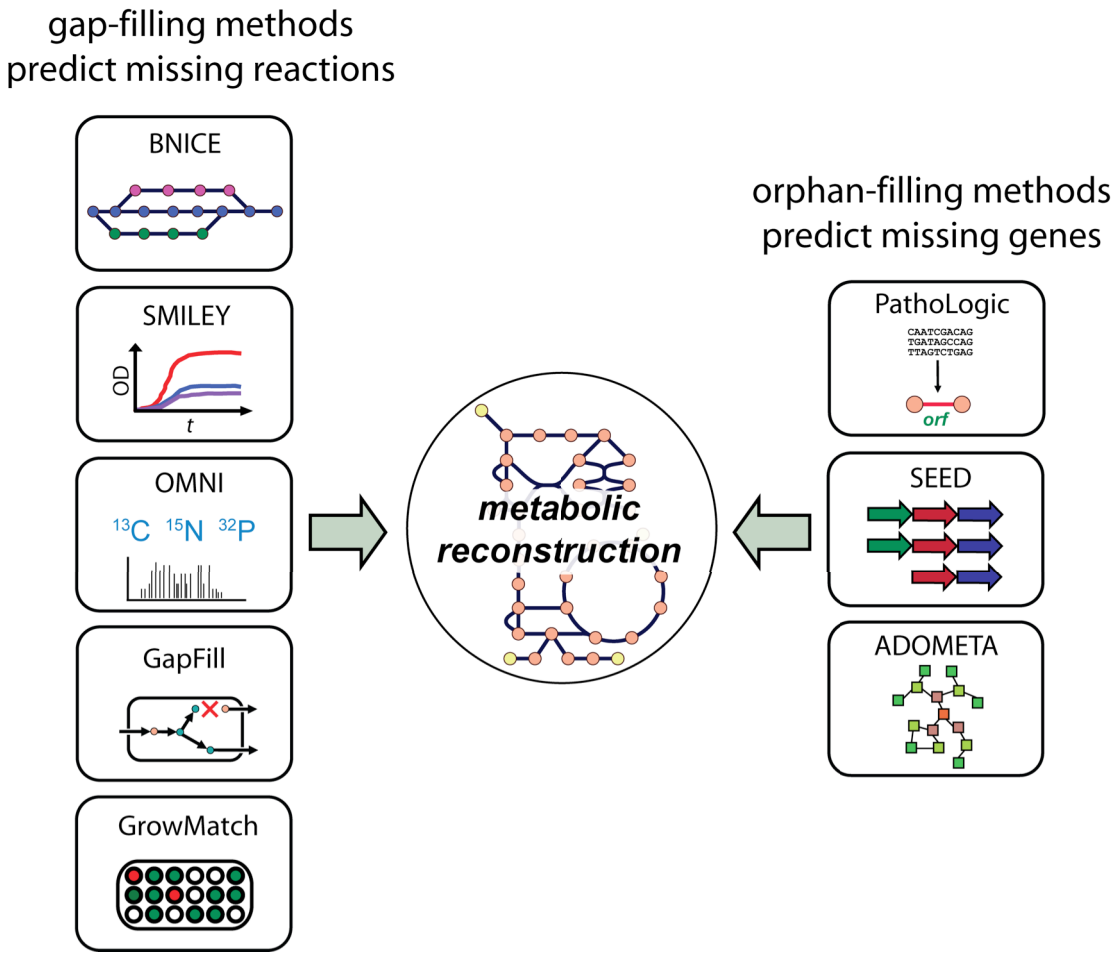

Figure 1.4 Overview of the different algorithms used for predicting gap-filling reactions and orphan-filling genes.

There are several reasons why there may be gaps in a metabolic network. First, it is possible that the actual biochemical network is missing an enzyme that is part of a completed pathway in related organisms. An example of this situation is in the O-antigen synthesis pathways of many *E. coli* K-12 strains, in which an IS5 insertion element has disrupted a rhamnosyltransferase gene, inactivating it [106, 107]. These strains are therefore unable to produce a functioning rhamnosyltransferase enzyme, blocking the entire downstream pathway that produces O-antigen. In cases such as this, there really is

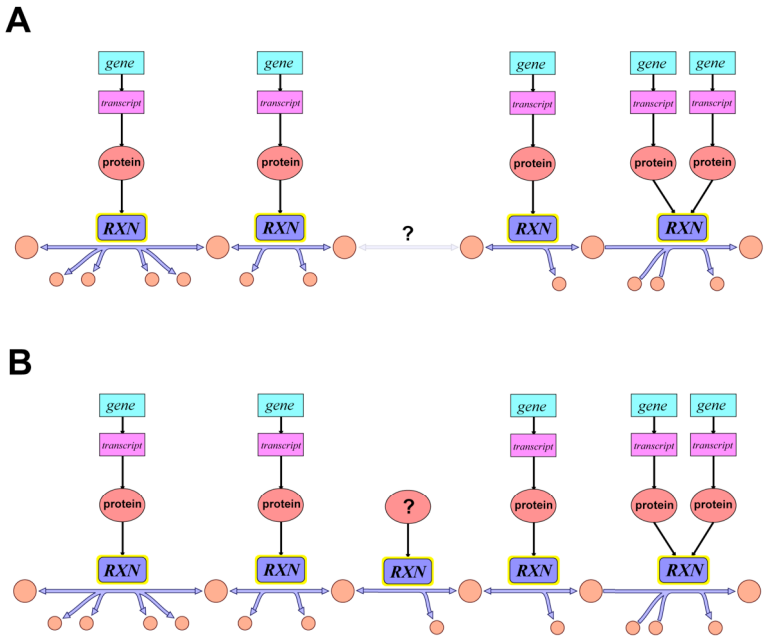

Figure 1.5 Examples of a gap (**a**) and an orphan reaction (**b**) in metabolic networks. Gaps occur when the reaction that consumes or produces a particular metabolite is completely unknown. Orphans occur when a particular reaction is known to occur, but it is not known which gene encodes the catalyzing enzyme. In this figure, metabolites are represented by orange circles and reactions by blue arrows connecting them. Gene-protein-reaction interactions are shown for each reaction.

a gap in the biochemical network, known as a *biological gap*. The *E. coli* K-12 MG1655 reconstructions *i*AF1260 and *i*JO1366 [8] contain a biologically realistic gap, and this pathway is unable to carry flux. Another reason for gaps in network models is the scope of the models themselves. To date, most large-scale metabolic network models do not include other systems such as signaling or transcription and translation. Metabolites that are produced in metabolism but then enter these other systems may be left as gaps in models, even though their biological functions are known. These are *scope gaps*. An example of scope gaps are the tRNAs in *i*AF1260. This model contains tRNA charging reactions but has no consuming reactions for these tRNAs even though it is well known how charged tRNAs are used in the process of translation. Finally, there may be a gap

because it is not known what biochemical reaction produces or consumes a certain metabolite. In this case, the gap is not biologically realistic; it is the result of limited knowledge. These are called *knowledge gaps*, and to fill them, new biological discoveries must be made.

The second category of missing information in a metabolic network is orphan reactions. These are biochemical reactions that are known to occur but are catalyzed by an unknown gene product. Reactions can be identified without genes through several different types of evidence, including biochemical assays of crude cell extracts, an observed phenotype such as uptake or secretion of a particular substrate, or the implied presence of a reaction due to the presence of other genes in a conserved pathway. Even the most well studied organisms have many genes with unknown functions, and many of these genes may code for orphan reactions. *E. coli* K-12 MG1655, for example, still has 981 partially or fully uncharacterized genes according to EcoCyc (version 13.6) [108]. It is also possible that some genes with currently known functions may also catalyze orphan reactions, so these genes must not be ignored when attempting to identify the gene associated with an orphan. Orphan reactions can be local, where the gene is unknown in one organism but known in at least one other. They can also be global orphan reactions, in which there are no known genes in any organism that code for a catalyzing enzyme. It has been determined that 30-40% of all known enzymatic activities are global orphans [109-111]. Global orphans are evenly distributed across the different types of known biochemical reactions [112]. Identifying the gene responsible for an orphan reaction in one organism increases the probability of identifying genes in other organisms, since it would provide a reference sequence for homology search methods. A database called

ORENZA (ORphan ENZyme Activities) at *http://www.orenza.u-psud.fr/* lists currently known global orphan reactions [113].

1.5.2 Methods for predicting gap-filling reactions

BNICE

When a gap exists in a biochemical network, there are potentially many sets of reactions that might fill the gap and restore connectivity. An algorithm called BNICE (Biochemical Network Integrated Computational Explorer) was developed to identify all biochemical reactions that could realistically link two metabolites [2]. In this algorithm, molecules are represented as bond-electron matrices (BEMs) and reactions are represented by matrix addition. It is possible to model the activity of an enzyme catalyzed reaction occurring with alternative substrates by adding the same reaction matrix to different BEMs, resulting in different product BEMs from each one. The enzyme catalyzed reactions in the KEGG database [63, 114] were analyzed, and it was found that there are fewer than 250 known generalized enzyme reactions, each of which can be represented by a reaction matrix. These generalized enzyme reactions typically correspond with third level Enzyme Classification (EC) numbers (e.g., EC 1.1.1._), while the fourth level EC numbers indicate the particular substrates and products of a reaction. BNICE then iteratively searches for all possible pathways of chemical transformations from one metabolite to another, offering gap-filling hypotheses. The pathways for synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan were analyzed with BNICE restricted to use only the generalized enzymatic reactions already known to occur in these pathways. Several dozen alternative pathways were identified

using these reaction mechanisms alone, some including metabolites not found in the KEGG or Chemical Abstracts Services databases. Thermodynamic analysis of the predicted pathways using the group contribution method showed that the currently known native pathways are the most favorable [2]. Although gap-filling predictions generated by BNICE have not yet been experimentally verified, it remains a potentially useful tool.

GapFind and GapFill

The algorithms GapFind and GapFill were developed to attempt to minimize the total number of gaps in a metabolic network model [5]. GapFind is a mixed-integer linear programming (MILP) algorithm that can identify every gap in a network by identifying blocked metabolites (i.e., those that cannot be produced or consumed at steady state under any conditions). Although it may be easy to identify some gaps by inspection of network or pathway maps, certain unusual model structures such as cycles can lead to non-intuitive sets of blocked metabolites, so GapFind can be very useful (**Figure 1.6**). In the *E. coli* model *i*AF1260, for example, the compounds in the ubiquinol-8 synthesis pathway are blocked even though the pathway appears to have a clear entry and exit. This is because ubiquinol-8 and ubiquinone-8 are interconverted in several reactions, but there are no reactions that consume either of these compounds, so no additional ubiquinol-8 can be produced at steady state [115]. GapFill is another MILP algorithm, and its objective is to minimize the total number of gaps by reversing the directionality of existing reactions, adding new reactions (from the MetaCyc database [116]), adding transport reactions for blocked metabolites, or adding intracellular transport reactions between compartments for multi-compartment models. GapFill attempts to reduce the

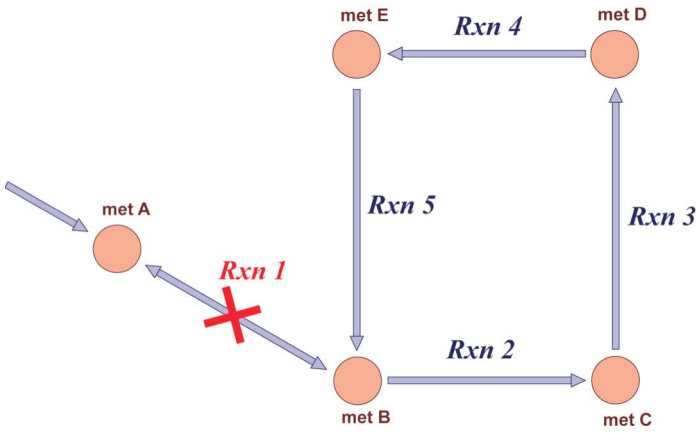

Figure 1.6 An example of how a cycle in a metabolic network can lead to a non-obvious gap. Reactions 2-5 form a cycle with reaction 1 as the only input or output. At steady state, reaction 1 is unable to carry flux without violating mass balance constraints because of this cycle. The four reactions in the cycle are free to carry any flux (as long as all reactions carry the same flux), but metabolite A is blocked even though it has both producing and consuming reactions. Algorithms such as GapFind are useful for identifying these types of gaps.

number of gaps with the smallest number of model modifications possible. These algorithms were used to make gap-filling predictions on the single-compartment *E. coli* metabolic model *i*JR904 [100] and the multi-compartment *S. cerevisiae* metabolic model *i*ND750 [77]. Although the hypotheses generated by GapFill were not experimentally verified, most were found to be in reasonable agreement with predicted free energy changes and database information.

SMILEY

With a constraint-based metabolic network model, FBA can be used to simulate the growth of microbial organisms on different substrates, including many different carbon and nitrogen sources. When an *in vivo* experiment shows that an organism can grow on a certain substrate but the model predicts that it cannot, the model is likely

missing one or more of the reactions required to consume the substrate. An algorithm called SMILEY was created to predict which reactions are likely missing from a network when the model predicts no growth but experiment indicates growth [11]. This algorithm was based on the OptStrain algorithm used in metabolic engineering [117]. SMILEY uses MILP to attempt to identify a flux distribution that leads to growth on the substrate of interest while minimizing the total number of reactions added from a universal database of reactions (based on the KEGG database). It thus predicts which reactions should be added to the model parsimoniously, adding the smallest number of reactions necessary to reconcile *in silico* and *in vivo* growth predictions. SMILEY does not directly target or attempt to fill gaps in a model, but the reactions it adds could end up filling gaps. SMILEY was used to predict reactions missing from the *E. coli* model *i*JR904, and made several predictions that have been experimentally verified [11].

GrowMatch

The newest algorithm for filling gaps in metabolic networks is called GrowMatch. This algorithm uses experimentally determined gene essentiality data to identify incorrect model predictions [12]. If the model predicts growth but an experiment shows that the organism cannot grow with a particular gene knockout (i.e., it is an essential gene), there is a growth/no growth inconsistency (GNG). When the model predicts no growth but an experiment shows that growth is possible, there is a no growth/growth inconsistency (NGG). GrowMatch attempts to correct these two types of disagreements differently. In the case of a GNG mutant, the model has some extra capabilities that are not realistic and should be removed or constrained. In cases of NGG inconsistency, the model is missing

one or more reactions that allow growth *in vivo*. GrowMatch then uses the three single compartment methods from GapFill (reversing reaction directions, adding reactions from a database, or adding transport reactions) to attempt to correct the model. Corrections to GNG or NGG mutants made by GrowMatch can be either global or conditional. Global corrections resolve at least one inconsistency while creating no new inconsistencies, and conditional corrections resolve one inconsistency while creating inconsistencies in other mutant strains. GrowMatch was used to analyze the *E. coli* model *i*AF1260 by comparing it to gene essentiality data from the Keio Collection [118]. 72 GNG mutants and 38 NGG mutants were identified, and GrowMatch suggested corrections for 56 GNG and 13 NGG mutants [12].

OMNI

Another algorithm related to SMILEY and GrowMatch that can also be used for gap-filling is called OMNI (Optimal Metabolic Network Identification) [14]. This is another MILP based algorithm that compares *in silico* predictions to experimental measurements in an attempt to improve a constraint-based model. In this case, OMNI uses measured metabolic flux data, which is obtained through ^{13}C labeling experiments [119]. OMNI compares experimentally measured fluxes to fluxes predicted by FBA, and then attempts to minimize the total difference between measured and predicted fluxes by adding or removing reactions while maintaining a predicted growth rate above a defined minimum. It uses a matrix **F** that contains fixed reactions which cannot be deleted and a matrix **D** that contains reactions that may be deleted. To improve a model by predicting missing reactions, a library of candidate reactions such as the one used by SMILEY can

be provided as **D** to the algorithm, which will then add reactions as needed to achieve a more realistic flux distribution. The OMNI algorithm has many potential uses other than gap-filling. Because it can remove reactions as well as add them, it can be used to remove reactions corresponding to poorly annotated genes if they do not match the experimental data. OMNI can also be used to identify alternative reaction mechanisms or bottleneck reactions in evolved strains or metabolically engineered strains.

1.5.3 Methods for predicting metabolic gene functions

PathoLogic Pathway Hole Filler

PathoLogic is a program for automatically constructing metabolic networks from annotated genomes. It uses EC numbers, Gene Ontology terms, or annotated gene names to map reactions to genes, and then assembles the reactions into pathways by comparing the reactions to the reference database MetaCyc and adding any missing reactions [69, 120]. After performing these steps, however, many of the reactions in the new pathways may be orphans. Pathologic includes a Pathway Hole Filler (PHFiller) program that attempts to identify the genes associated with these reactions [15]. The first step is to identify genes in other organisms that code for enzymes that catalyze an orphan reaction from the Swiss-Prot and PIR [121] databases. The sequences of these genes are then compared to the genome of the organism of interest using BLAST [122], and genes with similar sequences are identified. Because PHFiller needs sequence data from other organisms, it cannot predict the genes associated with global orphan reactions. After identifying candidate genes, the hole filler program then uses a simple Bayesian network to calculate the probability that each gene actually encodes an enzyme that catalyzes the

orphan activity. An extended PHFiller, called PHFiller-GC, which uses genome-context methods in addition to sequence comparisons to predict pathway genes, was released in 2007 [16]. PHFiller-GC uses proteins that are known to occur in the same complex, proteins from the same operon, or known transcription factor/regulatory target pairs to identify associated genes. Unlike the original PHFiller, PHFiller-GC is able to predict genes associated with global orphan reactions, but it requires that at least one enzyme in the target pathway have a known gene association. Both PHFiller and PHFiller-GC were tested on known *E. coli* reactions with homologous genes in other organisms, and both identified the true gene in the top ten predictions well (85.1% for PHFiller and 91% for PHFiller-GC). For reactions associated with genes with no known homologs, PHFiller-GC identified the true gene in the top ten predictions 58.9% of the time.

SEED: a subsystems approach

In response to the increasing number of fully sequenced microbial genomes, the Fellowship for Interpretation of Genomes (FIG) has launched the Project to Annotate 1000 Genomes. This large-scale project is using a subsystems approach to annotate these genomes [123]. Briefly, a subsystem is defined as a set of "functional roles" that act together to carry out a specific biological process such as a traditional biochemical pathway. Annotators focus on annotating one subsystem at a time in many different organisms using the SEED annotation environment [124], rather than the traditional approach of annotating one full genome at a time. The organization of genes in many different genomes into common subsystems can be beneficial in attempts to characterize genes and fill gaps. When one organism is missing a particular subsystem component

(reaction) that most other organisms with the same subsystem have, it is an indication of missing content. Bioinformatics methods can then be used to analyze the genes composing the subsystems in these other organisms to identify which genes in the target organism likely encode the missing reactions [18]. Analysis of chromosomal clustering has proven to be the most useful analysis method [17]. Genes that encode enzymes in the same pathways are often clustered close together on the genome, so genes and their orthologs that are found near each other in multiple genomes are likely to have related functions [125-127]. This approach followed by experimental verification has already led to several new discoveries [17, 128, 129].

ADOMETA

Since the first genomes were sequenced over a decade ago, numerous bioinformatics methods have been developed to attempt to identify the functions of genes. These methods include analysis of gene co-expression [130], phylogenetic profiles [131, 132], chromosomal clustering [126, 127], and protein fusions [133, 134]. More recently, a bioinformatics framework called ADOMETA (ADoption of Orphan METabolic Activities) that combines these types of information has been developed [19-21]. The genes in an organism with unknown functions (and genes with known functions as well) can be compared to the genes in the local metabolic network surrounding an orphan reaction using different types of functional association evidence. Genes that are close to each other in a metabolic network are more likely to have similar co-expression profiles, phylogenetic profiles, and so forth. In ADOMETA, the genes in an organism are compared to the surrounding genes using different methods, and the different association

scores are combined using the Adaboost method [135, 136]. The products of the highest scoring genes are the most likely to catalyze the orphan reaction.

The first type of functional association evidence used by ADOMETA is analysis of gene co-expression [130]. When two genes have similar expression patterns, they are likely to have related functions because they may be used in the same functional pathway. ADOMETA uses the Spearman's rank correlation coefficient to compare co-expression of genes using data from the Stanford Microarray Database [137]. Phylogenetic profiles compare the occurrence of genes in different organisms [131, 132]. A phylogenetic profile consists of a vector of 1's and 0's, with a 1 for every organism that contains an ortholog of a certain gene and a 0 for every organism that doesn't. Genes with related functions evolve together, so genes with the most similar phylogenetic profiles are therefore more likely to exist in the same pathways. ADOMETA uses a BLAST-based dataset of evolutionarily distinct genomes to determine phylogenetic profiles. Chromosomal clustering, which is used in the subsystems approach, can be a very powerful type of association evidence, especially in prokaryotic organisms [126, 127]. ADOMETA uses a dataset of orthologous gene clusters based on 108 genomes from KEGG. Finally, protein fusion data can be used [133, 134]. One gene in an organism may be homologous to two separate genes in other organisms, and it is likely that these two genes have similar functions. ADOMETA uses BLAST to identify genes containing homology to both candidate orphan-filling genes and network adjacent genes. ADOMETA was tested by comparing its predictions to the known enzyme encoding genes for reactions in both *E. coli* and *S. cerevisiae* using the *i*JR904 and *i*FF708 [138] models, respectively. 60% of the correct *E. coli* genes were in the top ten predictions for

their reactions, and 43% of the correct yeast genes were in the top ten predictions. ADOMETA is available as a web server at *http://vitkuplab.cu-genome.org/html/adometa/adometa.html* and can currently predict orphan associated genes for *E. coli*, *S. cerevisiae*, and *Bacillus subtilis*.

1.6 Metabolic engineering

One of the most useful practical applications of systems biology methods is in performing model-based metabolic engineering. The complete metabolic network models now available for many organisms allow the effects of genetic manipulations on these networks to be predicted. It is thus possible to design new strains that are capable of producing useful chemicals and can be used for industrial bioprocessing [36, 139]. *E. coli* is commonly used in metabolic engineering studies for several reasons. It has a high growth rate and many robust methods have been developed for genetic engineering and experimental analysis of this organism in the laboratory. Its metabolism is also very well studied, it naturally grows on a wide range of defined substrates, and it can naturally produce many different fermentation products.

1.6.1 Strategies for metabolic engineering

Many of the earliest microbial strains to be used for bioprocessing were not designed through rational engineering. Instead, random mutagenesis was used to create strains with different phenotypes. Cultures were exposed to UV radiation or chemicals to induce random genetic mutations, and the resulting strains were then screened for

desirable properties such as fermentation of non-native products. Once the tools of molecular biology became available for microbial organisms, it became possible to make beneficial genetic changes based on known biology, rather than relying on random mutations and screening.

Several common strategies were developed to engineer strains to produce useful compounds efficiently [140]. Methods to knockout, overexpress, or introduce heterologous genes could be utilized to rationally design strains. First, gene knockouts could be used to eliminate competing pathways in an organism's metabolic network. When an organism contains enzymes and pathways that allow it to produce multiple fermentation products, the elimination of one of these products could increase the yield of other products. The engineered organism would be unable to produce these other products, and would thus be forced to redirect metabolic flux to produce a targeted compound instead. Another common metabolic engineering strategy is to overexpress or upregulate the pathway for production of the desired target. If an organism naturally produces a desired compound, it may only do so at a very low rate or yield due to low expression of the pathway that produces this compound. By cloning the pathway genes on to a plasmid with promoters that allow for higher than usual expression, or by making changes to the cell's regulatory network, the number of enzymes available in the cell can be increased, increasing production. Of course, gene expression levels are not the only factor that determines flux through a particular pathway. Thermodynamic and kinetic factors affect reaction rates as well. These strategies can be overcome by introducing heterologous genes into an organism. By replacing native rate-limiting enzymes with others that have faster kinetics or require different cofactors, reaction rates and yields can

be increased. Heterologous genes can also add new capabilities to an organism, by adding a pathway for the production of an entirely new compound, for example. Less direct strategies can be used as well. Modifications can be made to increase the supply of precursors and cofactors, or to avoid accumulation of toxic intermediates.

With the advent of systems biology methods for the analysis of microbial organisms, systems metabolic engineering is now possible. Constraint-based modeling tools and metabolic network reconstructions can be used to determine the effects of targeted genetic changes on the entire metabolic network. Computational modeling allows for thousands of possible genetic alterations to be analyzed much faster than with *in vivo* screening [141]. Experimental tools such as genome-wide expression profiling and *in vivo* metabolomics can also be used to study the systemic effects of targeted genetic manipulations [36]. One of the most promising uses of systems biology in metabolic engineering is in the design of growth-coupled production strains.

1.6.2 Growth-coupled strain design

When the topology and other constraints of an organism's metabolic network force it to synthesize a certain product under specific conditions in order to grow as efficiently as possible, this product is coupled to growth. These growth-coupled products may be by-products of essential biomass producing pathways, or they may be the fermentation products that most efficiently balance intracellular redox levels. Growth-coupling is a useful feature in metabolic engineering strains because as mutations occur in an organism that increase production of the growth-coupled compound, growth rate will also increase. This allows for production strains to be selected for in adaptive

laboratory evolution (ALE) experiments. Through batch or continuous culture evolution experiments, strains with the highest growth rates can easily be selected for, and these strains will also have the highest production rates [142]. ALE can thus be used as a strain optimization tool. Growth-coupled metabolic engineering strains will also generally be more stable than non-growth-coupled production strains. In a non-growth-coupled strain, as random mutations occur to increase the cell growth rates, they will typically do so by decreasing production of the desired compound. A non-growth-coupled strain will tend to evolve to a less desirable phenotype over time, unless steps are taken to prevent evolution (**Figure 1.7 a**).

Constraint-based modeling of metabolic networks can be used to predict which compounds are growth-coupled in a particular strain. Simple FBA or flux variability analysis can be used to determine which compounds must be produced, which must not be produced, and which are optional at the maximum growth rate under specific conditions (**Figure 1.7 b**). When a metabolic network is changed through gene knockouts or knockins, the growth-coupled products can change [143].

1.6.3 Computational metabolic engineering strategies

A number of constraint-based computational tools have now been developed to design growth-coupled production strains. The first of these algorithms was OptKnock [38]. OptKnock is a bi-level optimization algorithm that seeks the set of reaction knockouts that will maximize flux through a particular target reaction (e.g., secretion of a desired compound) while also maximizing growth rate. This bi-level problem can be reformulated as an MILP problem, and has been implemented in the COBRA Toolbox

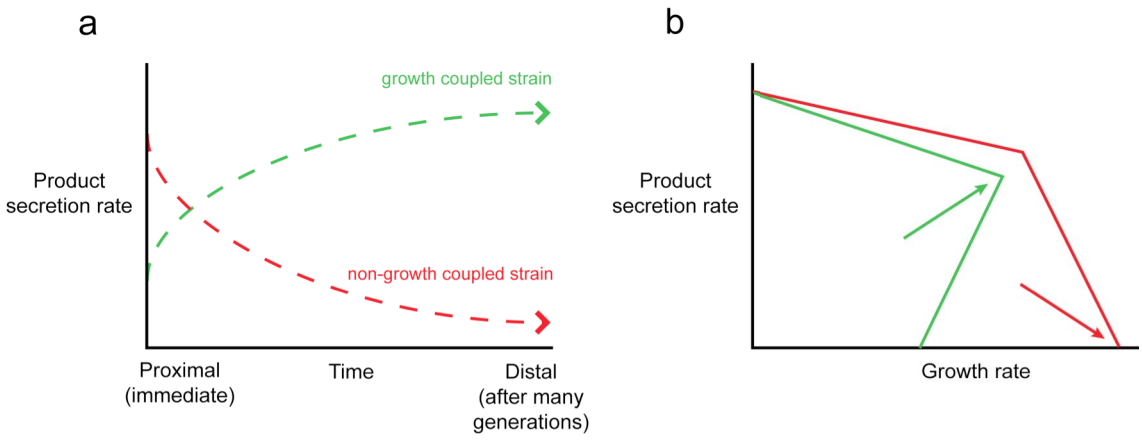

Figure 1.7 Growth-coupled and conventional strain designs. (**a**) The production rate of a growth-coupled strain will increase over time as a strain evolves to its highest possible growth rate. (**b**) FBA can be used to predict the production envelope of a strain (i.e., the solution space in two dimensions). Growth-coupled strains have a maximum growth rate with a significant non-zero production rate of a target compound (indicated by the arrow), while non-growth-coupled strains have a production rate of zero at the maximum growth rate.

[144]. A slightly modified version of OptKnock was developed to avoid identification of solutions with non-unique target secretion rates [143]. It is possible that OptKnock will identify a set of knockouts that leads to a production rate with a range of values including zero. Such a strain would not actually be growth-coupled. The modified version of OptKnock uses a tilted objective function, which includes both maximization of biomass production and minimization of product yield multiplied by a small constant. This ensures the OptKnock algorithm will identify the lowest possible production rate and try to maximize it. An independently developed method called RobustKnock was also developed to avoid these zero-production solutions [145].

Another algorithm for designing growth-coupled strains is OptGene [39]. Like OptKnock, OptGene seeks the set of gene knockouts that leads to the highest possible growth rate and target compound production rate. Unlike OptKnock, OptGene is a

genetic algorithm that simulates sets of random gene knockouts that are gradually changed to increase a defined objective function. As it uses random sets of knockouts, OptGene is not guaranteed to find the globally optimal solution. The main advantage of OptGene is that it can use any mathematical combination of growth rate and reaction fluxes as its objective, including nonlinear functions. Thus, features such as strength of growth-coupling and penalties for unnecessary knockouts can be included.

Many other constraint-based strain design algorithms are also now available. OptStrain can predict the optimal heterologous reactions to add to a network to produce non-native products [117]. An algorithm called OptReg predicts which genes should be up- or downregulated in a network to achieve optimal production [146]. OptFlux is an updated version of OptGene that uses both a genetic algorithm and simulated annealing [147]. A local search algorithm called Genetic Design through Local Search (GDLS) was designed to find knockout designs faster than the MILP based OptKnock [148]. Flux Scanning based on Enforced Objective Flux (FSEOF) identifies genes that should be upregulated to increase target compound production [149]. OptORF is an algorithm closely related to OptKnock that identifies the set of gene knockouts rather than reaction knockouts, while also predicting upregulation targets [150]. The newest algorithm is called Enhancing Metabolism with Iterative Linear Optimization (EMILiO), and was designed to be the fastest available strain design algorithm [151].

Acknowledgements

Chapter 1 is, in part, adapted from a chapter that appeared in EcoSal – *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (www.ecosal.org), Chapter 10.2.1, ASM Press, Washington D.C., February 18, 2010. The dissertation author was the primary author of this chapter which was coauthored by Ronan M.T. Fleming and Bernhard Ø. Palsson.

Chapter 1 is also, in part, adapted from a paper that appeared in Nature Biotechnology, Volume 28, Number 3, Pages 245-8, March 2010. The dissertation author was the primary author of this paper which was coauthored by Ines Thiele and Bernhard Ø. Palsson.

Chapter 1 is also, in part, adapted from a paper that appeared in Biotechnology and Bioengineering, Volume 107, Number 3, Pages 403-12, October 15, 2010. The dissertation author was the primary author of this paper which was coauthored by Bernhard Ø. Palsson.

We would like to thank Byung-Kwan Cho, Adam Feist, Nathan Lewis, and Karsten Zengler for their helpful comments and insights.

References

1. Palsson, B.Ø., *Systems biology: properties of reconstructed networks.* 2006, New York: Cambridge University Press.

2. Hatzimanikatis, V., et al., *Exploring the diversity of complex metabolic networks.* Bioinformatics, 2005. **21**(8): p. 1603-9.

3. Hua, Q., et al., *Responses of the central metabolism in Escherichia coli to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts.* J Bacteriol, 2003. **185**(24): p. 7053-67.

4. Feist, A.M. and B.Ø. Palsson, *The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.* Nat Biotech, 2008. **26**(6): p. 659-667.

5. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.

6. Varma, A. and B.Ø. Palsson, *Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors.* Journal of Theoretical Biology, 1993. **165**(4): p. 477-502.

7. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.

8. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**(121).

9. Nogales, J., B.Ø. Palsson, and I. Thiele, *A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory.* BMC Syst Biol, 2008. **2**: p. 79.

10. Resendis-Antonio, O., et al., *Metabolic reconstruction and modeling of nitrogen fixation in Rhizobium etli.* PLoS Comput Biol, 2007. **3**(10): p. 1887-95.

11. Reed, J.L., et al., *Systems approach to refining genome annotation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17480-4.

12. Kumar, V.S. and C.D. Maranas, *GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.* PLoS Comput Biol, 2009. **5**(3): p. e1000308.

13. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.

14. Herrgård, M.J., S.S. Fong, and B.Ø. Palsson, *Identification of genome-scale metabolic network models using experimentally measured flux profiles.* PLoS Comput Biol, 2006. **2**(7): p. e72.

15. Green, M.L. and P.D. Karp, *A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.* BMC Bioinformatics, 2004. **5**: p. 76.

16. Green, M.L. and P.D. Karp, *Using genome-context data to identify specific types of functional associations in pathway/genome databases.* Bioinformatics, 2007. **23**(13): p. i205-11.

17. Osterman, A., *A hidden metabolic pathway exposed.* Proc Natl Acad Sci U S A, 2006. **103**(15): p. 5637-8.

18. Osterman, A. and R. Overbeek, *Missing genes in metabolic pathways: a comparative genomics approach.* Curr Opin Chem Biol, 2003. **7**(2): p. 238-51.

19. Chen, L.F. and D. Vitkup, *Predicting genes for orphan metabolic activities using phylogenetic profiles.* Genome Biology, 2006. **7**(2).

20. Kharchenko, P., et al., *Identifying metabolic enzymes with multiple types of association evidence.* BMC Bioinformatics, 2006. **7**(177).

21. Kharchenko, P., D. Vitkup, and G.M. Church, *Filling gaps in a metabolic network using expression information.* Bioinformatics, 2004. **20 Suppl 1**: p. I178-I185.

22. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends in Biotechnology, 2003. **21**(4): p. 162-169.

23. Pal, C., B. Papp, and M.J. Lercher, *Horizontal gene transfer depends on gene content of the host.* Bioinformatics, 2005. **21 Suppl 2**: p. ii222-ii223.

24. Pal, C., B. Papp, and M.J. Lercher, *Adaptive evolution of bacterial metabolic networks by horizontal gene transfer.* Nat Genet, 2005. **37**(12): p. 1372-5.

25. Pal, C., et al., *Chance and necessity in the evolution of minimal metabolic networks.* Nature, 2006. **440**(7084): p. 667-70.

26. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.

27. Burgard, A.P., et al., *Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions.* Genome. Res., 2004. **14**(2): p. 301-12.

28. Barrett, C.L., et al., *The global transcriptional regulatory network for metabolism in Escherichia coli attains few dominant functional states.* Proc Natl Acad Sci U S A, 2005. **102**(52): p. 19103-19108.

29. Samal, A. and S. Jain, *The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response.* BMC Syst Biol, 2008. **2**(1): p. 21.

30. Almaas, E., et al., *Global organization of metabolic fluxes in the bacterium Escherichia coli.* Nature, 2004. **427**(6977): p. 839-843.

31. Edwards, J.S., R.U. Ibarra, and B.Ø. Palsson, *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nat Biotechnol, 2001. **19**(2): p. 125-130.

32. Segre, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15112-7.

33. Shlomi, T., O. Berkman, and E. Ruppin, *Regulatory on/off minimization of metabolic flux changes after genetic perturbations.* Proc Natl Acad Sci U S A, 2005. **102**(21): p. 7695-700.

34. Ibarra, R.U., J.S. Edwards, and B.Ø. Palsson, *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, 2002. **420**(6912): p. 186-9.

35. Joyce, A.R., et al., *Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli.* J Bacteriol, 2006. **188**(23): p. 8259-8271.

36. Park, J.H., et al., *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**(19): p. 7797-802.

37. Fong, S.S., et al., *In silico design and adaptive evolution of Escherichia coli for production of lactic acid.* Biotechnol Bioeng, 2005. **91**(5): p. 643-8.

38. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

39. Patil, K.R., et al., *Evolutionary programming as a platform for in silico metabolic engineering.* BMC Bioinformatics, 2005. **6**: p. 308.

40. Famili, I., R. Mahadevan, and B.Ø. Palsson, *k-Cone Analysis: Determining All Candidate Values for Kinetic Parameters on a Network Scale.* Biophys J, 2005. **88**(3): p. 1616-25.

41. Segre, D., et al., *From annotated genomes to metabolic flux models and kinetic parameter fitting.* Omics, 2003. **7**(3): p. 301-16.

42. Rizzi, M., et al., *In vivo analysis of metabolic dynamics in Saccharomyces cerevisiae .2. Mathematical model.* Biotechnology and Bioengineering, 1997. **55**(4): p. 592-608.

43. Teusink, B., et al., *Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry.* Eur J Biochem, 2000. **267**(17): p. 5313-29.

44. Vaseghi, S., et al., *In vivo Dynamics of the pentose phosphate pathway in Saccharomyces cerevisiae.* Metabolic Engineering, 1999. **1**: p. 128-140.

45. Wright, B.E. and P.J. Kelly, *Kinetic models of metabolism in intact cells, tissues, and organisms.* Curr Top Cell Regul, 1981. **19**: p. 103-58.

46. Jamshidi, N. and B.Ø. Palsson, *Formulating genome-scale kinetic models in the post-genome era.* Mol Syst Biol, 2008. **4**: p. 171.

47. Covert, M.W., et al., *Metabolic modeling of microbial strains in silico.* Trends Biochem. Sci., 2001. **26**: p. 179-186.

48. Edwards, J.S., M. Covert, and B.Ø. Palsson, *Metabolic modeling of microbes: the flux-balance approach.* Environmental Microbiology, 2002. **4**(3): p. 133-40.

49. Price, N.D., J.L. Reed, and B.Ø. Palsson, *Genome-scale models of microbial cells: evaluating the consequences of constraints.* Nat Rev Microbiol, 2004. **2**(11): p. 886-897.

50. Varma, A. and B.Ø. Palsson, *Metabolic Flux Balancing: Basic concepts, Scientific and Practical Use.* Nat Biotechnol, 1994. **12**: p. 994-998.

51. Oberhardt, M.A., B.Ø. Palsson, and J.A. Papin, *Applications of genome-scale metabolic reconstructions.* Mol Syst Biol, 2009. **5**: p. 320.

52. Thiele, I. and B.Ø. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.

53. Durot, M., P.Y. Bourguignon, and V. Schachter, *Genome-scale models of bacterial metabolism: reconstruction and applications.* FEMS Microbiol Rev, 2009. **33**(1): p. 164-90.

54. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox.* Nat. Protocols, 2007. **2**(3): p. 727-738.

55. Karp, P.D., et al., *Multidimensional annotation of the Escherichia coli K-12 genome.* Nucleic Acids Res, 2007.

56. Stein, L., *Genome annotation: from sequence to biology.* Nat Rev Genet, 2001. **2**(7): p. 493-503.

57. Christie, K.R., et al., *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**: p. D311-4.

58. Guldener, U., et al., *CYGD: the Comprehensive Yeast Genome Database.* Nucleic Acids Res, 2005. **33**(Database issue): p. D364-8.

59. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2005. **33**(Database issue): p. D54-8.

60. Peterson, J.D., et al., *The Comprehensive Microbial Resource.* Nucleic Acids Res, 2001. **29**(1): p. 123-5.

61. Stoesser, G., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 1999. **27**(1): p. 18-24.

62. Markowitz, V.M., et al., *The integrated microbial genomes (IMG) system.* Nucleic Acids Res, 2006. **34**(Database issue): p. D344-8.

63. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

64. Chang, A., et al., *BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D588-92.

65. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.

66. Krieger, C.J., et al., *MetaCyc: a multiorganism database of metabolic pathways and enzymes*, 2004. p. D438-442.

67. DeJongh, M., et al., *Toward the automated generation of genome-scale metabolic networks in the SEED.* BMC Bioinformatics, 2007. **8**: p. 139.

68. Ren, Q., K. Chen, and I.T. Paulsen, *TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels.* Nucleic Acids Res, 2007. **35**(Database issue): p. D274-9.

69. Paley, S.M. and P.D. Karp, *Evaluation of computational metabolic-pathway predictions for Helicobacter pylori.* Bioinformatics, 2002. **18**(5): p. 715-24.

70. Claudel-Renard, C., et al., *Enzyme-specific profiles for genome annotation: PRIAM.* Nucleic Acids Res, 2003. **31**(22): p. 6633-9.

71. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

72. Reed, J.L., et al., *Towards multidimensional genome annotation.* Nat Rev Genet, 2006. **7**(2): p. 130-41.

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

74. Edwards, J.S. and B.Ø. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities.* Proc Natl Acad Sci U S A., 2000. **97**(10): p. 5528-5533.

75. Varma, A. and B.Ø. Palsson, *Metabolic capabilities of Escherichia coli: II. Optimal growth patterns.* Journal of Theoretical Biology, 1993. **165**(4): p. 503-522.

76. Varma, A. and B.Ø. Palsson, *Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110.* Applied and Environmental Microbiology, 1994. **60**(10): p. 3724-3731.

77. Duarte, N.C., M.J. Herrgård, and B.Ø. Palsson, *Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model.* Genome Res, 2004. **14**(7): p. 1298-309.

78. Breitling, R., D. Vitkup, and M.P. Barrett, *New surveyor tools for charting microbial metabolic maps.* Nat Rev Microbiol, 2008. **6**(2): p. 156-61.

79. Bonneau, R., et al., *A predictive model for transcriptional control of physiology in a free living cell.* Cell, 2007. **131**(7): p. 1354-65.

80. Herrgård, M.J., et al., *Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae.* Genome Research, 2006. **16**(5): p. 627-635.

81. Workman, C.T., et al., *A systems approach to mapping DNA damage response pathways.* Science, 2006. **312**(5776): p. 1054-9.

82. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

83. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.

84. Hu, Z., P.J. Killion, and V.R. Iyer, *Genetic reconstruction of a functional transcriptional regulatory network.* Nat Genet, 2007. **39**(5): p. 683-7.

85. Segal, E., et al., *Module networks: identifying regulatory modules and their condition- specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.

86. Cho, B.K., et al., *Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19462-7.

87. Cho, B.K., et al., *Genome-wide Analysis of Fis Binding in Escherichia coli Indicates a Causative Role for A-/AT-tracts.* Genome Res., 2008. **18**(6): p. 900-910.

88. Cho, B.K., E.M. Knight, and B.Ø. Palsson, *Genomewide identification of protein binding locations using chromatin immunoprecipitation coupled with microarray.* Methods Mol Biol, 2008. **439**: p. 131-45.

89. Ideker, T.E., V. Thorsson, and R.M. Karp, *Discovery of regulatory interactions through perturbation: inference and experimental design.* Pacific Symposium on Biocomputing, 2000. **292**(5518): p. 305-16.

90. Cho, B.K., E.M. Knight, and B.Ø. Palsson, *Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA.* Microbiology, 2006. **152**(Pt 8): p. 2207-19.

91. Grainger, D.C., et al., *Transcription factor distribution in Escherichia coli: studies with FNR protein.* Nucleic Acids Res, 2007. **35**(1): p. 269-78.

92. Shimada, T., et al., *The Escherichia coli RutR transcription factor binds at targets within genes as well as intergenic regions.* Nucleic Acids Res, 2008. **36**(12): p. 3950-5.

93. Covert, M.W., C.H. Schilling, and B.Ø. Palsson, *Regulation of gene expression in flux balance models of metabolism.* Journal of Theoretical Biology, 2001. **213**(1): p. 73-88.

94. Gianchandani, E.P., et al., *Matrix Formalism to Describe Functional States of Transcriptional Regulatory Systems.* PLoS Comput Biol., 2006. **2**(8): p. e101.

95. Gianchandani, E.P., et al., *Functional States of the genome-scale Escherichia coli transcriptional regulatory system.* PLoS Comput Biol, 2009. **5**(6): p. e1000403.

96. Thiele, I., et al., *Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization.* PLoS Comput Biol, 2009. **5**(3): p. e1000312.

97. Majewski, R.A. and M.M. Domach, *Simple constrained optimization view of acetate overflow in E. coli.* Biotechnology and Bioengineering, 1990. **35**: p. 732-738.

98. Varma, A., B.W. Boesch, and B.Ø. Palsson, *Biochemical production capabilities of Escherichia coli.* Biotechnology and Bioengineering, 1993. **42**(1): p. 59-73.

99. Pramanik, J. and J.D. Keasling, *Stoichiometric model of Escherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements.* Biotechnology and Bioengineering, 1997. **56**(4): p. 398-421.

100. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biology, 2003. **4**(9): p. R54.1-R54.12.

101. Jankowski, M.D., et al., *Group contribution method for thermodynamic analysis of complex metabolic networks.* Biophys J, 2008. **95**(3): p. 1487-99.

102. Mizuno, T., *Compilation of all genes encoding two-component phosphotransfer signal transducers in the genome of Escherichia coli.* DNA Res, 1997. **4**(2): p. 161-8.

103. Laub, M.T. and M. Goulian, *Specificity in two-component signal transduction pathways.* Annu Rev Genet, 2007. **41**: p. 121-45.

104. Bijlsma, J.J. and E.A. Groisman, *Making informed decisions: regulatory interactions between two-component systems.* Trends Microbiol, 2003. **11**(8): p. 359-66.

105. Palsson, B.Ø., *Metabolic systems biology.* FEBS Lett, 2009. **583**(24): p. 3900-4.

106. Liu, D. and P.R. Reeves, *Escherichia coli K12 regains its O antigen.* Microbiology, 1994. **140 (Pt 1)**: p. 49-57.

107. Rubires, X., et al., *A gene (wbbL) from Serratia marcescens N28b (O4) complements the rfb-50 mutation of Escherichia coli K-12 derivatives.* J Bacteriol, 1997. **179**(23): p. 7581-6.

108. Keseler, I.M., et al., *EcoCyc: a comprehensive view of Escherichia coli biology.* Nucleic Acids Res, 2009. **37**(Database issue): p. D464-70.

109. Karp, P.D., *Call for an enzyme genomics initiative.* Genome Biol, 2004. **5**(8): p. 401.

110. Lespinet, O. and B. Labedan, *Orphan enzymes could be an unexplored reservoir of new drug targets.* Drug Discov Today, 2006. **11**(7-8): p. 300-5.

111. Pouliot, Y. and P.D. Karp, *A survey of orphan enzyme activities.* BMC Bioinformatics, 2007. **8**: p. 244.

112. Chen, L. and D. Vitkup, *Distribution of orphan metabolic activities.* Trends Biotechnol, 2007. **25**(8): p. 343-8.

113. Lespinet, O. and B. Labedan, *ORENZA: a web resource for studying ORphan ENZyme activities.* BMC Bioinformatics, 2006. **7**: p. 436.

114. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.

115. Orth, J.D., et al., *A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011.* Mol Syst Biol, 2011. **7**: p. 535.

116. Caspi, R., et al., *The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases.* Nucleic Acids Res, 2008. **36**(Database issue): p. D623-31.

117. Pharkya, P., A.P. Burgard, and C.D. Maranas, *OptStrain: a computational framework for redesign of microbial production systems.* Genome Res, 2004. **14**(11): p. 2367-76.

118. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol Syst Biol, 2006. **2**: p. 2006.0008.

119. Fong, S.S., et al., *Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes.* J Biol Chem, 2006. **281**(12): p. 8024-33.

120. Karp, P.D., S. Paley, and P. Romero, *The Pathway Tools software.* Bioinformatics, 2002. **18 Suppl 1**: p. S225-32.

121. Wu, C.H., et al., *The Protein Information Resource.* Nucleic Acids Res, 2003. **31**(1): p. 345-7.

122. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

123. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.

124. Overbeek, R., T. Disz, and R. Stevens, *The SEED: a peer-to-peer environment for genome annotation.* Commun ACM, 2004. **47**(11): p. 46-51.

125. Dandekar, T., et al., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends Biochem Sci, 1998. **23**(9): p. 324-8.

126. Lee, J.M. and E.L. Sonnhammer, *Genomic gene clustering analysis of pathways in eukaryotes.* Genome Res, 2003. **13**(5): p. 875-82.

127. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.

128. Loh, K.D., et al., *A previously undescribed pathway for pyrimidine catabolism.* Proc Natl Acad Sci U S A, 2006. **103**(13): p. 5114-9.

129. Rodionov, D.A., et 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**(11): p. 4279-84.

130. Wu, L.F., et al., *Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters.* Nat Genet, 2002. **31**(3): p. 255-65.

131. Huynen, M.A. and P. Bork, *Measuring genome evolution.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 5849-56.

132. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.

133. Enright, A.J., et al., *Protein interaction maps for complete genomes based on gene fusion events.* Nature, 1999. **402**(6757): p. 86-90.

134. Yanai, I., A. Derti, and C. DeLisi, *Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes.* Proc Natl Acad Sci U S A, 2001. **98**(14): p. 7940-5.

135. Freund, Y. and R. Schapire, *A decision-theoretic generalization of online learning and an application to boosting.* J Computer and System Sci, 1997. **55**(1): p. 119-139.

136. Schapire, R., *The boosting approach to machine learning: An overview.*, in *MSRI Workshop on Nonlinear Estimation and Classification*2002.

137. Sherlock, G., et al., *The Stanford Microarray Database.* Nucleic Acids Res, 2001. **29**(1): p. 152-5.

138. Forster, J., et al., *Genome-Scale Reconstruction of the Saccharomyces cerevisiae Metabolic Network.* Genome Research, 2003. **13**(2): p. 244-53.

139. Lee, K.H., et al., *Systems metabolic engineering of Escherichia coli for L-threonine production.* Mol Syst Biol, 2007. **3**: p. 149.

140. Bailey, J.E., *Toward a science of metabolic engineering.* Science, 1991. **252**(5013): p. 1668-75.

141. Kim, T.Y., et al., *Strategies for systems-level metabolic engineering.* Biotechnol J, 2008. **3**(5): p. 612-23.

142. Conrad, T.M., N.E. Lewis, and B.Ø. Palsson, *Microbial laboratory evolution in the era of genome-scale science.* Mol Syst Biol, 2011. **7**: p. 509.

143. Feist, A.M., et al., *Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli.* Metab Eng, 2010. **12**(3): p. 173-86.

144. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.* Nature protocols, 2011. **6**(9): p. 1290-307.

145. Tepper, N. and T. Shlomi, *Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways.* Bioinformatics, 2010. **26**(4): p. 536-43.

146. Pharkya, P. and C.D. Maranas, *An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems.* Metab Eng, 2006. **8**(1): p. 1-13.

147. Rocha, M., et al., *Natural computation meta-heuristics for the in silico optimization of microbial strains.* BMC Bioinformatics, 2008. **9**: p. 499.

148. Lun, D.S., et al., *Large-scale identification of genetic design strategies using local search.* Mol Syst Biol, 2009. **5**: p. 296.

149. Choi, H.S., et al., *In silico identification of gene amplification targets for improvement of lycopene production.* Appl Environ Microbiol, 2010. **76**(10): p. 3097-105.

150. Kim, J. and J.L. Reed, *OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.* BMC Syst Biol, 2010. **4**: p. 53.

151. Yang, L., W.R. Cluett, and R. Mahadevan, *EMILiO: a fast algorithm for genome-scale strain design.* Metab Eng, 2011. **13**(3): p. 272-81.

Chapter 2: The core *Escherichia coli* metabolic network reconstruction

Most metabolic network reconstructions used in current research are genome-scale, that is, they contain the functions of all known metabolic genes in the entire genome of an organism. Due to the great size and complexity of these models, interpretation of constraint-based computations can be difficult. The core *Escherichia coli* metabolic reconstruction is a small-scale reconstruction, containing only 95 reactions and 72 metabolites. The core reconstruction also contains a core regulatory network reconstruction that can be modeled in a Boolean format. This chapter gives a complete biochemical description of this reconstruction and also presents a tutorial that uses the COBRA Toolbox to demonstrate several constraint-based analysis procedures.

2.1 Construction and content of the core *E. coli* metabolic network

The core *Escherichia coli* model is a small-scale model that can be used for educational purposes. It is intended to be used by senior undergraduate and first year graduate students learning about constraint-based modeling and systems biology. This model has enough reactions and pathways to enable interesting and insightful calculations, but it is also simple enough that the results of such calculations can be easily understood. This model is also useful for testing and evaluating new constraint-based

analysis methods, since its small scope makes troubleshooting and interpretation of results easier.

The reactions and pathways in the core model were chosen to represent the most well known and widely studied metabolic pathways of *E. coli*. These pathways are often the subjects of textbook chapters and should be familiar to most readers with a basic biochemistry background. As much as possible, the reactions and gene-protein-reaction associations (GPRs) were taken directly from the *i*AF1260 genome-scale *E. coli* reconstruction [1]. Some pathways, such as the electron transport system, were greatly simplified in order to limit the scope of the model and ensure that every reaction is understandable. This model contains a total of 72 metabolites and 95 reactions (**Figure 2.1**). There are 20 extracellular metabolites and 52 intracellular metabolites, with a total of 54 unique metabolites (most extracellular metabolites are just extracellular versions of intracellular metabolites). There are 20 exchange reactions, one for each extracellular metabolite. The model also has 25 transport reactions, 49 metabolic reactions, and one biomass reaction. In this chapter, reaction abbreviations are listed uppercase and italicized (e.g., *PFK* is phosphofructokinase), and metabolite abbreviations are lowercase and may be either cytosolic ([c]) or extracellular ([e]).

2.1.1 Glycolysis

One of the most important and widely studied metabolic pathways is glycolysis, a series of ten chemical reactions that convert one 6-carbon glucose molecule into two 3-carbon pyruvate molecules. In these reactions, a net two molecules of adenosine triphosphate (ATP) are produced by substrate level phosphorylation, and two molecules

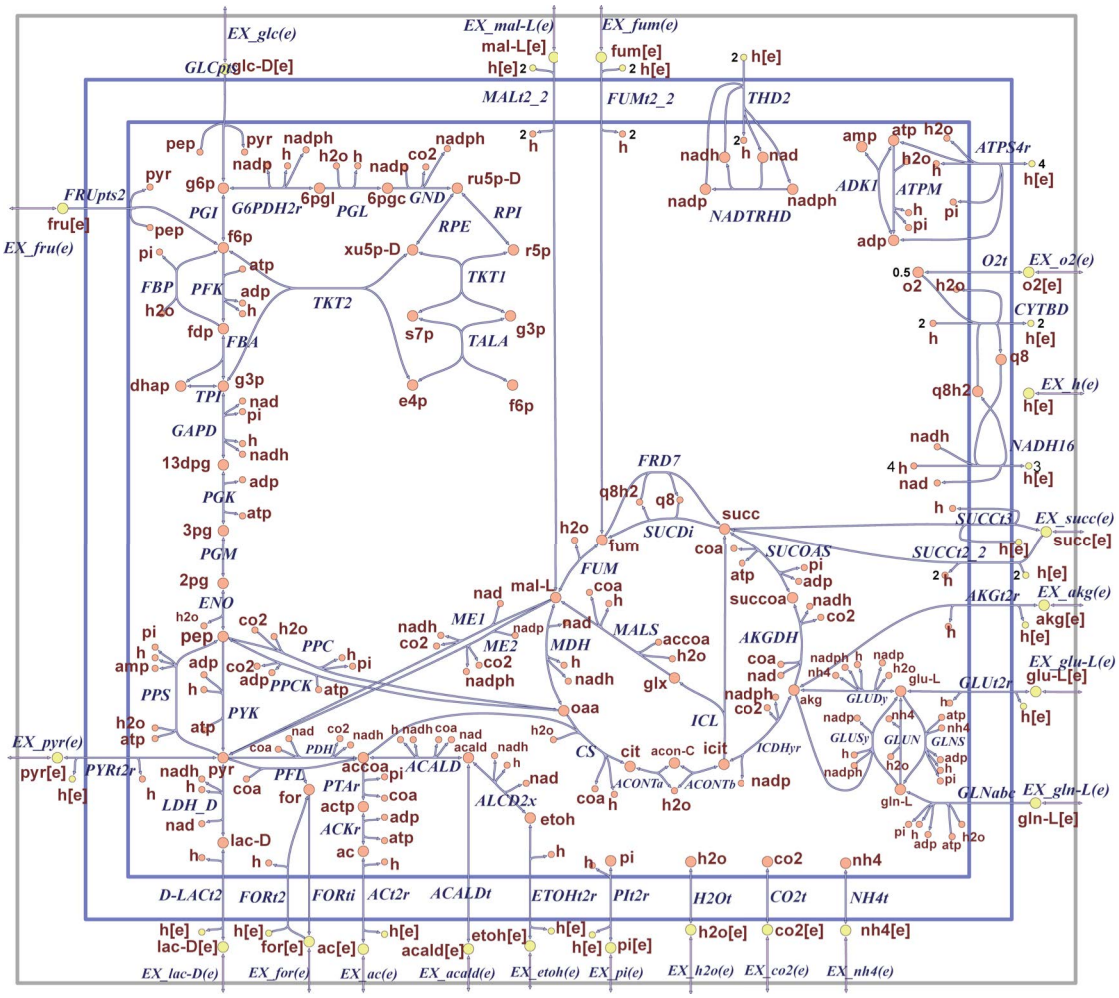

Figure 2.1 The core *E. coli* metabolic reconstruction includes 95 reactions and 72 metabolites in two compartments: cytosol and extracellular.

of reduced nicotinamide adenine dinucleotide (NADH) are also produced. In addition to producing these energy and redox carriers, glycolysis also produces several compounds that are precursors for *E. coli* biomass. Glycolysis begins at the phosphoenolpyruvate:pyruvate phosphotransferase protein complex, which actively translocates hexoses across the inner cytoplasmic membrane [2]. Certain proteins of the phosphoenolpyruvate:pyruvate phosphotransferase complex are carbohydrate specific, but in each case, transport is driven by transfer of the phosphate group of

phosphoenolpyruvate to the carbohydrate. In the reaction D-glucose transport via PEP:Pyr PTS, *GLCpts*, phosphoenolpyruvate donates the phosphate group to glucose to form D-glucose-6-phosphate and the dephosphorylated remainder of phosphoenolpyruvate is pyruvate [3]. The same general procedure generates D-fructose-6-phosphate from fructose in the reaction fructose transport via PEP:Pyr PTS (f6p generating), *FRUpts2* [4]. The interconversion of D-glucose-6-phosphate and D-fructose-6-phosphate is catalyzed by glucose-6-phosphate isomerase, *PGI* [5]. Phosphofructokinase, *PFK*, catalyzes the transfer of a phosphate group from ATP to D-fructose-6-phosphate to form D-fructose-1,6-bisphosphate and adenosine diphosphate (ADP) [6, 7]. This reaction is effectively thermodynamically irreversible. However, in the reverse direction, the dephosphorylation of D-fructose-1,6-bisphosphate to form D-fructose-6-phosphate is catalyzed by fructose-bisphosphatase, *FBP* [8].

Fructose-bisphosphate aldolase, *FBA*, splits the 6-carbon D-fructose-1,6-bisphosphate into two 3-carbon molecules, dihydroxyacetone-phosphate and glyceraldehyde-3-phosphate [8-11]. Triose-phosphate isomerase, *TPI*, rapidly and reversibly structurally rearranges dihydroxyacetone-phosphate to glyceraldehyde-3-phosphate [12]. A linear sequence of four reversible reactions catalyzed by glyceraldehyde-3-phosphate dehydrogenase, *GAPD* [13], phosphoglycerate kinase, *PGK* [14], phosphoglycerate mutase, *PGM* [15], and enolase, *ENO* [16], converts glyceraldehyde-3-phosphate to phosphoenolpyruvate. This sequence of reactions also reduces one NAD^+ to form NADH in the glyceraldehyde-3-phosphate dehydrogenase reaction and yields one high-energy currency metabolite, ATP, in the phosphoglycerate kinase reaction. In the final step of glycolysis, pyruvate kinase, *PYK*, catalyzes the

transfer of the phosphate group of phosphoenolpyruvate to ADP resulting in the production of pyruvate and ATP [17, 18]. Pyruvate is an important precursor involved in many pathways. It can be converted into acetyl-CoA, which provides carbon for the tricarboxylic acid cycle, and can also be converted into lactate as part of fermentation.

2.1.2 Pentose phosphate pathway

The primary function of the pentose phosphate shunt is to provide the 5-carbon and 4-carbon biosynthetic precursors alpha-D-ribose-5-phosphate and D-erythrose-4-phosphate. Alpha-D-ribose-5-phosphate and D-erythrose-4-phosphate can be produced by either of two parallel pathways, the decarboxylating oxidative pathway or the non-oxidative pathway. In anaerobic conditions there is a greater flux through the non-oxidative pathway than the oxidative pathway [19, 20].

The decarboxylating oxidative pathway is effectively irreversible. It consumes D-glucose-6-phosphate and after three reactions catalyzed by glucose-6-phosphate dehydrogenase, *G6PDH2r* [21], 6-phosphogluconolactonase, *PGL* [22], and phosphogluconate dehydrogenase, *GND* [23], produces D-ribulose-5-phosphate. The first and third reactions in this pathway each reduce one nicotinamide adenine dinucleotide phosphate ($NADP^+$) to NADPH. D-ribulose-5-phosphate is then reversibly structurally rearranged into alpha-D-ribose-5-phosphate in the reaction ribose-5-phosphate isomerase, *RPI* [24]. The oxidative branch of the pentose phosphate shunt is important for production of reducing power in the form of NADPH. However, the pentose phosphate shunt is not the only source of NADPH [25]. The reactions catalyzed by NAD(P)

transhydrogenase, *THD2*, isocitrate dehydrogenase (NADP), *ICDHyr*, and malic enzyme (NADP), *ME2*, can also supply *E. coli* with NADPH.

The non-oxidative reversible rearrangement of the glycolytic sugar monophosphates to the pentose phosphate shunt sugar monophosphates is a simple mechanism for creating these precursors [26], but does not contribute to reducing power. This rearrangement requires three steps. First, transketolase, *TKT2* [27, 28], catalyzes the conversion of a 6-carbon compound, D-fructose-6-phosphate, plus a 3-carbon compound, glyceraldehyde-3-phosphate, into a 5-carbon compound, D-xylulose-5-phosphate, plus a 4-carbon precursor, D-erythrose-4-phosphate. Then transaldolase, *TALA*, catalyzes the conversion of this 4-carbon compound plus another 6-carbon D-fructose-6-phosphate, into a 7-carbon compound, sedoheptulose-7-phosphate, plus another molecule of 3-carbon glyceraldehyde-3-phosphate. Next, the multifunctional enzyme transketolase, *TKT1*, catalyzes the conversion of this 7-carbon plus this 3-carbon into two different 5-carbon compounds, D-xylulose-5-phosphate and the precursor alpha-D-ribose-5-phosphate. In addition to ribose-5-phosphate isomerase, ribulose-5-phosphate 3-epimerase, *RPE* [29], provides another reversible catalytic link between the oxidative and non-oxidative branches since it interconverts D-xylulose-5-phosphate and D-ribulose-5-phosphate.

2.1.3 Tricarboxylic acid cycle

The tricarboxylic acid (TCA) cycle is a well-studied pathway with a variety of functions depending on the environment. During aerobic growth on 6-carbon sugars such as glucose, the TCA cycle functions to create the precursors oxaloacetate, 2-oxoglutarate

(also commonly called α-ketoglutarate), and succinyl-CoA. The aerobic production of biomass precursors is carried out primarily by the oxidative arm of the TCA cycle, from oxaloacetate to 2-oxoglutarate. This is the counter-clockwise, lower part of the cycle in **Figure 2.1**. The full TCA cycle, continuing counter-clockwise, can totally oxidize acetyl-CoA. Cycle intermediates are still required as biosynthetic precursors, so flux from anaplerotic pathways is required to maintain the pool of dicarboxylic intermediates. Under anaerobic conditions, the TCA cycle functions not as a cycle, but as two separate pathways. The oxidative pathway, the counterclockwise lower part of the cycle, forms the precursor 2-oxoglutarate. The reductive pathway, the clockwise upper part of the cycle, forms the precursor succinyl-CoA. *E. coli* can grow in an environment where the only carbon substrate is one of the TCA cycle intermediates. This is enabled by proton symport transport reactions that translocate either 2-oxoglutarate, succinate, fumarate or L-malate into the cell.

Pyruvate dehydrogenase, *PDH*, catalyzes the synthesis of acetyl-CoA from pyruvate and coenzyme A with concomitant reduction of NAD^+ to NADH [30]. Citrate synthase, *CS* [31], catalyzes the condensation reaction of the 2-carbon acetate residue from acetyl-CoA and a molecule of 4-carbon oxaloacetate to form the 6-carbon compound citrate, with the release of coenzyme A. Next, the reactions aconitase A, *ACONTa*, and aconitase B, *ACONTb*, isomerize citrate to isocitrate via the metabolite cis-aconitate [32, 33]. Isocitrate sits at a branch point in the TCA cycle, where carbon flux either continues in the oxidative branch to the reaction catalyzed by isocitrate dehydrogenase, *ICDHyr* [34], or is diverted into the glyoxylate cycle by isocitrate lyase, *ICL*.

Isocitrate dehydrogenase catalyzes the decarboxylation of isocitrate, producing 2-oxoglutarate and CO_2 while reducing $NADP^+$ to NADPH. 2-oxoglutarate provides a carbon backbone for synthesis of glutamate and glutamine, the central metabolites in nitrogen metabolism. 2-oxoglutarate dehydrogenase, *AKGDH*, catalyzes the decarboxylation of 2-oxoglutarate, producing CO_2, reducing NAD^+ to NADH, and transferring coenzyme A to the decarboxylated compound to form succinyl-CoA [35].

During aerobic growth, the TCA cycle continues counterclockwise from succinyl-CoA. Succinyl-CoA synthetase, *SUCOAS*, generates ATP by separating succinate and coenzyme A [36]. Succinate dehydrogenase, *SUCDi*, is a multiprotein enzyme complex which straddles the cytoplasmic membrane allowing it to couple the TCA cycle to the electron transport chain [37]. The succinate dehydrogenase complex catalyzes the irreversible oxidation of succinate to fumarate while reducing ubiquinone-8, q8[c], to ubiquinol-8, q8h2[c] [38]. Ubiquinol-8 is then released from the enzyme complex and free to diffuse through the cytoplasmic membrane to interact with subsequent components of the electron transport chain. After succinate is converted to fumarate, fumarase, *FUM*, reversibly catalyzes the conversion of fumarate and water into L-malate [39-41]. Finally, to complete the TCA cycle, malate dehydrogenase, *MDH*, reversibly catalyzes the conversion of malate into oxaloacetate while reducing NAD^+ to NADH [42]. When this set of reactions is used in reverse as the reductive pathway of the TCA cycle, a reversing reaction is catalyzed by fumarate reductase, *FRD7* [43]. In the model, this reaction oxidizes ubiquinol-8 to ubiquinone-8, although in actual *E. coli*, fumarate reductase oxidizes the electron carrier menaquinol-8 instead [44]. This reaction had to be

included in an unrealistic form because the simplified electron transport system in the model includes only ubiquinone-8/ubiquinol-8 as an electron carrier.

2.1.4 Glyoxylate shunt, gluconeogenesis, and anaplerotic reactions

When growing on some substrates, the glyoxylate shunt is used instead of the full TCA cycle because it bypasses the reactions that lose carbon in the form of CO_2. The glyoxylate cycle consists of some of the reactions in the TCA cycle as well as other reactions used only by the glyoxylate cycle. It overlaps with the TCA cycle from the incorporation of acetyl-CoA to the production of isocitrate at the aforementioned branch point in the oxidative arm of the TCA cycle. Isocitrate lyase, *ICL*, catalyzes the cleavage of 6-carbon isocitrate into 4-carbon succinate and 2-carbon glyoxylate [45]. Malate synthase, *MALS*, then catalyzes the condensation of glyoxylate with another acetyl-CoA, yielding malate [46]. Succinate generated in the first step can continue along the TCA cycle to eventually form oxaloacetate.

In addition to growing on hexoses or pentoses, *E. coli* can also grow on 2-, 3-, or 4-carbon sources, such as lactate or malate, but in this situation, some 6-carbon metabolites in the glycolytic pathway are still required as precursors for biomass components. With only 2-, 3-, or 4-carbon sources in the environment, the glycolytic pathway can actually be reversed to produce net flux from pyruvate to glucose-6-phosphate. This reversal of glycolytic flux is referred to as gluconeogenesis. The two reactions of glycolysis that are effectively irreversible, catalyzed by phosphofructokinase, *PFK*, and pyruvate kinase, *PYK*, are replaced with two reversing reactions, catalyzed by fructose-bisphosphatase, *FBP*, and phosphoenolpyruvate synthase, *PPS*, respectively.

Phosphoenolpyruvate synthase [47] catalyzes the conversion of pyruvate to phosphoenolpyruvate and in the process hydrolyzes one ATP to AMP [48].

Anaplerotic reactions replenish TCA cycle intermediates drained off for biosynthesis. The TCA cycle operating cyclically can completely oxidize acetate to carbon dioxide without net consumption or production of intermediates. However, intermediates of the TCA cycle such as oxaloacetate and 2-oxoglutarate are consumed in the production of macromolecules. TCA cycle intermediate generation from the glycolytic metabolites is accomplished by the irreversible carbon dioxide-fixing conversion of the 3-carbon phosphoenolpyruvate to the 4-carbon oxaloacetate, catalyzed by the enzyme phosphoenolpyruvate carboxylase, *PPC*.

Growth on 4-carbon dicarboxylic acid intermediates of the TCA cycle, such as malate, requires that the cell be able to produce phosphoenolpyruvate for gluconeogenesis. There are two pathways existing to fulfill these phosphoenolpyruvate demands [49, 50]. One pathway involves the conversion of malate to pyruvate by malic enzyme, *ME1* or *ME2* [51, 52], followed by the synthesis of phosphoenolpyruvate from pyruvate by phosphoenolpyruvate synthase, *PPS* [53]. Malic enzyme, *ME1*, reduces one molecule of NAD^+ to NADH while converting malate to pyruvate. A second parallel reaction, malic enzyme (NADP), *ME2*, reduces one molecule of $NADP^+$ to NADPH. The other pathway from the TCA cycle to glycolytic intermediates is the conversion of oxaloacetate to phosphoenolpyruvate by the action of phosphoenolpyruvate carboxykinase, *PPCK* [54]. Phosphoenolpyruvate carboxykinase catalyzes the reverse reaction to the anaplerotic enzyme, phosphoenolpyruvate carboxylase, *PPC* [55]. The former reaction consumes a high-energy phosphate bond in ATP and produces CO_2,

whereas phosphoenolpyruvate carboxylase releases inorganic phosphate and consumes CO_2. Although the reactions catalyzed by the enzymes phosphoenolpyruvate carboxykinase and malic enzyme are thermodynamically reversible, physiologically they are found to operate unidirectionally [56, 57].

2.1.5 Electron transport chain, oxidative phosphorylation, and transfer of reducing equivalents

The electron transport chain and oxidative phosphorylation are used to produce the bulk of the cell's ATP under aerobic conditions. The electron transport chain translocates protons (H^+) from the cytoplasm, across the cytoplasmic membrane into the periplasmic space. Since the cytoplasmic membrane is effectively impermeable to protons and hydroxyl ions (OH^-), this establishes a difference in concentration of protons and a difference in electrical charge across the cytoplasmic membrane. This thermodynamic potential difference gives rise to a proton motive force which can be utilized to drive a myriad of endergonic reactions, such as synthesis of high energy currency metabolites like ATP. In the model, protons are translocated into the extracellular medium as a simplification, but this is a reasonable approximation given that the pH of the periplasm and extracellular medium is the same [58].

The electron transport chain of *E. coli* consists of several different respiratory dehydrogenases, quinones, and terminal reductases. There are 15 different dehydrogenases which accept electrons from donors such as NADH or succinate, then pass the electrons to one of three different quinones which then deliver the electrons to one of at least 14 different terminal reductases. The reductases complete the chain by

reducing a terminal electron acceptor such as oxygen or fumarate. Some, but not all dehydrogenases and reductases pump protons into the periplasm. It is possible for the dehydrogenases, quinones, and reductases to be used in many different combinations, so the entire system can be very complicated [59]. The core *E. coli* model condenses the sequence of steps in the electron transport chain into two reactions, representing generic NADH dehydrogenase and cytochrome oxidase reactions, connected by only one quinone. First, an NADH dehydrogenase, *NADH16*, catalyzes the oxidation of NADH to form NAD^+ while removing four protons from the cytoplasm, translocating three protons to the extracellular space and combining the fourth with a proton plus two electrons from NADH with ubiquinone-8 to form the reduced ubiquinol-8 [60, 61]. Ubiquinone-8 and ubiquinol-8 are oil soluble coenzymes which diffuse freely within the lipid environment of the cytoplasmic membrane. The next condensed step is when cytochrome oxidase, *CYTBD*, catalyzes the oxidation of ubiquinol-8 back to ubiquinone-8, which drives the translocation of two more protons into the extracellular space [62]. The two spare electrons are then combined with two cytoplasmic protons and an oxygen atom to form water. Oxygen spontaneously diffuses from the environment into the cell down a concentration gradient, *O2t*, and represents an exogenous source of terminal electron acceptor.

The enzyme ATP synthase, *ATPS4r*, catalyzes the synthesis of ATP from ADP, forming a high energy phosphate bond by coupling catalysis to the import of four protons that were pumped out by the electron transport chain [63, 64]. The exact number of high-energy phosphate bonds that are generated per oxygen atom used as a terminal acceptor is the P/O ratio. This value varies depending on periplasmic pH and other environmental

conditions, but the core model P/O ratio is stoichiometrically fixed at 1.25. The ATP maintenance reaction, *ATPM*, is not a real biochemical reaction. It is included for modeling purposes since the scope of the *E. coli* core model does not extend to all of the reactions that consume ATP in the cell. Adenylate kinase, *ADK1*, is a phosphotransferase enzyme that catalyzes the interconversion of adenine nucleotides, and plays an important role in cellular energy homeostasis [65, 66].

NADH is used for the catabolic activities of the cell, for instance driving the export of protons into the periplasmic space in the electron transport chain in the reaction catalyzed by NADH dehydrogenase, *NADH16*. In contrast NADPH is essential for anabolic metabolism such as the biosynthesis of building blocks for polymerization reactions from precursor metabolites produced by the fueling pathways of core metabolism. Maintaining the proper balance between anabolic reduction charge, $NADPH/NADP^+$, and catabolic reduction charge, $NADH/NAD^+$, is achieved by reactions catalyzed by transhydrogenase enzymes. NAD(P) transhydrogenase, *THD2*, catalyzes the transfer of a hydride ion from NADH to create NADPH, in a reaction coupled to the proton motive force [67]. The opposite transfer, of a hydride ion from NADPH to create NADH, is catalyzed by another enzyme, NAD transhydrogenase, *NADTRHD*, but it is not coupled to the translocation of protons [68]. This pair of reactions effectively allows the transfer of reducing equivalents between anabolic and catabolic reduction charge.

2.1.6 Fermentation

During aerobic respiration, oxygen is the terminal electron acceptor for the electron transport chain, and the ATP required for biosynthesis is produced by ATP

synthase. Under anaerobic conditions, *E. coli* can generate ATP by substrate level phosphorylation in the process of fermentation, where excess carbon is secreted as various organic by-products. Glycolysis results in the net production of two ATP per glucose by substrate level phosphorylation, but this is very low compared to the 17.5 ATP per glucose generated during aerobic respiration (in the model 17.5 ATP per glucose is generated, but this number can vary *in vivo*). The substrates of fermentation are typically sugars, so during fermentative growth, each cell must maintain a large flux through glycolysis to generate sufficient ATP to drive the constitutive biosynthesis, polymerization, and assembly reactions required for growth. This necessitates a large efflux of fermentative end products since there is insufficient ATP to assimilate all carbon as biomass. Approximately 10% of carbon substrate is assimilated due to the poor energy yield of fermentation [69]. Glycolysis also produces two molecules of NADH for each glucose; therefore NADH must be reoxidized by fermentation in order to regenerate NAD^+ to maintain the oxidation-reduction balance of the cell.

E. coli fermentation normally generates a mixture of end products anaerobically from sugars. The major soluble products are acetate, ethanol, lactate and formate, with a smaller amount of succinate [69]. In addition, fermentation results in the production of substantial quantities of carbon dioxide and hydrogen. Depending on the pH of the culture medium, and the redox state of the fermentation substrate, a cell may vary the relative flux through each of the fermentation pathways branching from pyruvate. When the pH of the environment drops due to increased concentrations of other acidic fermentative end products, such as acetic, formic or succinic acid, then flux may be increased through the reaction catalyzed by D-lactate dehydrogenase, *LDH_D* [70, 71].

This results in the reduction of pyruvate to form lactate while oxidizing NADH to form NAD^+.

Pyruvate formate lyase, *PFL*, catalyzes the non-oxidative cleavage of pyruvate to acetyl-CoA and formate, with the incorporation of coenzyme A into acetyl-CoA [72, 73]. Acetyl-CoA can then lead to either of the fermentative end products, acetate or ethanol. Both pathways involve two-step reversible mechanisms. In the conversion to acetate, phosphotransacetylase, *PTAr* [74], catalyzes transfer of a phosphate group onto the acetyl moiety of acetyl-CoA, to form acetyl-phosphate and release coenzyme A to be recycled for the previous reaction catalyzed by pyruvate formate lyase, *PFL*. Then acetate kinase, *ACKr*, catalyzes the conversion of acetyl-phosphate to acetate, in the process forming a much needed high energy phosphate bond by converting ADP to ATP [75]. The reactions catalyzed by phosphotransacetylase and acetate kinase are thermodynamically reversible. This allows *E. coli* to grow aerobically on acetate by reversing the flux through this fermentative pathway. During the conversion of acetyl-CoA to ethanol, two molecules of NADH are reoxidized, one by the first reaction catalyzed by acetaldehyde dehydrogenase (acetylating), *ACALD* [76], and the second by the subsequent reaction catalyzed by alcohol dehydrogenase (ethanol), *ALCD2x* [77]. Fermentation of many different substrates is possible because ethanol is more reduced than sugars, whereas acetate is more oxidized than ethanol. Redox balancing is achieved by varying the ratio of ethanol to acetate secreted. The end products of each fermentation pathway exit the cell along a concentration gradient and in the process transport a proton from the cytoplasm to the periplasmic space.

2.1.7 Nitrogen metabolism

Nitrogen is the fourth most abundant element in *E. coli* and enters the cell either by ammonium ion uptake, *NH4t*, or as a moiety within organic molecules such as the amino acids L-glutamine or L-glutamate. Glutamate is an extremely abundant metabolite, with a measured concentration of 100.55 μmol/gDW and production rate of 4.77 mmol/gDW/h. Glutamine is less abundant, with a measured concentration of 3.92 μmol/gDW and production rate of 3.36 mmol/gDW/h [78]. Glutamate is a central component of the nitrogen metabolism of *E. coli*, as it is a nitrogen donor in many reactions. In particular it is required for the synthesis of the other amino acids because its amino group is transferred to other compounds in transamination reactions. Glutamine is actively transported across the inner cytoplasmic membrane by an ATP binding cassette transporter, *GLNabc*, which imports one molecule of glutamine while consuming one ATP [79]. Glutamate is actively transported across the inner cytoplasmic membrane by proton symport, *GLUt2r* [80].

Direct ammonia assimilation is catalyzed by NADPH specific reductive amination of 2-oxoglutarate to glutamate by glutamate dehydrogenase (NADP), *GLUDy* [78]. Indirect ammonia assimilation is by a cyclic pair of sequential reactions catalyzed by glutamine synthetase, *GLNS* [81], and glutamate synthase (NADPH), *GLUSy* [78]. Glutamine synthetase first catalyzes the assimilation of ammonia by converting glutamate, with one amino moiety, into glutamine, with two amino moieties. Then glutamate synthase, *GLUSy*, catalyzes the transfer of the amide group of glutamine to 2-oxoglutarate, generating a second molecule of glutamate.

2.1.8 Biomass reaction

In order to represent growth, the core *E. coli* model includes a biomass reaction which drains precursor metabolites from the network at stoichiometrically fixed relative rates while also producing several by-product metabolites [1] (**Table 2.1**). These precursors are used to produce the lipids, proteins, nucleic acids, and other macromolecules required to replicate a cell. 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} g/cell [82]. Since most of the subunits of the cellular macromolecules, such as nucleic acids and amino acids, are not present in the core model, they could not be directly accounted for in the biomass reaction. The precursor metabolites in the core model that those macromolecular subunits are synthesized from are included instead. For example, the amino acid L-alanine is synthesized from pyruvate and L-glutamate, so both of these metabolites are consumed in the biomass reaction. Several metabolites are actually produced by the biomass reaction. ADP, protons, and inorganic phosphate are produced by the hydrolysis of ATP in the balanced reaction "atp + h2o → adp + h + pi." 2-oxoglutarate is produced during the synthesis of amino acids, when L-glutamate transfers its amino group to another compound in a transamination reaction. Coenzyme A is produced when acetyl-CoA is consumed, and NAD^+ is reduced to NADH and NADPH is oxidized to $NADP^+$ during biomass synthesis.

Table 2.1 In the biomass reaction, 23 different metabolites are consumed or produced in order to simulate growth. The metabolites that are consumed have negative stoichiometric coefficients and the metabolites that are produced have positive coefficients.

Abbr.	Metabolite	Stoichiometry
3pg	3-phospho-D-glycerate	-1.496
accoa	acetyl-CoA	-3.7478
adp	adenosine diphosphate	59.81
akg	2-oxoglutarate	4.1182
atp	adenosine triphosphate	-59.81
coa	coenzyme A	3.7478
e4p	D-erythrose 4-phosphate	-0.361
f6p	D-fructose 6-phosphate	-0.0709
g3p	glyceraldehyde 3-phosphate	-0.129
g6p	D-glucose 6-phosphate	-0.205
gln-L	L-glutamine	-0.2557
glu-L	L-glutamate	-4.9414
h	H^+	59.81
h2o	H_2O	-59.81
nad	nicotinamide adenine dinucleotide (NAD^+)	-3.547
nadh	nicotinamide adenine dinucleotide-reduced	3.547
nadp	nicotinamide adenine dinucleotide phosphate ($NADP^+$)	13.0279
nadph	nicotinamide adenine dinucleotide phosphate-reduced	-13.0279
oaa	oxaloacetate	-1.7867
pep	phosphoenolpyruvate	-0.5191
pi	phosphate	59.81
pyr	pyruvate	-2.8328
r5p	alpha-D-ribose 5-phosphate	-0.8977

Additional energetic requirements exist for growth beyond what is needed to generate the macromolecular content of the cell. These energetic maintenance requirements are for growth associated maintenance (e.g., protein polymerization costs) and non-growth-associated maintenance (e.g., membrane leakage). To represent growth associated maintenance, ATP is converted to ADP at 59.81 mmol/gDW/h, accounting for energy used in cell division and other growth processes. Non-growth associated

maintenance is not part of the biomass reaction. Instead, it is represented with a lower bound of 8.39 mmol/gDW/h on the ATP maintenance reaction (*ATPM*), simulating energy used for protein turnover and other processes that do not change with growth.

2.2 Construction and content of the core *E. coli* regulatory network

In addition to the metabolic reconstruction, the core *E. coli* model also contains a Boolean representation of part of the associated transcriptional regulatory network. This network is a modified subset of the genome-scale transcriptional regulatory reconstruction *i*MC1010 [83]. In response to external and internal stimuli, *in silico* transcription factors either activate or repress genes associated with metabolic reactions. This regulation improves the predictive fidelity of the metabolic model by imposing additional context dependent constraints on certain reactions. The transcriptional regulatory reconstruction consists of a set of Boolean rules that dictate whether a gene is either fully induced or fully repressed. If the genes associated with an enzyme or transport protein/complex are repressed, then *in silico* flux is constrained to zero for the corresponding reaction. The solution space of the network shrinks when these additional constraints are imposed. Reactions that are not used due to regulatory effects are thus restricted, so when using flux balance analysis (FBA), the optimal flux distribution will be consistent with known regulation. This optimal flux distribution may be different from the flux distribution of an unregulated model. In this case, the flux distribution of the unregulated model violated at least one regulatory constraint, making it biologically

unrealistic. The use of computationally implemented Boolean rules in a genome-scale model has been shown to lead to more accurate FBA predictions [83].

A gene is considered to be induced when evaluation of the corresponding Boolean rule gives "true." In contrast, a gene is considered to be repressed when evaluation of the corresponding Boolean rule gives "false." Boolean logic is used to evaluate each Boolean rule. For example, consider the enzyme phosphoenolpyruvate synthase, *PPS*, which catalyzes the first step of gluconeogenesis, the conversion of pyruvate to phosphoenolpyruvate. The gene for phosphoenolpyruvate synthase is *pps* and its Boolean rule is simply "FruR." That is, if FruR is "true" then the *pps* gene is induced allowing *in silico* flux through the reaction catalyzed by phosphoenolpyruvate synthase, *PPS*. FruR is a transcriptional regulator which is active when the cytoplasmic concentration of D-fructose-1,6-bisphosphate, *FDP*, is low [84]. The Boolean rule for FruR is "NOT *surplusFDP*." That is, if there is no surplus of D-fructose-1,6-bisphosphate, then FruR is "true" and therefore the *pps* gene is induced, allowing *in silico* gluconeogenic flux through the reaction catalyzed by phosphoenolpyruvate synthase, *PPS*. In contrast, if *surplusFDP* is "true," then FruR is "false" and therefore *pps* is repressed.

Regulatory conditions, such as *surplusFDP*, are variables that represent a complex regulatory rule for a transcription factor that cannot be accurately represented with only one variable. The regulatory rule for *surplusFDP* is "((NOT *FBP*) AND (NOT (*TKT2* OR *TALA* OR *PGI*))) OR fru[e]." If fru[e] is "true," then *surplusFDP* is "true," independent of the state of the other variables. If fructose-bisphosphatase, *FBP*, is "false," and any one of transketolase, *TKT2*, transaldolase, *TALA*, or glucose-6-phosphate isomerase, *PGI*, is "false," then *surplusFDP* is "true," therefore FruR is "false" and *pps* is

repressed. By using Boolean logic, all rules in a regulatory network can be reduced to either "true" or "false," and ultimately this dictates whether each metabolic gene is induced or repressed. Not every gene in the metabolic network is controlled by the regulatory network, so the unregulated genes are assumed to always be active, and their fluxes are never constrained to zero. An abstract overview of part of the regulatory network is depicted in **Figure 2.2**.

2.3 Computational characterization of the core *E. coli* model

In this section, the use of the core *E. coli* computational model to predict various properties and phenotypes of *E. coli* metabolism is presented. The COBRA Toolbox [85] was used to perform all calculations, and the code used is presented in **Bold Arial** text. Several different constraint-based analysis procedures are presented and explained here. Most of these methods rely on FBA or related constraint-based procedures (see **Section 1.1.3 Flux balance analysis**).

2.3.1 Determination of growth rates on different substrates: simple FBA

Growth of *E. coli* on glucose can be simulated under aerobic conditions. To set the maximum glucose uptake rate to 18.5 mmol/gDW/h (millimoles per gram dry cell weight per hour, the default flux units used in the COBRA Toolbox), enter into Matlab:

```
model = changeRxnBounds(model,'EX_glc(e)',-18.5,'l');
```

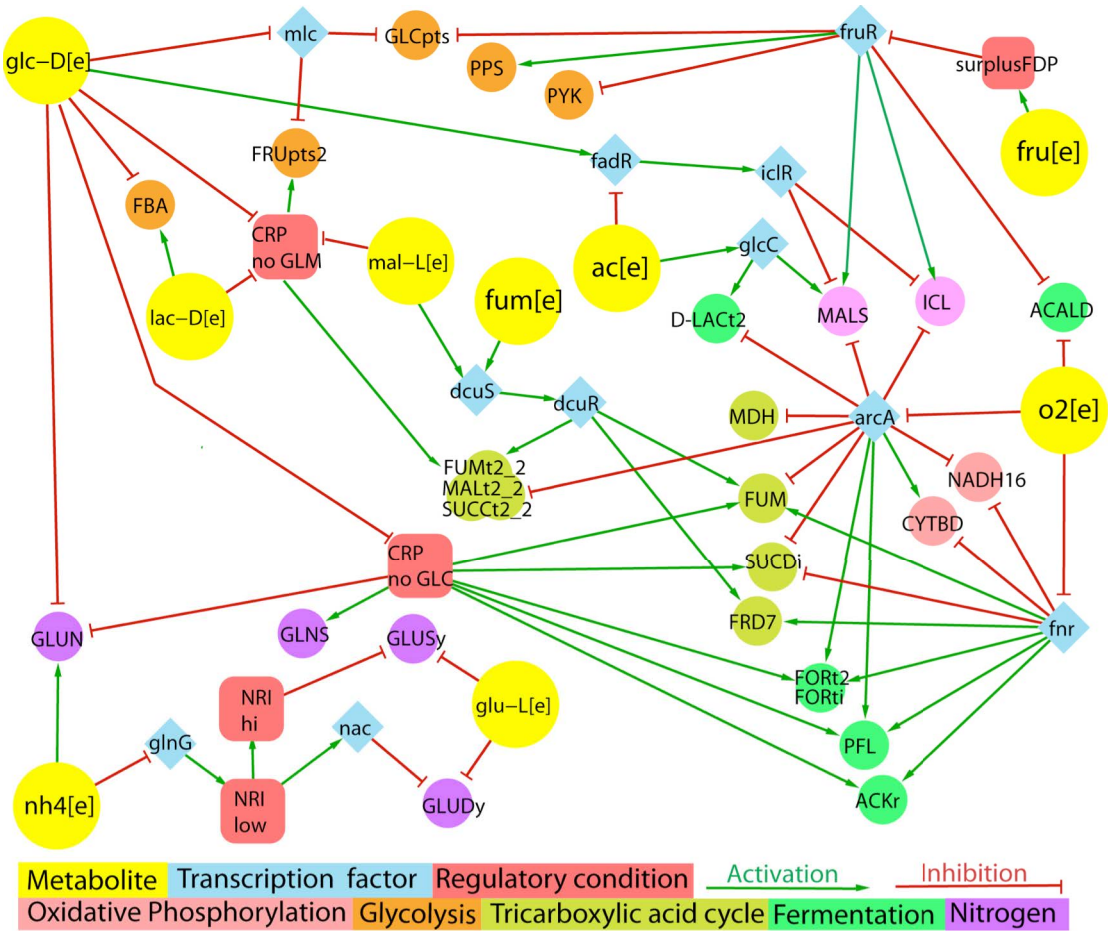

Figure 2.2 A schematic overview of part of the *E. coli* core regulatory network from an environmental stimulus-response perspective. For the sake of clarity, reaction abbreviations are displayed rather than the gene(s) which code for the associated protein or protein complex. The Boolean state of *in silico* reactions, true = flux, or false = zero flux, may be determined by traversing the activation or inhibition links, beginning at an environmental metabolite stimulus (yellow circles).

This changes the lower bound (**'l'**) of the glucose exchange reaction to -18.5, a biologically realistic uptake rate. By convention, exchange reactions are written as export reactions (e.g., **'glc[e] <==>'**), so import of a metabolite is a negative flux. To allow unlimited oxygen uptake, enter:

```
model = changeRxnBounds(model,'EX_o2(e)',-1000,'l');
```

By setting the lower bound of the oxygen uptake reaction to such a large number, it is practically unbounded. Next, to ensure that the biomass reaction is set as the objective function, enter:

```
model = changeObjective(model,'Biomass_Ecoli_core_w_GAM');
```

To perform FBA with maximization of the biomass reaction as the objective, enter:

```
FBAsolution = optimizeCbModel(model,'max');
```

FBAsolution.f then gives the value of the objective function (Z) as 1.6531. This means that the model predicts a growth rate of 1.6531 h^{-1}. Inspection of the flux distribution vector **FBAsolution.x** (**v**) shows that there is high flux in the glycolysis, pentose phosphate, TCA cycle, and oxidative phosphorylation pathways, and that no organic by-products are secreted (**Figure 2.3 a**).

Next, the same simulation is performed under anaerobic conditions. With the same model, enter:

```
model = changeRxnBounds(model,'EX_o2(e)',0,'l');
```

The lower bound of the oxygen exchange reaction is now 0, so oxygen may not enter the system. When **optimizeCbModel** is used as before, the resulting growth rate is now much lower, 0.4706 h^{-1}. The flux distribution shows that oxidative phosphorylation is not used in these conditions, and that acetate, formate, and ethanol are produced by fermentation pathways (**Figure 2.3 b**).

Just as FBA was used to calculate growth rates of *E. coli* on glucose, it can also be used to simulate growth on other substrates. The core *E. coli* model contains exchange reactions for 13 different organic compounds, each of which can be used as the sole carbon source under aerobic conditions. For example, to simulate growth on succinate

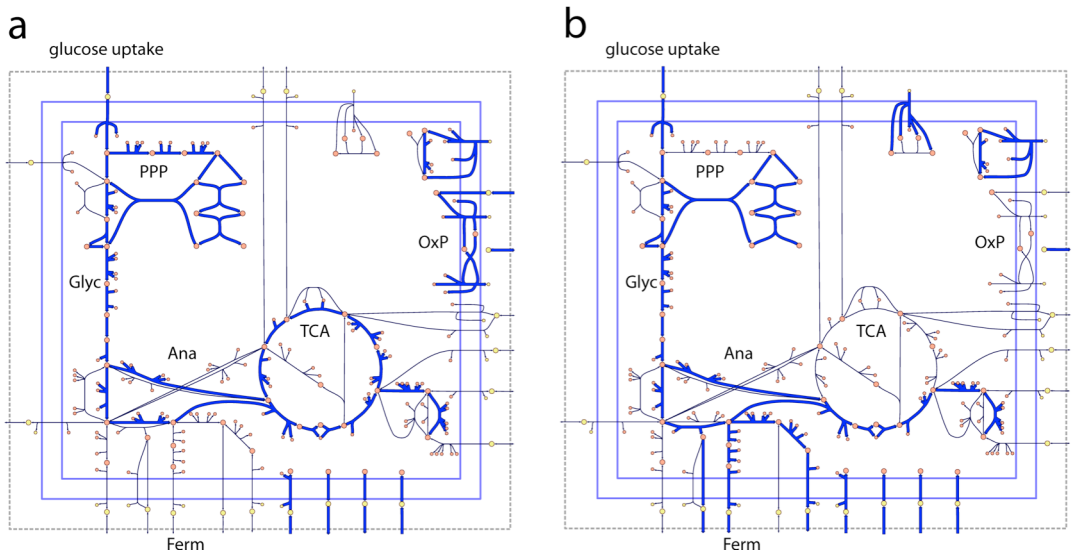

Figure 2.3 Flux distributions computed by FBA can be visualized on network maps. In these two examples, the thick blue arrows represent reactions carrying flux, and the thin black arrows represent unused reactions. These maps show the state of the *E. coli* core model with maximum growth rate as the objective (*Z*) under aerobic (**a**) and anaerobic (**b**) conditions. The metabolic pathways shown in these maps are glycolysis (Glyc), pentose phosphate pathway (PPP), TCA cycle (TCA), oxidative phosphorylation (OxP), anaplerotic reactions (Ana), and fermentation pathways (Ferm).

instead of glucose, first use the **changeRxnBounds** function to set the lower bound of the glucose exchange reaction (*EX_glc(e)*) to 0. Then use **changeRxnBounds** to set the lower bound of the succinate exchange reaction (*EX_succ(e)*) to -20 mmol/gDW/h, an arbitrary uptake rate). As in the glucose examples, make sure that *Biomass_Ecoli_core_w_GAM* is set as the objective (the function **checkObjective** can be used to identify the objective reaction(s)), and use **optimizeCbModel** to perform FBA. The growth rate, given by **FBAsolution.f**, will be 0.8401 h^{-1}. The full code to calculate growth on succinate (with the model starting with its default bounds and objective) is:

```
model = changeRxnBounds(model,'EX_glc(e)',0,'l');
model = changeRxnBounds(model,'EX_succ(e)',-20,'l');
FBAsolution = optimizeCbModel(model,'max');
```

Growth can also be simulated under anaerobic conditions with any substrate by using **changeRxnBounds** to set the lower bound of the oxygen exchange reaction (*EX_o2(e)*) to 0 mmol/gDW/h, so no oxygen can enter the system. When this constraint is applied and succinate is the only organic substrate, **optimizeCbModel** returns a growth rate of 0 h^{-1}, and does not calculate a flux vector **v** (depending on which linear programming solver is used with the COBRA Toolbox, a growth rate may not be calculated at all). In this case, FBA predicts that growth is not possible on succinate under anaerobic conditions. Because the maximum amount of ATP that can be produced from this amount of succinate is less than the minimum bound of 8.39 mmol/gDW/h of the ATP maintenance reaction, *ATPM*, there is no feasible solution. FBA predicted growth rates for all 13 organic substrates in the *E. coli* core model under both aerobic and anaerobic conditions are shown in **Table 2.2**. The growth rates are all much lower anaerobically (0 h^{-1} in most cases) because the electron transport chain cannot be used to fully oxidize the substrates and generate ATP.

2.3.2 Production of cofactors and biomass precursors

FBA can also be used to determine the maximum yields of important cofactors and biosynthetic precursors from glucose and other substrates [86]. In this example, the maximum yields of the cofactors ATP, NADH, and NADPH from glucose under aerobic conditions are calculated. To calculate optimal ATP production, first use **changeRxnBounds** to constrain the glucose exchange reaction (*EX_glc(e)*) to exactly -1 mmol/gDW/h by setting both the lower and upper bounds to -1 (**'b'**). Next, set the ATP maintenance reaction (*ATPM*) as the objective to be maximized using **changeObjective**.

Table 2.2 The maximum growth rate of the core *E. coli* model on its 13 different organic substrates, computed by FBA. Growth rate was calculated for both aerobic and anaerobic conditions for each substrate, and the maximum substrate uptake rate was set to 20 mmol/gDW/h for every substrate.

	Growth Rate (h^{-1})	
Substrate	**Aerobic**	**Anaerobic**
acetate	0.3893	0
acetaldehyde	0.6073	0
2-oxoglutarate	1.0982	0
ethanol	0.6996	0
D-fructose	1.7906	0.5163
fumarate	0.7865	0
D-glucose	1.7906	0.5163
L-glutamine	1.1636	0
L-glutamate	1.2425	0
D-lactate	0.7403	0
L-malate	0.7865	0
pyruvate	0.6221	0.0655
succinate	0.8401	0

The reaction *ATPM* is a stoichiometrically balanced reaction that hydrolyzes ATP (atp[c]) and produces ADP (adp[c]), inorganic phosphate (pi[c]), and a proton (h[c]). It works as an objective for maximizing ATP production because in order to consume the maximum amount of ATP, the network must first produce ATP by the most efficient pathways available by recharging the produced ADP. The constraint on this reaction should be removed by using **changeRxnBounds** to set the lower bound to 0. By default, this reaction has a lower bound of 8.39 mmol/gDW/h to simulate non-growth associated maintenance costs. Use **optimizeCbModel** to calculate the maximum yield of ATP, which is 17.5 mol ATP/mol glucose. The full COBRA Toolbox code to perform this calculation (with the model starting with its default bounds and objective) is:

```
model = changeRxnBounds(model,'EX_glc(e)',-1,'b');
model = changeObjective(model,'ATPM');
model = changeRxnBounds(model,'ATPM',0,'l');
FBAsolution = optimizeCbModel(model,'max');
```

Calculation of the yields of NADH and NADPH one at a time can be performed in a similar manner. First, constrain *ATPM* to 0 mmol/gDW/h flux (**'b'**) so the cell is not required to produce ATP, and also cannot consume any ATP using this reaction. Add stoichiometrically balanced NADH and NADPH consuming reactions using the function **addReaction**, and set these as the objectives using **changeObjective**. The maximum yields of ATP, NADH, and NADPH are shown in **Table 2.3**. The full code to calculate the maximum NADH yield is:

```
model = changeRxnBounds(model,'EX_glc(e)',-1,'b');
model = changeRxnBounds(model,'ATPM',0,'b');
model = addReaction(model,'NADH_drain','nadh[c] -> nad[c] + h[c]');
model = changeObjective(model,'NADH_drain');
FBAsolution = optimizeCbModel(model,'max');
```

The sensitivity of an FBA solution is indicated by either shadow prices or reduced costs. Shadow prices are the derivative of the objective function with respect to the exchange flux of a metabolite. They indicate how much the addition of that metabolite will increase or decrease the objective. Reduced costs are the derivatives of the objective function with respect to an internal reaction with 0 flux, indicating how much each particular reaction affects the objective. In the COBRA Toolbox, shadow prices and reduced costs can be calculated by **optimizeCbModel**. The vector of *m* shadow prices is **FBAsolution.y** and the vector of *n* reduced costs is **FBAsoltuion.w**.

ATP production is limited by cellular proton balancing. The shadow price of cytosolic protons (h[c]) is -0.25, indicating that the addition of 4 mol protons/mol glucose to the system reduces ATP yield by 1 mol ATP/mol glucose. Protons are produced by

Table 2.3 The maximum yields of the cofactors ATP, NADH, and NADPH from glucose under aerobic conditions. "ATP Shadow Price" is the shadow price of the metabolite atp[c], and indicates how much the addition of ATP to the system will increase the yield of the cofactor. "Constraint" indicates what the limiting constraints on the yields are. Energy constraints are due to the limited availability of ATP, while stoichiometry constraints indicate that the structure of the network prevents maximum yield.

Cofactor	Yield (mol/mol glc)	ATP Shadow Price	Constraint
ATP	17.5	0	H^+ balancing
NADH	10	0.5714	Energy, Stoichiometry
NADPH	8.778	0.4444	Energy, Stoichiometry

various metabolic reactions and are also pumped into the cell by the ATP synthase reaction (*ATPS4r*). In order for the system to be at steady state (**Sv** = 0), an equal number of protons must be pumped out by the electron transport chain reactions or by excreting metabolites with symporters. If more ATP were to be produced by ATP synthase, it would import additional protons that have no way to escape the cell. The flux distribution for optimal ATP production is shown in **Figure 2.4**.

NADH and NADPH production are also ultimately limited by proton balancing. For maximum NADH yield, the proton shadow price is -0.1429. For maximum NADPH yield, the proton shadow price is -0.1111. The protons produced in metabolism are removed by *ATPS4r* in reverse (with a negative flux), which consumes ATP. The stoichiometry of the network also limits the production of NADH and NADPH. Glucose has 12 electron pairs that can be donated to redox carriers such as NAD^+ or $NADP^+$, but when the TCA cycle is used, two of these electron pairs are used to reduce the quinone q8[c] in the succinate dehydrogenase reaction (*SUCDi*), and thus cannot be used to produce NADH or NADPH. The only pathway that can reduce 12 redox carriers with one

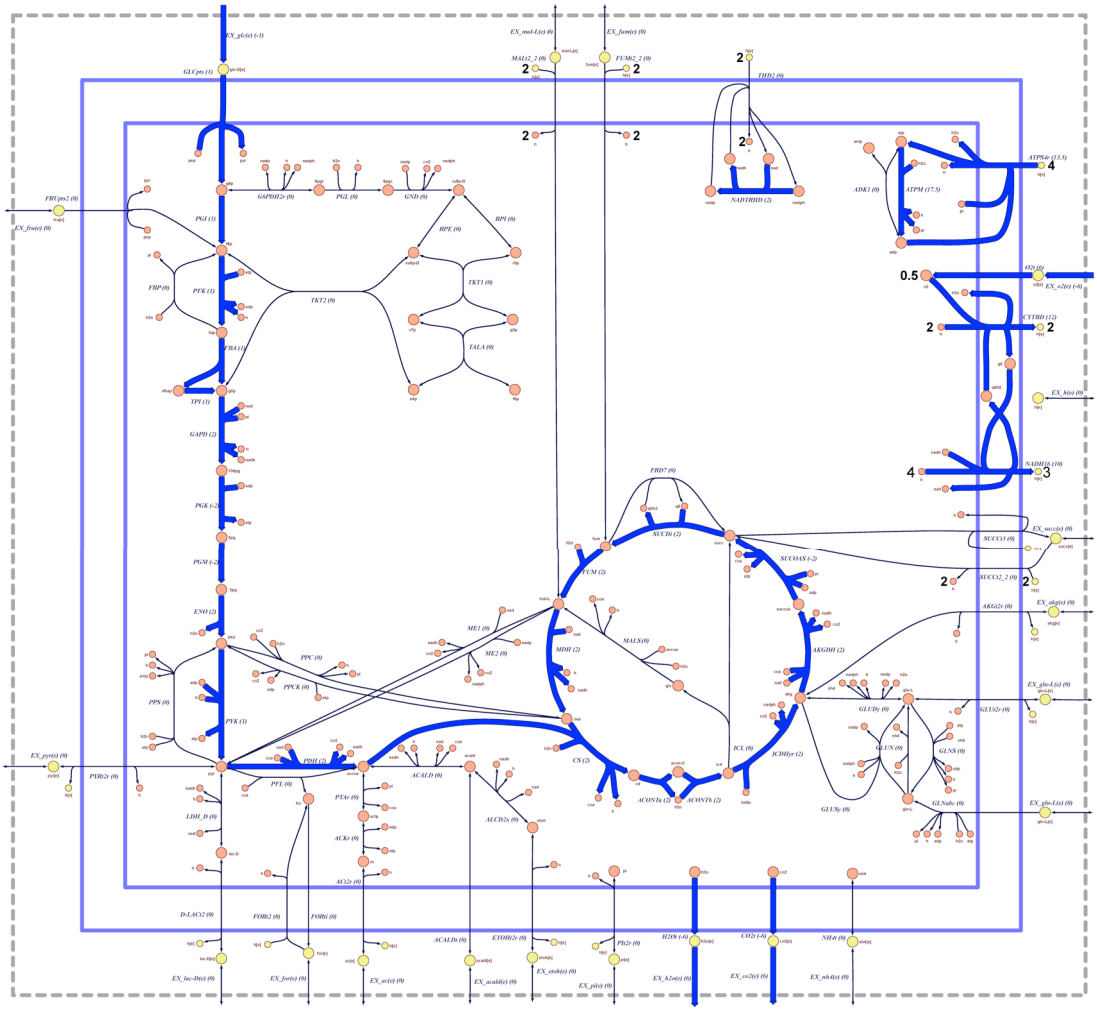

Figure 2.4 Flux map for maximum ATP yield from glucose under aerobic conditions. Thick blue lines indicate reactions carrying flux in this particular solution vector. This is a unique solution.

molecule glucose is the pentose phosphate pathway, but this is infeasible because this pathway generates no ATP, which is needed to pump out the protons that are produced.

The production of these cofactors can also be computed under anaerobic conditions by setting the lower bound of the oxygen exchange reaction (*EX_o2(e)*) to 0 mmol/gDW/h. The results of these calculations are shown in **Table 2.4**.

Table 2.4 The maximum yields of the cofactors ATP, NADH, and NADPH from glucose under anaerobic conditions. "ATP Shadow Price" is the shadow price of the metabolite atp[c], and indicates how much the addition of ATP to the system will increase the yield of the cofactor. "Constraint" indicates what the limiting constraints on the yields are.

Cofactor	Yield (mol/mol glc)	ATP Shadow Price	Constraint
ATP	2.75	0	H^+ balancing
NADH	6	1	Energy
NADPH	4	1.333	Energy

The core *E. coli* model contains 12 basic biosynthetic precursor compounds that are used to build macromolecules such as nucleic acids and proteins. The maximum yield of each of these precursor metabolites from glucose can be calculated by adding a demand reaction for each one (a reaction that consumes the compound without producing anything) and setting these as the objectives for FBA. Maximum yields of each of the 12 precursors are listed in **Table 2.5**. Note that the drain reactions for acetyl-CoA (accoa[c]) and succinyl-CoA (succoa[c]) produce coenzyme A (coa[c]), since this carrier molecule is not produced from glucose in the core model.

2.3.3 Alternate optimal solutions

The flux distribution calculated by FBA is often not unique. In many cases, it is possible for a biological system to achieve the same objective value by using alternate pathways, so phenotypically different alternate optimal solutions are possible. One method that uses FBA to identify alternate optimal solutions is Flux Variability Analysis (FVA) [87]. This is a method that identifies the maximum and minimum possible fluxes through a particular reaction with the objective value constrained to be close to or equal

Table 2.5 The maximum yields of different biosynthetic precursors from glucose under aerobic conditions. The precursors are 3pg (3-phospho-D-glycerate), pep (phosphoenolpyruvate), pyr (pyruvate), oaa (oxaloacetate), g6p (D-glucose-6-phosphate), f6p (D-fructose-6-phosphate), r5p (α-D-ribose-5-phosphate), e4p (D-erythrose-4-phosphate), g3p (glyceraldehyde-3-phosphate), accoa (acetyl-CoA), akg (2-oxoglutarate), and succoa (succinyl-CoA). "Carbon Conversion" indicates what percentage of the carbon atoms in glucose are converted to the precursor compound. "ATP Shadow Price" is the shadow price of the metabolite atp[c]. "Constraint" indicates what the limiting constraints on the yields are, preventing a yield of at least 100%. Some precursors have a yield of over 100% because carbon from CO_2 can be fixed in some reactions.

Precursor	Yield (mol/mol glc)	Carbon Conversion	ATP Shadow Price	Constraint
3pg	2	100%	0	-
pep	2	100%	0	-
pyr	2	100%	0	-
oaa	2	133.33%	0	-
g6p	0.8916	89.16%	0.0482	Energy
f6p	0.8916	89.16%	0.0482	Energy
r5p	1.0571	88.10%	0.0571	Energy
e4p	1.2982	86.55%	0.0702	Energy
g3p	1.6818	84.09%	0.0909	Energy
accoa	2	66.67%	0	Stoichiometry
akg	1	83.33%	0	Stoichiometry
succoa	1.64	109.33%	0	-

to its optimal value. Performing FVA on a single reaction using the basic COBRA Toolbox functions is simple. First, use functions **changeRxnBounds**, **changeObjective**, and **optimizeCbModel** to perform FBA as described in the previous examples. Get the optimal objective value *Z* (**FBAsolution.f**), and then set both the lower and upper bounds of the objective reaction to exactly this value. Next, set the reaction of interest as the objective, and use FBA to minimize and maximize this new objective in two separate steps. This will give the minimum and maximum possible fluxes through this reaction while contributing to the optimal objective value.

For example, consider the variability of the malic enzyme reaction (*ME1*) in *E. coli* growing on succinate. The minimum possible flux through this reaction is 0 mmol/gDW/h and the maximum is 6.49 mmol/gDW/h. In one alternate optimal solution, the *ME1* reaction is used, but in another, it is not used at all. The full code to set the model to these conditions and perform FVA on this reaction is:

```
model = changeRxnBounds(model,'EX_glc(e)',0,'l');
model = changeRxnBounds(model,'EX_succ(e)',-20,'l');
FBAsol = optimizeCbModel(model,'max');
model = changeRxnBounds(model,'Biomass_Ecoli_core_w_GAM',FBAsol.f,'b');
model = changeObjective(model,'ME1');
FBAsolutionMin = optimizeCbModel(model,'min');
FBAsolutionMax = optimizeCbModel(model,'max');
```

The COBRA Toolbox includes a built in function for performing FVA called **fluxVariability**. This function is useful because it performs FVA on every reaction in a model. When FVA is performed on every reaction in the *E. coli* core model for growth on succinate, eight reactions are found to be variable (**Table 2.6**). Inspection of the variable reactions shows that conversion of L-malate to pyruvate may occur through several different pathways, each leading to the same maximum growth rate. Flux distributions using combinations of these pathways are also valid solutions. Two of these alternate solutions are shown in **Figure 2.5**.

2.3.4 Robustness analysis

Another method that uses FBA to analyze network properties is robustness analysis [88]. In this method, the flux through one reaction is varied and the optimal objective value is calculated as a function of this flux. This reveals how sensitive the

Table 2.6 Variable reactions for growth on succinate (uptake rate = 20 mmol/gDW/h) under aerobic conditions. The minimum and maximum possible flux for every reaction was calculated at the maximum growth rate and only reactions with variable fluxes are shown here. *FRD7* (fumarate reductase) and *SUCDi* (succinate dehydrogenase) always have highly variable fluxes in this model because they form a cycle that can carry any flux. Physiologically, these fluxes are not relevant. The other variable reactions are *MDH* (malate dehydrogenase), *ME1* (malic enzyme (NAD)), *ME2* (malic enzyme (NADP)), *NADTRHD* (NAD transhydrogenase), *PPCK* (phosphoenolpyruvate carboxykinase), and *PYK* (pyruvate kinase).

Reaction	Minimum Flux (mmol/gDW/h)	Maximum Flux (mmol/gDW/h)
FRD7	0	972.77
MDH	13.56	20.06
ME1	0	6.49
ME2	7.17	13.67
NADTRHD	0	6.49
PPCK	3.93	10.42
PYK	0	6.49
SUCDi	27.23	1000

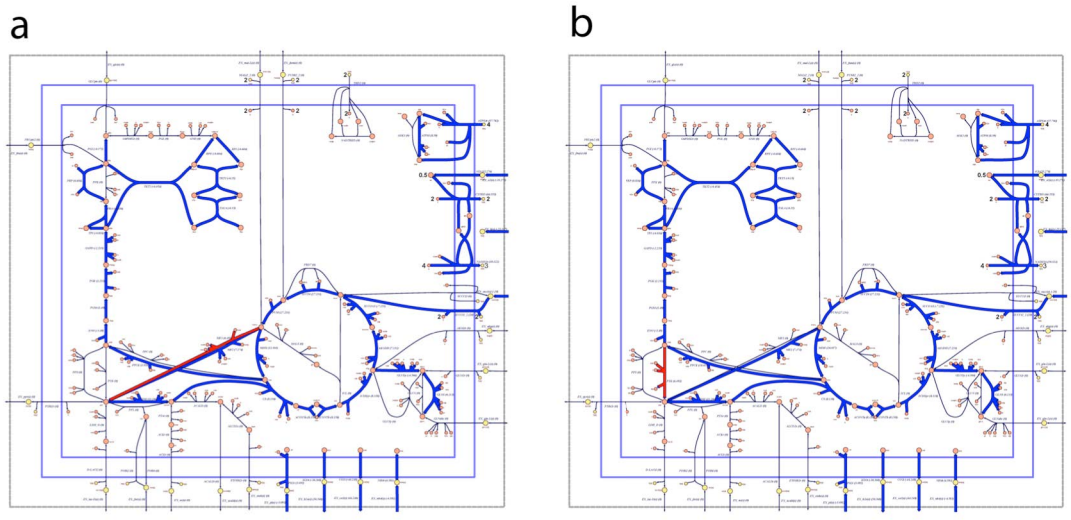

Figure 2.5 Flux maps for two alternate solutions for maximum aerobic growth on succinate. In (**a**), the reaction *ME1* is used to convert L-malate to pyruvate, but in (**b**), this reaction is not used at all, and the reaction *PYK* is used. The two alternative reactions are highlighted in red.

objective is to a particular reaction. There are many informative combinations of reaction and objective that can be investigated with robustness analysis. One example is examination of the effects of nutrient uptake on growth rate. To determine the effect of varying glucose uptake on growth, first use **changeRxnBounds** to set the oxygen uptake rate (*EX_o2(e)*) to 17 mmol/gDW/h, a realistic uptake rate. Then, use a for loop to set both the upper and lower bounds of the glucose exchange reaction to values between 0 and -20 mmol/gDW/h, and use **optimizeCbModel** to perform FBA with each uptake rate. Be sure to store each growth rate in a vector or other type of Matlab list. The COBRA Toolbox also contains a function for performing robustness analysis (**robustnessAnalysis**) that can perform these functions. The full code to perform this robustness analysis is:

```
model = changeRxnBounds(model,'EX_o2(e)',-17,'b');
growthRates = zeros(21,1);
for i = 0:20
        model = changeRxnBounds(model,'EX_glc(e)',-i,'b');
        FBAsolution = optimizeCbModel(model,'max');
        growthRates(i+1) = FBAsolution.f;
end
```

The results can then be plotted, as in **Figure 2.6**. As expected, with a glucose uptake of 0 mmol/gDW/h, the maximum possible growth rate is 0 h^{-1}. Growth remains at 0 h^{-1} until a glucose uptake rate of about 2.83 mmol/gDW/h, because with such a small amount of glucose, the system cannot make 8.39 mmol/gDW/h ATP to meet the default lower bound of the ATP maintenance reaction (*ATPM*). Once enough glucose is available to meet this ATP requirement, growth rate increases rapidly as glucose uptake increases. Above an uptake rate of about 7.59 mmol/gDW/h, growth does not increase as rapidly. It

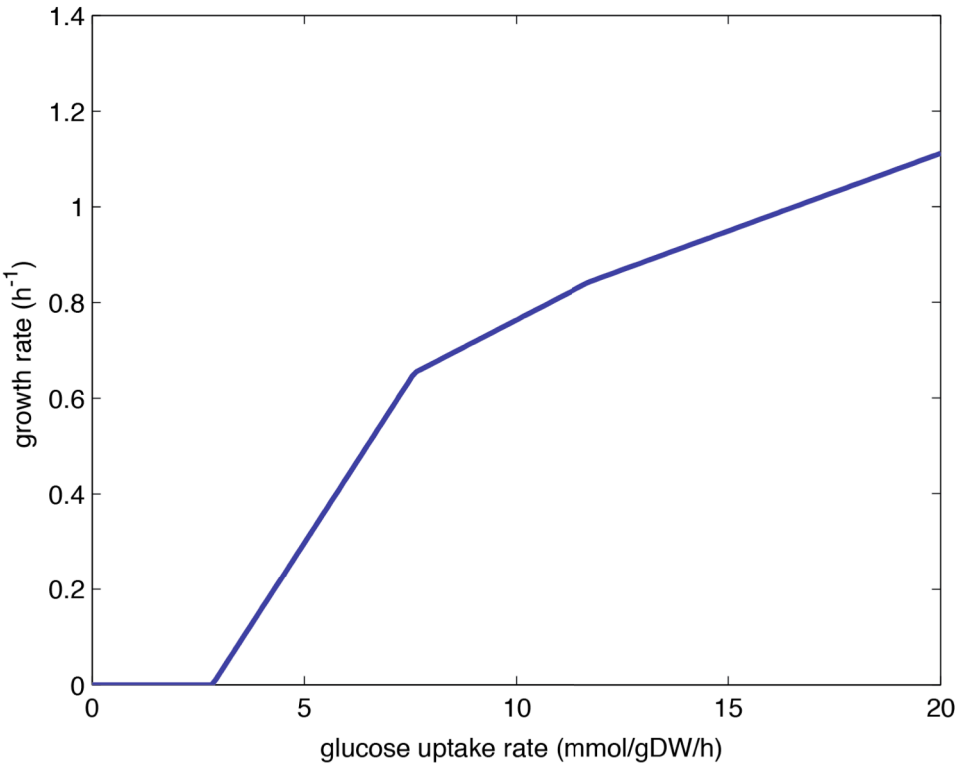

Figure 2.6 Robustness analysis for maximum growth rate while varying the glucose uptake rate with the oxygen uptake fixed at 17 mmol/gDW/h.

is at this point that oxygen and not glucose limits growth. Excess glucose cannot be fully oxidized, so the fermentation pathways are used.

The oxygen uptake rate can also be varied with the glucose uptake rate held constant. With glucose uptake fixed at 10 mmol/gDW/h, growth rate increases steadily as oxygen uptake increases (**Figure 2.7**). At an oxygen uptake of about 21 mmol/gDW/h, growth actually begins to decrease as oxygen uptake increases. This is because glucose becomes limiting at this point, and glucose that would have been used to produce biomass must instead be used to reduce excess oxygen. In the previous example, the growth rate continues to increase once oxygen become limiting because *E. coli* can grow

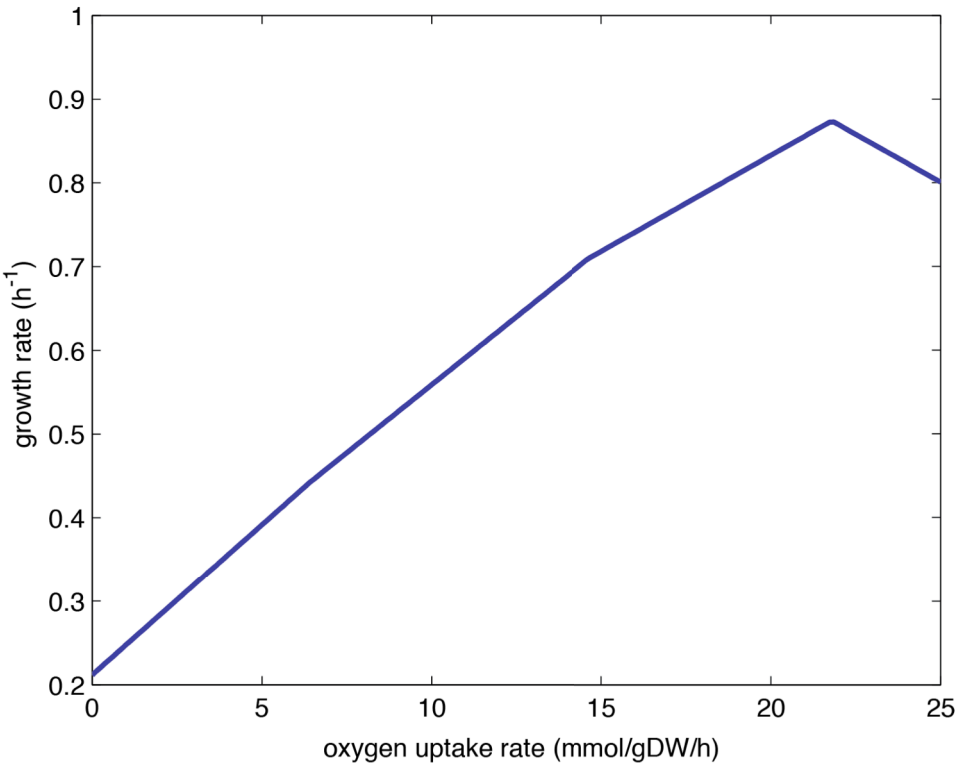

Figure 2.7 Robustness analysis for maximum growth rate while varying oxygen uptake rate with glucose uptake fixed at 10 mmol/gDW/h.

on glucose without oxygen. In this example, *E. coli* cannot grow with oxygen but not glucose (or another carbon source), so growth decreases when excess oxygen is added.

2.3.5 Phenotypic phase plane analysis

When performing robustness analysis, one parameter is varied and the network state is calculated. It is also possible to vary two parameters simultaneously and plot the results as a phenotypic phase plane [89]. These plots can reveal the interactions between two reactions in interesting ways. As an example, the phenotypic phase plane for maximum growth while varying glucose and oxygen uptake rates will be calculated. Although more sophisticated methods for computing phenotypic phase planes exist [90],

they can be easily computed in a manner similar to the calculations for robustness analysis. Instead of using one for loop to vary one reaction, two nested for loops are used to vary two reactions. In this case, use for loops to vary the bounds of the glucose exchange reaction (*EX_glc(e)*) and oxygen exchange reaction (*EX_o2(e)*) between 0 and -20 mmol/gDW/h. Use **optimizeCbModel** to perform FBA at each combination of glucose and oxygen uptake rates. The full code to perform the calculations is:

```
growthRates = zeros(21);
for i = 0:20
        for j = 0:20
                model = changeRxnBounds(model,'EX_glc(e)',-i,'b');
                model = changeRxnBounds(model,'EX_o2(e)',-j,'b');
                FBAsolution = optimizeCbModel(model,'max');
                growthRates(i+1,j+1) = FBAsolution.f;
        end
end
```

The resulting growth rates can be plotted as a 2-D graph or as a 3-D surface (**Figure 2.8**). It is clear from these plots that this surface has 5 distinct regions, and each one is a flat plane. This is a general feature of phenotypic phase planes. They do not form curved surfaces or other shapes. Each of these phases has a qualitatively distinct phenotype, and all of the shadow prices (**FBAsolution.y**) are constant within each phase. Phase 1 (on the far left of the plots) is characterized by 0 growth. There is not enough glucose to meet the ATP maintenance requirement imposed by the *ATPM* reaction. In phase 2, growth is limited by oxygen. o2[e] has a shadow price of -0.0229 because there is not enough glucose to reduce all of the oxygen and produce biomass optimally. The line between phase 2 and phase 3 is where glucose and oxygen are perfectly balanced and growth yield is highest. In phases 3, 4, and 5, oxygen and glucose are both limiting

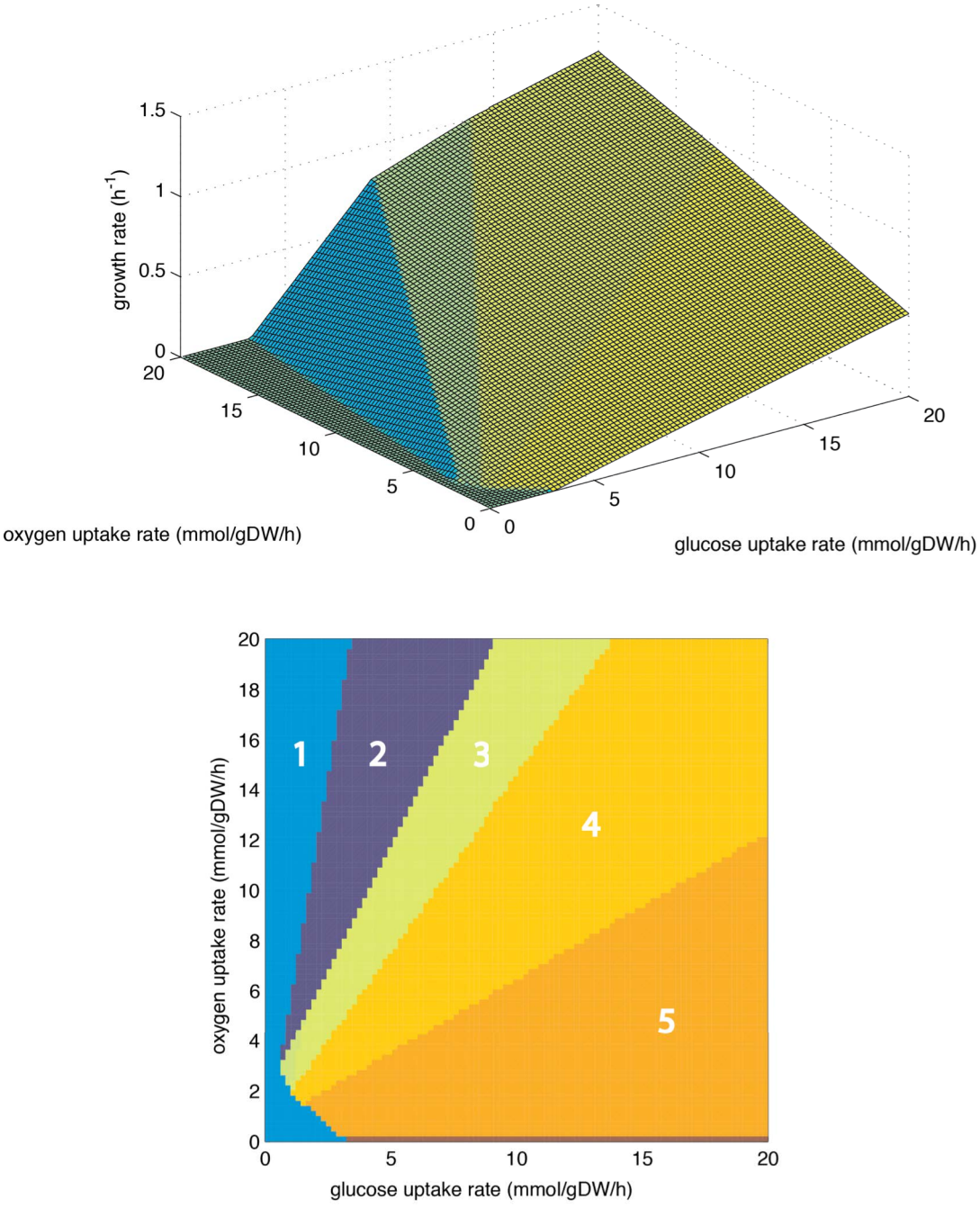

Figure 2.8 Phenotypic phase planes for growth with varying glucose and oxygen uptake rates. In phase 1, no growth is possible. In phase 2, growth is limited by excess oxygen. In phase 3, acetate is secreted; in phase 4, acetate and formate are secreted; and in phase 5, acetate, formate, and ethanol are secreted.

growth. There is not enough oxygen to fully oxidize glucose, so various compounds are produced by fermentation.

2.3.6 Gene knockout analysis

Just as growth in different environments can be simulated with FBA, gene knockouts can also be simulated by changing reaction bounds. To simulate the knockout of any gene, its associated reaction or reactions can simply be constrained to not carry flux. By setting both the upper and lower bounds of a reaction to 0 mmol/gDW/h, a reaction is essentially knocked out, and is restricted from carrying flux. The *E. coli* core model, like many other constraint-based models, contains a list of GPRs, a list of Boolean rules that dictate which genes are connected with each reaction in the model. When a reaction is catalyzed by isozymes (two different enzymes that catalyze the same reaction), the associated GPR contains an "or" rule, where either of two or more genes may be knocked out but the reaction will not be constrained. For example, the GPR for phosphofructokinase (*PFK*) is "b1723 (*pfkB*) or b3916 (*pfkA*)," so according to this Boolean rule, both *pfkB* and *pfkA* must be knocked out to restrict this reaction. When a reaction is catalyzed by a protein with multiple essential subunits, the GPR contains an "and" rule, and if any of the genes are knocked out the reaction will be constrained to 0 flux. Succinyl-CoA synthetase (*SUCOAS*), for example, has the GPR "b0728 (*sucC*) and b0729 (*sucD*)," so knocking out either of these genes will restrict this reaction. Some reactions are catalyzed by a single gene product, while others may be associated with ten or more genes in complex associations.

The COBRA Toolbox contains a function called **deleteModelGenes** that uses the GPRs to constrain the correct reactions. Then FBA may be used to predict the model phenotype with gene knockouts. For example, growth can be predicted for *E. coli* growing aerobically on glucose with the gene b1852 (*zwf*), corresponding to the reaction glucose-6-phospahte dehydrogenase (*G6PDH2r*), knocked out. The FBA predicted growth rate of this strain is 1.6329 h^{-1}, which is slightly lower than the growth rate of 1.6531 h^{-1} for wild-type *E. coli* because the cell can no longer use the oxidative branch of the pentose phosphate pathway to generate NADPH. Using FBA to predict the phenotypes of gene knockouts is especially useful in predicting essential genes. When the gene b2779 (*eno*), corresponding to the enolase reaction (*ENO*), is knocked out, the resulting growth rate on glucose is 0 h^{-1}. Growth is no longer possible because there is no way to convert glucose into TCA cycle intermediates without this glycolysis reaction, so this gene is predicted to be essential. Because FBA can compute phenotypes very quickly, it is feasible to use it for large-scale computational screens for gene essentiality, including screens for two or more simultaneous knockouts. **Figure 2.9** shows the results of a double knockout screen, in which every pairwise combination of the 136 genes in the *E. coli* core model were knocked out. The code to produce this figure is:

```
[grRatio,grRateKO,grRateWT] = doubleGeneDeletion(model);
imagesc(grRatio);
xlabel('gene knockout 1');
ylabel('gene knockout 2');
```

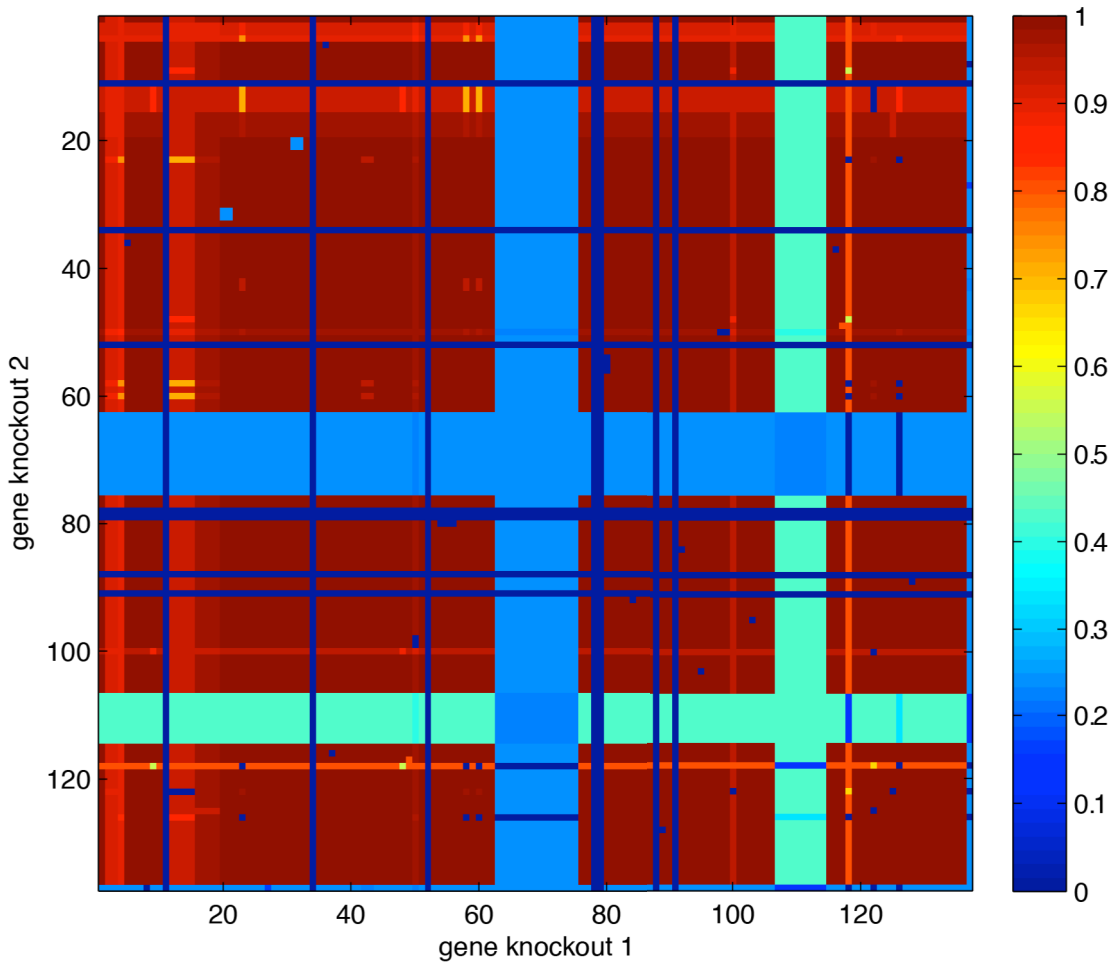

Figure 2.9 Gene knockout screen on glucose under aerobic conditions. Each of the 136 genes in the core *E. coli* model were knocked out in pairs, and the resulting relative growth rates were plotted. In this figure, genes are ordered by their b number. Some genes are always essential, and result in a growth rate of 0 when knocked out no matter which other gene is also knocked out. Other genes form synthetic lethal pairs, where knocking out only one of the genes has no effect on growth rate, but knocking both out is lethal. Growth rates in this figure are relative to wild-type.

2.3.7 Gene essentiality for biomass precursors

Here, the gene knockout analysis will be repeated, but instead of optimizing for the biomass production, we will optimize for the synthesis of all biomass precursors individually. Therefore, we have to add a demand reaction for each biomass precursor to the model and perform a gene deletion study for each demand reaction. First, the

components of the biomass reaction can be identified and demand reactions can be added by using the **addDemandReaction** function:

```
[biomassComponents,biomassFraction] = printBiomass(model,13);
[modelBiomass,rxnNames] =
addDemandReaction(model,biomassComponents);
```

Next, gene knockout screens can be performed with each of these demand reactions set as the objective, one at a time:

```
for i = 1:length(rxnNames)
        modelBiomass = changeObjective(modelBiomass,rxnNames{i});
        [grRatio,grRateKO,grRateWT,hasEffect,delRxns,fluxSolution] =
        singleGeneDeletion(modelBiomass);
        biomassPrecursorGeneEss(:,i) = grRateKO;
        biomassPrecursorGeneEssRatio(:,i) = grRatio;
end
```

The resulting matrix **biomassPrecursorGeneEssRatio** is plotted with the **imagesc** function in **Figure 2.10**, and indicates which biomass precursors become blocked by certain gene knockouts. Some precursors, such as atp[c], cannot be produced by any of the gene knockout strains because of the demand reactions that consume them. The **addDemandReacton** function produces a demand reaction that does not regenerate ADP when ATP is consumed or NAD^+ when NADH is consumed, so these reactions cannot carry flux at steady state.

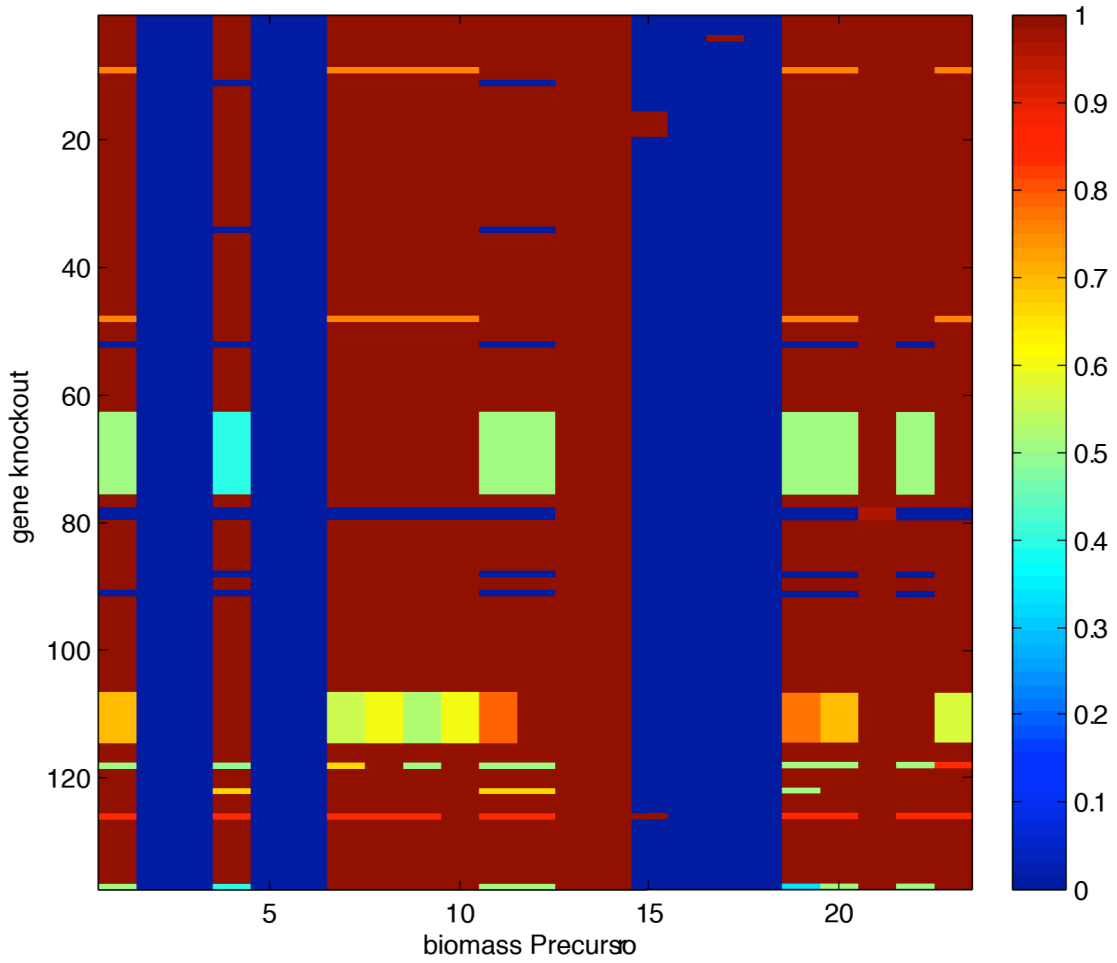

Figure 2.10 Gene essentiality for biomass precursor synthesis. Heat map shows the relative biomass precursor synthesis rate of mutant compared to wild-type. The 23 biomass precursors are 3pg, accoa, adp, akg, atp, coa, e4p, f6p, g3p, g6p, gln-L, glu-L, h2o, h, nad, nadh, nadp, nadph, oaa, pep, pi, pyr, r5p.

Acknowledgements

Chapter 2 is, in part, adapted from a chapter that appeared in EcoSal – *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (www.ecosal.org), Chapter 10.2.1, ASM Press, Washington D.C., February 18, 2010. The dissertation author was the primary author of this chapter, which was coauthored by Ronan M.T. Fleming and Bernhard Ø. Palsson.

Chapter 2 is also, in part, adapted from a paper that appeared in Nature Biotechnology, Volume 28, Number 3, Pages 245-8, March 2010. The dissertation author was the primary author of this paper, which was coauthored by Ines Thiele and Bernhard Ø. Palsson.

We would like to thank Byung-Kwan Cho, Neema Jamshidi, Nathan Lewis, and Karsten Zengler for their helpful comments and insights.

References

1. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**(121).

2. Stolz, B., et al., *The mannose transporter of Escherichia coli. Structure and function of the IIABMan subunit.* J Biol Chem, 1993. **268**(36): p. 27094-9.

3. Reidl, J. and W. Boos, *The malX malY operon of Escherichia coli encodes a novel enzyme II of the phosphotransferase system recognizing glucose and maltose and an enzyme abolishing the endogenous induction of the maltose system.* J Bacteriol, 1991. **173**(15): p. 4862-76.

4. Eberstadt, M., et al., *Solution structure of the IIB domain of the glucose transporter of Escherichia coli.* Biochemistry, 1996. **35**(35): p. 11286-92.

5. Froman, B.E., R.C. Tait, and L.D. Gottlieb, *Isolation and characterization of the phosphoglucose isomerase gene from Escherichia coli.* Mol Gen Genet, 1989. **217**(1): p. 126-31.

6. Daldal, F., *Nucleotide sequence of gene pfkB encoding the minor phosphofructokinase of Escherichia coli K-12.* Gene, 1984. **28**(3): p. 337-42.

7. Rypniewski, W.R. and P.R. Evans, *Crystal structure of unliganded phosphofructokinase from Escherichia coli.* J Mol Biol, 1989. **207**(4): p. 805-21.

8. Hines, J.K., H.J. Fromm, and R.B. Honzatko, *Novel allosteric activation site in Escherichia coli fructose-1,6-bisphosphatase.* J Biol Chem, 2006. **281**(27): p. 18386-93.

9. Alefounder, P.R., et al., *Cloning, sequence analysis and over-expression of the gene for the class II fructose 1,6-bisphosphate aldolase of Escherichia coli.* Biochem J, 1989. **257**(2): p. 529-34.

10. Baldwin, S.A. and R.N. Perham, *Novel kinetic and structural properties of the class-I D-fructose 1,6-bisphosphate aldolase from Escherichia coli (Crookes' strain).* Biochem J, 1978. **169**(3): p. 643-52.

11. Thomson, G.J., et al., *The dhnA gene of Escherichia coli encodes a class I fructose bisphosphate aldolase.* Biochem J, 1998. **331 (Pt 2)**: p. 437-45.

12. Pichersky, E., L.D. Gottlieb, and J.F. Hess, *Nucleotide sequence of the triose phosphate isomerase gene of Escherichia coli.* Mol Gen Genet, 1984. **195**(1-2): p. 314-20.

13. Branlant, G. and C. Branlant, *Nucleotide sequence of the Escherichia coli gap gene. Different evolutionary behavior of the NAD+-binding domain and of the catalytic domain of D-glyceraldehyde-3-phosphate dehydrogenase.* Eur J Biochem, 1985. **150**(1): p. 61-6.

14. Nellemann, L.J., et al., *Cloning and characterization of the Escherichia coli phosphoglycerate kinase (pgk) gene.* Gene, 1989. **77**(1): p. 185-91.

15. Fraser, H.I., M. Kvaratskhelia, and M.F. White, *The two analogous phosphoglycerate mutases of Escherichia coli.* FEBS Lett, 1999. **455**(3): p. 344-8.

16. Kuhnel, K. and B.F. Luisi, *Crystal structure of the Escherichia coli RNA degradosome component enolase.* J Mol Biol, 2001. **313**(3): p. 583-92.

17. Garrido-Pertierra, A. and R.A. Cooper, *Evidence for two distinct pyruvate kinase genes in Escherichia coli K-12.* FEBS Lett, 1983. **162**(2): p. 420-2.

18. Muirhead, H., *Isoenzymes of pyruvate kinase.* Biochem Soc Trans, 1990. **18**(2): p. 193-6.

19. Josephson, B.L. and D.G. Fraenkel, *Transketolase mutants of Escherichia coli.* J Bacteriol, 1969. **100**(3): p. 1289-95.

20. Josephson, B.L. and D.G. Fraenkel, *Sugar metabolism in transketolase mutants of Escherichia coli.* J Bacteriol, 1974. **118**(3): p. 1082-9.

21. Peyru, G. and D.G. Fraenkel, *Genetic mapping of loci for glucose-6-phosphate dehydrogenase, gluconate-6-phosphate dehydrogenase, and gluconate-6-phosphate dehydrase in Escherichia coli.* J Bacteriol, 1968. **95**(4): p. 1272-8.

22. Thomason, L.C., et al., *Identification of the Escherichia coli K-12 ybhE gene as pgl, encoding 6-phosphogluconolactonase.* J Bacteriol, 2004. **186**(24): p. 8248-53.

23. Veronese, F.M., E. Boccu, and A. Fontana, *Isolation and properties of 6-phosphogluconate dehydrogenase from Escherichia coli. Some comparisons with the thermophilic enzyme from Bacillus stearothermophilus.* Biochemistry, 1976. **15**(18): p. 4026-33.

24. Essenberg, M.K. and R.A. Cooper, *Two ribose-5-phosphate isomerases from Escherichia coli K12: partial characterisation of the enzymes and consideration of their possible physiological roles.* Eur J Biochem, 1975. **55**(2): p. 323-32.

25. Csonka, L.N. and D.G. Fraenkel, *Pathways of NADPH Formation in Escherichia coli.* J. Biological Chemistry, 1977. **252**(10): p. 3382-3391.

26. Melendez-Hevia, E. and A. Isidoro, *The game of the pentose phosphate cycle.* Journal of Theoretical Biology, 1985. **117**(2): p. 251-63.

27. Iida, A., S. Teshiba, and K. Mizobuchi, *Identification and characterization of the tktB gene encoding a second transketolase in Escherichia coli K-12.* J Bacteriol, 1993. **175**(17): p. 5375-83.

28. Sprenger, G.A., et al., *Transketolase A of Escherichia coli K12. Purification and properties of the enzyme from recombinant strains.* Eur J Biochem, 1995. **230**(2): p. 525-32.

29. Lyngstadaas, A., G.A. Sprenger, and E. Boye, *Impaired growth of an Escherichia coli rpe mutant lacking ribulose-5- phosphate epimerase activity.* Biochim Biophys Acta, 1998. **1381**(3): p. 319-30.

30. Reed, L.J., et al., *Reconstitution of the Escherichia coli pyruvate dehydrogenase complex.* Proc Natl Acad Sci U S A, 1975. **72**(8): p. 3068-72.

31. Nguyen, N.T., et al., *Comparative analysis of folding and substrate binding sites between regulated hexameric type II citrate synthases and unregulated dimeric type I enzymes.* Biochemistry, 2001. **40**(44): p. 13177-87.

32. Brock, M., et al., *Oxidation of propionate to pyruvate in Escherichia coli. Involvement of methylcitrate dehydratase and aconitase.* Eur J Biochem, 2002. **269**(24): p. 6184-94.

33. Prodromou, C., M.J. Haynes, and J.R. Guest, *The aconitase of Escherichia coli: purification of the enzyme and molecular cloning and map location of the gene (acn).* J Gen Microbiol, 1991. **137**(11): p. 2505-15.

34. Burke, W.F., R.A. Johanson, and H.C. Reeves, *NADP+-specific isocitrate dehydrogenase of Escherichia coli. II. Subunit structure.* Biochim Biophys Acta, 1974. **351**(2): p. 333-40.

35. Perham, R.N. and L.C. Packman, *2-Oxo acid dehydrogenase multienzyme complexes: domains, dynamics, and design.* Ann N Y Acad Sci, 1989. **573**: p. 1-20.

36. Bridger, W.A., et al., *The subunits of succinyl-coenzyme A synthetase--function and assembly.* Biochem Soc Symp, 1987. **54**: p. 103-11.

37. Yankovskaya, V., et al., *Architecture of succinate dehydrogenase and reactive oxygen species generation.* Science, 2003. **299**(5607): p. 700-4.

38. Condon, C., et al., *The succinate dehydrogenase of Escherichia coli. Immunochemical resolution and biophysical characterization of a 4-subunit enzyme complex.* J Biol Chem, 1985. **260**(16): p. 9427-34.

39. Bell, P.J., et al., *Nucleotide sequence of the FNR-regulated fumarase gene (fumB) of Escherichia coli K-12.* J Bacteriol, 1989. **171**(6): p. 3494-503.

40. Flint, D.H., *Initial kinetic and mechanistic characterization of Escherichia coli fumarase A.* Arch Biochem Biophys, 1994. **311**(2): p. 509-16.

41. Woods, S.A., S.D. Schwartzbach, and J.R. Guest, *Two biochemically distinct classes of fumarase in Escherichia coli.* Biochim Biophys Acta, 1988. **954**(1): p. 14-26.

42. Sutherland, P. and L. McAlister-Henn, *Isolation and expression of the Escherichia coli gene encoding malate dehydrogenase.* J Bacteriol, 1985. **163**(3): p. 1074-9.

43. Cole, S.T., et al., *Molecular biology, biochemistry and bioenergetics of fumarate reductase, a complex membrane-bound iron-sulfur flavoenzyme of Escherichia coli.* Biochim Biophys Acta, 1985. **811**(4): p. 381-403.

44. Iverson, T.M., et al., *Structure of the Escherichia coli fumarate reductase respiratory complex.* Science, 1999. **284**(5422): p. 1961-6.

45. Hoyt, J.C., et al., *Escherichia coli isocitrate lyase: properties and comparisons.* Biochim Biophys Acta, 1988. **966**(1): p. 30-5.

46. Molina, I., et al., *Molecular characterization of Escherichia coli malate synthase G. Differentiation with the malate synthase A isoenzyme.* Eur J Biochem, 1994. **224**(2): p. 541-8.

47. Narindrasorasak, S. and W.A. Bridger, *Phosphoenolypyruvate synthetase of Escherichia coli: molecular weight, subunit composition, and identification of phosphohistidine in phosphoenzyme intermediate.* J Biol Chem, 1977. **252**(10): p. 3121-7.

48. Cooper, R.A. and H.L. Kornberg, *Net formation of phosphoenolpyruvate from pyruvate by Escherichia coli.* Biochim Biophys Acta, 1965. **104**(2): p. 618-20.

49. Hansen, E.J. and E. Juni, *Two routes for synthesis of phosphoenolpyruvate from C4-dicarboxylic acids in Escherichia coli.* Biochem Biophys Res Commun, 1974. **59**(4): p. 1204-10.

50. Hansen, E.J. and E. Juni, *Isolation of mutants of Escherichia coli lacking NAD- and NADP-linked malic.* Biochem Biophys Res Commun, 1975. **65**(2): p. 559-66.

51. Iwakura, M., et al., *Studies on regulatory functions of malic enzymes. VI. Purification and molecular properties of NADP-linked malic enzyme from Escherichia coli W.* J Biochem, 1979. **85**(5): p. 1355-65.

52. Mahajan, S.K., et al., *Physical analysis of spontaneous and mutagen-induced mutants of Escherichia coli K-12 expressing DNA exonuclease VIII activity.* Genetics, 1990. **125**(2): p. 261-73.

53. Niersbach, M., et al., *Cloning and nucleotide sequence of the Escherichia coli K-12 ppsA gene, encoding PEP synthase.* Mol Gen Genet, 1992. **231**(2): p. 332-6.

54. Delbaere, L.T., et al., *Structure/function studies of phosphoryl transfer by phosphoenolpyruvate carboxykinase.* Biochim Biophys Acta, 2004. **1697**(1-2): p. 271-8.

55. Kai, Y., H. Matsumura, and K. Izui, *Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms.* Arch Biochem Biophys, 2003. **414**(2): p. 170-9.

56. Kornberg, H.L., *The coordination of metabolic routes*, in *Function and Structure in Microorganisms: Fifteenth Symposium of the Society for General Microbiology*1965, University Press: London.

57. Kornberg, H.L., *Anapleroric sequences and their role in metabolism.* Essays Biochem, 1966. **2**: p. 1-31.

58. Wilks, J.C. and J.L. Slonczewski, *pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry.* J Bacteriol, 2007. **189**(15): p. 5601-7.

59. Unden, G. and P. Dunnwald, *Module 3.2.2, the aerobic and anaerobic respiratory chain of Escherichia coli and Salmonella enterica: enzymes and energetics*, in *EcoSal - Escherichia coli and Salmonella: cellular and molecular biology*2008, ASM Press.

60. Schneider, D., et al., *Assembly of the Escherichia coli NADH:ubiquinone oxidoreductase (complex I).* Biochim Biophys Acta, 2008. **1777**(7-8): p. 735-9.

61. Spehr, V., et al., *Overexpression of the Escherichia coli nuo-operon and isolation of the overproduced NADH:ubiquinone oxidoreductase (complex I).* Biochemistry, 1999. **38**(49): p. 16261-7.

62. Kobayashi, K., S. Tagawa, and T. Mogi, *Electron transfer process in cytochrome bd-type ubiquinol oxidase from Escherichia coli revealed by pulse radiolysis.* Biochemistry, 1999. **38**(18): p. 5913-7.

63. Cain, B.D. and R.D. Simoni, *Proton translocation by the F1F0ATPase of Escherichia coli. Mutagenic analysis of the a subunit.* J Biol Chem, 1989. **264**(6): p. 3292-300.

64. Kasimoglu, E., et al., *Transcriptional regulation of the proton-translocating ATPase (atpIBEFHAGDC) operon of Escherichia coli: control by cell growth rate.* J Bacteriol, 1996. **178**(19): p. 5563-7.

65. Berry, M.B., et al., *Crystal structure of ADP/AMP complex of Escherichia coli adenylate kinase.* Proteins, 2006. **62**(2): p. 555-6.

66. Brune, M., R. Schumann, and F. Wittinghofer, *Cloning and sequencing of the adenylate kinase gene (adk) of Escherichia coli.* Nucleic Acids Res, 1985. **13**(19): p. 7139-51.

67. Bizouarn, T., et al., *Proton translocating nicotinamide nucleotide transhydrogenase from E. coli. Mechanism of action deduced from its structural and catalytic properties.* Biochim Biophys Acta, 2000. **1457**(3): p. 211-28.

68. Sauer, U., et al., *The soluble and membrane-bound transhydrogenases UdhA and PntAB have divergent functions in NADPH metabolism of Escherichia coli.* J Biol Chem, 2004. **279**(8): p. 6613-9.

69. Clark, D.P., *The fermentation pathways of Escherichia coli.* FEMS Microbiol Rev, 1989. **5**(3): p. 223-34.

70. Dym, O., et al., *The crystal structure of D-lactate dehydrogenase, a peripheral membrane respiratory enzyme.* Proc Natl Acad Sci U S A, 2000. **97**(17): p. 9413-8.

71. Jiang, G.R., S. Nikolova, and D.P. Clark, *Regulation of the ldhA gene, encoding the fermentative lactate dehydrogenase of Escherichia coli.* Microbiology, 2001. **147**(Pt 9): p. 2437-46.

72. Knappe, J. and G. Sawers, *A radical-chemical route to acetyl-CoA: the anaerobically induced pyruvate formate-lyase system of Escherichia coli.* FEMS Microbiol Rev, 1990. **6**(4): p. 383-98.

73. Sawers, G. and G. Watson, *A glycyl radical solution: oxygen-dependent interconversion of pyruvate formate-lyase.* Mol Microbiol, 1998. **29**(4): p. 945-54.

74. Suzuki, T., *Phosphotransacetylase of Escherichia coli B, activation by pyruvate and inhibition by NADH and certain nucleotides.* Biochim Biophys Acta, 1969. **191**(3): p. 559-69.

75. Skarstedt, M.T. and E. Silverstein, *Escherichia coli acetate kinase mechanism studied by net initial rate, equilibrium, and independent isotopic exchange kinetics.* J Biol Chem, 1976. **251**(21): p. 6775-83.

76. Ferrandez, A., J.L. Garcia, and E. Diaz, *Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of Escherichia coli K-12.* J Bacteriol, 1997. **179**(8): p. 2573-81.

77. Kessler, D., W. Herth, and J. Knappe, *Ultrastructure and pyruvate formate-lyase radical quenching property of the multienzymic AdhE protein of Escherichia coli.* J Biol Chem, 1992. **267**(25): p. 18073-9.

78. Yuan, J., et al., *Kinetic flux profiling of nitrogen assimilation in Escherichia coli.* Nat Chem Biol, 2006. **2**(10): p. 529-30.

79. Wu, L.F. and M.A. Mandrand-Berthelot, *A family of homologous substrate-binding proteins with a broad range of substrate specificity and dissimilar biological functions.* Biochimie, 1995. **77**(9): p. 744-50.

80. Wallace, B., et al., *Cloning and sequencing of a gene encoding a glutamate and aspartate carrier of Escherichia coli K-12.* J Bacteriol, 1990. **172**(6): p. 3214-20.

81. Rhee, S.G., et al., *Catalytic cycle of the biosynthetic reaction catalyzed by adenylylated glutamine synthetase from Escherichia coli.* J Biol Chem, 1982. **257**(1): p. 289-97.

82. Neidhardt, F.C. and H.E. Umbarger, *Chemical Composition of Escherichia coli*, in *Escherichia coli and Salmonella : cellular and molecular biology*, F.C. Neidhardt, Editor 1996, ASM Press: Washington, D.C. p. 13-16.

83. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

84. Bledig, S.A., T.M. Ramseier, and M.H. Saier, Jr., *Frur mediates catabolite activation of pyruvate kinase (pykF) gene expression in Escherichia coli.* J Bacteriol, 1996. **178**(1): p. 280-3.

85. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox.* Nat. Protocols, 2007. **2**(3): p. 727-738.

86. Varma, A. and B.Ø. Palsson, *Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors.* Journal of Theoretical Biology, 1993. **165**(4): p. 477-502.

87. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.

88. Edwards, J.S. and B.Ø. Palsson, *Robustness analysis of the Escherichia coli metabolic network.* Biotechnology Progress, 2000. **16**(6): p. 927-39.

89. Edwards, J.S., R. Ramakrishna, and B.Ø. Palsson, *Characterizing the metabolic phenotype: a phenotype phase plane analysis.* Biotechnol Bioeng, 2002. **77**(1): p. 27-36.

90. Bell, S.L. and B.Ø. Palsson, *Phenotype phase plane analysis using interior point methods.* Computers & Chemical Engineering, 2005. **29**(3): p. 481-486.

Chapter 3: Updating the genome-scale metabolic network reconstruction of *Escherichia coli*, *i*JO1366

The initial genome-scale reconstruction of the metabolic network of *Escherichia coli* K-12 MG1655 was assembled in 2000. It has been updated and periodically released since then based on new and curated genomic and biochemical knowledge. An update has now been built, named *i*JO1366, which accounts for 1366 genes, 2251 metabolic reactions, and 1136 unique metabolites. *i*JO1366 was converted to a computational model and can be used to predict growth rates and metabolic flux distributions under realistic conditions. It was compared to its predecessor and to experimental datasets to confirm that it continues to make accurate phenotypic predictions of growth on different substrates and for gene knockout strains. Like its predecessors, the *i*JO1366 reconstruction is expected to be widely deployed for studying the systems biology of *E. coli* and for metabolic engineering applications.

3.1 Introduction

A common denominator for systems biology studies of a target organism is a high-quality genome-scale metabolic network reconstruction. A network reconstruction represents a biochemically, genetically, and genomically (BiGG) structured

knowledgebase that contains detailed information about an organism in a structured format [1]. Metabolic network reconstructions contain information such as exact stoichiometry of metabolic reactions, chemical formulas and charges of metabolites, and the associations between genes, proteins, and reactions. These reconstructions form a basis for the formulation of mechanistic, and thus computable, genome-scale genotype-phenotype relationships [2].

The most detailed and complete metabolic reconstruction of any organism to date is for the common laboratory strain *Escherichia coli* K-12 MG1655. The first genome-scale reconstruction of *E. coli* was *i*JE660 [3]. This network was constructed through extensive searches of literature and databases to ensure correct stoichiometry and cofactor usage, and was the most extensive metabolic network reconstruction in existence at that time. An updated version of this reconstruction, *i*JR904 [4], had an expanded scope, including pathways for the consumption of alternate carbon sources and more specific quinone usage in the electron transport system. Hundreds of new genes and reactions were added, gene-protein-reaction associations (GPRs) were included for the first time to connect reactions with genes, and all reactions were elementally and charged balanced through the inclusion of protons. In the next update, *i*AF1260 [5], the scope of the network was expanded again, now including many reactions for the synthesis of cell wall components, and all metabolites were assigned to the cytoplasm, periplasm, or extracellular space. The thermodynamic properties of each reaction were calculated, and this was used to set lower bounds on predicted irreversible reactions. *i*AF1260 contained 2077 reactions, 1039 metabolites, and 1260 genes. A core model version of *i*AF1260,

useful for testing and debugging new constraint-based algorithms and for educational use, has also been published [6].

Here, we present an updated version of the *E. coli* metabolic network reconstruction. This new version, titled *i*JO1366, includes many newly characterized genes and reactions. Since the *i*AF1260 model was a very complete representation of the known metabolism of *E. coli*, only minor expansions in the scope of the network were made. Still, new discoveries since 2007 have made this model update necessary. Several genes were added based on the results of an experimental screen of *E. coli* knockout strains in four different media conditions. The gaps in the *i*AF1260 network were identified and characterized, and new reactions and genes were added to reduce the total number of gaps. The *i*JO1366 reconstruction can serve as a basis for metabolic network reconstructions of other *E. coli* strains and closely related organisms. *i*JO1366 is the most complete *E. coli* metabolic reconstruction to date, and like its predecessors, it will likely aid in many new discoveries [7].

3.2 Results

3.2.1 Process for updating the reconstruction and its content

An updated and expanded metabolic network reconstruction of *E. coli* K-12 MG1655 was assembled and named *i*JO1366. This new network reconstruction is an updated version of the *i*AF1260b network [8], a slightly updated version of the *i*AF1260 network published in 2007 [5]. In order to identify incorrect model predictions to improve the *E. coli* reconstruction, we experimentally determined conditional essentiality

for most of the genes in the *i*AF1260b model. A set of 1075 gene knockout strains from the Keio Collection [9] were grown on four different conditions. They were grown on glucose, L-lactate, and succinate under aerobic conditions and on glucose under anaerobic conditions. By comparing model predicted growth phenotypes to the measured cell densities, errors in the reconstruction were found and several updates were made. The *pyrI* (b4244) gene encodes the nonessential regulatory subunit of the aspartate carbamoyltransferase enzyme, but was incorrectly assigned in *i*AF1260 as being essential for catalytic activity in the GPR. This gene has been changed to a nonessential component of the aspartate carbamoyltransferase GPR in *i*JO1366. For the case of *epd* (b2927), evidence for an isozyme, *gapA* (b1779), was found in the literature [10], and the GPR has been corrected in *i*JO1366. Next, literature and database searches were performed to add newly characterized genes and reactions since 2007. The EcoCyc [11] and KEGG [12] databases were used extensively for this purpose.

After this first round of updates, the reconstruction contained 1274 genes. The network gaps in this version of the reconstruction were then investigated using a modified version of the GapFind algorithm [13]. This algorithm was modified to find root no-consumption gaps as well as their upstream blocked metabolites, in addition to the no-production gaps the original algorithm could identify (see Section **1.5 Gap-filling of metabolic networks**). All orphan reactions in the reconstruction were also identified from the model GPRs. A total of 55 root no-production gaps, 58 downstream blocked metabolites, 67 root no-consumption gaps, 61 upstream blocked metabolites, and 145 orphan reactions were identified. These gaps and orphans were then manually curated one at a time. Gaps that were found to be due to the production of non-metabolic entities,

such as charged tRNAs or metal ions, were classified as “scope gaps.” The biological functions of all metabolites blocked by scope gaps are known, but are not further connected in the *E. coli* metabolic reconstruction as they are outside the defined scope of the reconstruction. Gaps involving metabolic compounds were classified as "knowledge gaps." These gaps exist because of our incomplete knowledge of the metabolism of *E. coli*. Targeted literature and database searches were performed for each knowledge gap to try to identify any known metabolic reactions missing from the reconstruction. Newly published metabolic information continued to be added to the reconstruction during this gap-filling process, and the reconstruction was ultimately updated to *i*JO1366. New content in this reconstruction eliminated 17 root gaps and 27 downstream or upstream blocked metabolites.

*i*JO1366 represents a significant expansion of the *E. coli* reconstruction, as it contains 1366 genes, 2251 metabolic reactions, and 1136 unique metabolites. A summary of the content of *i*JO1366 and its predecessor, *i*AF1260, is presented in **Table 3.1**. Like *i*AF1260, *i*JO1366 contains a wide range of metabolic functions (**Figure 3.1**). Also like its predecessor, *i*JO1366 accounts for three specific cellular compartments: the cytoplasm, periplasm, and extracellular space. In total, 107 new genes were added to the reconstruction, while one gene, *prpE* (b0335), was removed. The previously *prpE* associated acetate-CoA ligase reaction is now associated with *acs* (b4069) [14]. A total of 254 new reactions were added, including exchange reactions for 25 extracellular metabolites, while 59 reactions from *i*AF1260 were removed. Most of the removed reactions were replaced by similar reactions with slightly different cofactor or primary substrate usage. Finally, 150 new metabolites were added while ten metabolites were

Table 3.1 Properties of *i*JO1366 and *i*AF1260

	***i*JO1366 (this study)**	***i*AF1260 (Feist et al.)**
Included Genes	1366 (32%)[d]	1260 (29%)
Experimentally-based function	1328 (97%)	1227 (97%)
Computationally-predicted function	38 (3%)	33 (3%)
Unique Functional Proteins	1254	1148
Multigene complexes	185	167
Genes involved in complexes	483	415
Instances of isozymes[a]	380	346
Reactions	2251	2077
Metabolic Reactions	1473	1387
Unique metabolic reactions[b]	1424	1339
Cytoplasmic	1272	1187
Periplasmic	193	192
Extracellular	8	8
Transport Reactions	778	690
Cytoplasm to periplasm	447	390
Periplasm to extracellular	329	298
Cytoplasm to extracellular	2	2
Gene-protein-reaction associations		
Gene associated (metabolic/transport)	1382/706	1294/625
Spontaneous/diffusion reactions[c]	21/14	16/9
Total (gene associated and no association needed)	1403/720 (94%)	1310/634 (94%)
No gene association (metabolic/transport)	70/58 (6%)	77/56 (6%)
Exchange reactions	330	304
Metabolites		
Unique metabolites	1136	1039
Cytoplasmic	1039	951
Periplasm	442	418
Extracellular	324	299

a. Tabulated on a reaction basis, not including outer membrane nonspecific 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 gene coverage based on 4325 total ORFs in *Escherichia coli* (Annotation U00096.2, downloaded from ecogene.org); 2851 of these ORFs have been experimentally verified.

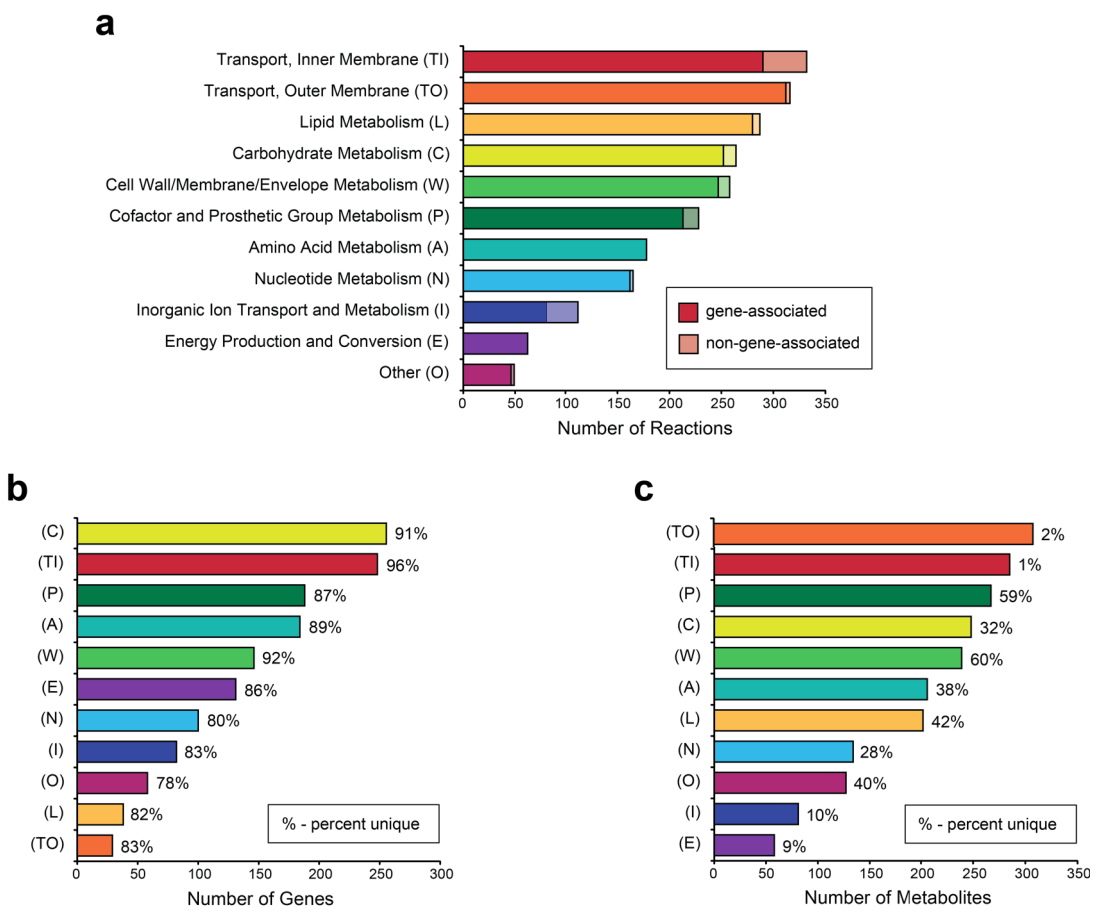

Figure 3.1 Properties of *i*JO1366. (**a**) The number of reactions in each of eleven functional categories. Non-gene-associated (orphan) reactions are indicated by the lighter portion at the far right of each bar. (**b**) The number of genes with associated reactions in each category. The number of genes unique to each category (i.e., associated only with genes in one category) is given as a percentage. (**c**) The number of unique metabolites that participate in at least one reaction in each category, with the number of metabolites unique to each category indicated.

removed when their associated reactions were removed from the network. Most new genes added have been characterized since *i*AF1260 was published in 2007 (**Figure 3.2 a**). The fact that some references predate previous versions of the *E. coli* reconstruction [3-5] does not necessarily mean that they were previously missed. Rather, as genes and reactions are often added on a pathway basis, complete functional pathways are typically fully elucidated over time from multiple sources. Thus, the citations in **Figure 3.2 a** are

spread out over time. The new genes mostly add new pathways and systems to the network, but a significant number of them fill gaps and orphan reactions in existing systems (**Figure 3.2 b**).

A number of newly discovered *E. coli* metabolic pathways and reactions were added to the network. For example, a pathway for the degradation of phenylacetic acid that had only been partially characterized previously [15] has now been fully characterized [16] and has been added to *i*JO1366. This pathway, encoded by the *paa* operon (b1388-98), breaks down phenylacetic acid to the central metabolic compounds acetyl-CoA and succinyl-CoA. Another newly discovered pathway, catalyzed by the enzymes of the *rut* operon (b1006-12), degrades pyrimidines [17]. Specifically, the reactions that convert uracil to 3-hydroxypropionic acid and excrete this by-product have been added to the reconstruction. PreQ1 is a precursor of the modified nucleoside queosine, and the pathway that produces this compound from 6-carboxy-5,6,7,8-tetrahydropterin was recently discovered [18].

The complete pathway for the synthesis of the essential cofactor biotin has also recently been characterized [19]. Beginning with malonyl-CoA, the same enzymes that catalyze fatty acid elongation (the *fab* genes) are used in two successive cycles to produce pimeloyl-ACP, which is then converted to biotin. It was previously believed that pimeloyl-CoA was the precursor of biotin, but its source was unknown, leaving a gap *i*AF1260 and other microbial metabolic network reconstructions. Because of this gap, it was impossible for biotin to be produced in this network, and the *i*AF1260 core biomass reaction did not include biotin even though it was known to be essential [5].

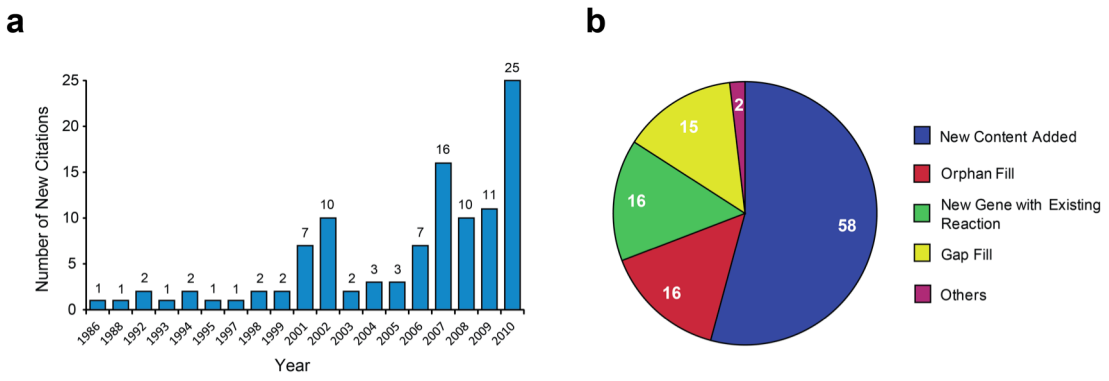

Figure 3.2 New content added to *i*JO1366. (**a**) Histogram of the years in which the function of each of the 107 new genes was first unambiguously identified. (**b**) Classification of each of the 107 new genes in *i*JO1366. "New Content Added" includes genes associated with new (non-gap-filling) pathways and systems in the model. "Orphan Fill" includes genes associated with orphan reactions from *i*AF1260. "New Gene with Existing Reaction" includes new isozymes for existing gene-associated reactions in *i*AF1260. "Gap Fill" includes genes associated with new gap-filling reactions. "Others" includes genes that are associated both with new, non-gap-filling reactions, and with a previous orphan reaction or as a new isozyme.

Other new content in the *i*JO1366 network was known prior to the construction of *i*AF1260, but had not been included in metabolic network reconstructions for various reasons. For example, the reactions for the synthesis of [2Fe-2S] and [4Fe-4S] iron-sulfur clusters by the ISC and SUF systems [20-22] have been added to *i*JO1366. These clusters can be found in many *E. coli* enzymes and typically act the in the catalysis of redox reactions. Pathways for the production of the molybdenum cofactors [23, 24] have also been added, as has the pathway for uptake and incorporation of selenium [25]. In addition to the pathways mentioned here, many other new systems have been added to the network.

3.2.2 Updating the biomass composition and growth requirements

The "core" and "wild-type" biomass reactions of *i*AF1260 have been updated in *i*JO1366. These are reactions that drain biomass precursor compounds in experimentally determined ratios to simulate growth [26, 27]. Each component of the biomass reaction has the units mmol/gDW (milli-moles per gram cell dry weight), and flux through the biomass reaction has the units h^{-1}, and is equivalent to the exponential growth rate of the organism [1]. The "wild-type" biomass reaction contains the precursors to all the typical wild-type cellular components of *E. coli*, while the "core" biomass reaction contains the precursors only to essential components. Now that the complete biotin synthesis pathway has been added to the reconstruction, biotin has been added to the biomass reactions along with the related cofactor lipoate. The [2Fe-2S] and [4Fe-4S] iron sulfur clusters have also been added, along with the molybdenum cofactors. Based on a recent study in which the metal content of *E. coli* was measured [28], the compositions of Cu, Mn, Zn, Ni, Mo, and Co in the biomass reactions have been adjusted.

Growth-associated maintenance (GAM) and non-growth-associated maintenance (NGAM) are the amounts of ATP consumed during cell growth and by non-growth associated processes such as maintenance of membrane gradients, respectively. GAM is a component of the biomass reaction, while NGAM is manifest as a lower bound on the separate ATP draining reaction "ATPM." These two parameters were recalculated for *i*JO1366 based on a new experimental dataset for *E. coli* K-12 MG1655 growing in a glucose minimal media chemostat [29]. This dataset accounts for cell lysis when determining growth rate, and thus includes a slightly higher growth rate and lower apparent maintenance requirements than in the previously used datasets [5]. For GAM

and NGAM determination, the P/O ratio of the model was constrained to 1.375, a physiologically realistic ratio [30], by enforcing a flux split through the two primary NADH dehydrogenases. GAM was determined to be 53.95 mmol ATP/gDW, while NGAM was determined to be 3.15 mmol ATP/gDW/h. It should be noted that the GAM and NGAM in a strain specific biomass reaction can vary given the experimental data set from which they were calculated. As such, these values should be based on the experimental data that most closely matches the field of use for a particular modeling application.

3.2.3 Conversion to a computational model

In order to perform computations such as the phenotypic predictions performed above, the *i*JO1366 metabolic network reconstruction was converted into a constraint-based mathematical model [31]. This model consists of an **S** matrix of size 1805 × 2583, representing 1805 different metabolites and 2583 reactions. Identical metabolites in different compartments are represented by separate rows in **S**. The complete set of reactions includes 324 exchange reactions, in which an extracellular metabolite can enter or exit the system. These exchange reactions define the boundary of the model system. The *i*JO1366 computational model also includes drain reactions for six cytoplasmic metabolites without known consuming reactions that must be drained from the system to allow simulation of steady-state cell growth. These metabolites are p-cresol, 5'-deoxyribose, aminoacetaldehyde, S-adenosyl-4-methylthio-2-oxobutanoate, (2R,4S)-2-methyl-2,3,3,4-tetrahydroxytetrahydrofuran, and oxamate. The core and wild-type biomass reactions are also included in the model. Each reaction has an upper and lower

bound on its possible flux. GPRs are represented by Boolean functions that describe which genes are required for each reaction. With the fully defined constraint-based computational model, flux balance analysis (FBA) and related methods can be used to study *E. coli* metabolism.

3.2.4 Comparison of *i*JO1366 to the Model SEED *E. coli* model

Automated tools and methods for the assembly of metabolic network reconstructions are beginning to appear, and one of the most comprehensive new tools is the Model SEED [32]. It is based on a strong annotation tool, RAST [33]. This framework combines the subsystem-based SEED genome annotations with gap-filling methods to create fully functional constraint-based metabolic models. In order to assess the completeness of *i*JO1366, it was compared to the Seed83333.1 V20.21 model of *E. coli* K-12 MG1655, a model that contains 1139 genes. Specifically, the set of genes contained in *i*JO1366 was compared to the genes in the SEED model. It was found that the SEED model contains 133 genes not contained in *i*JO1366, and that *i*JO1366 contains 362 not contained in the SEED model. The genes unique to the SEED model were investigated one at a time to determine if they have known metabolic functions and should be included in *i*JO1366. At the time of the initial comparison between these two models, four genes with characterized metabolic functions were identified in the SEED model and added to *i*JO1366: *btuE* (b1710), *yggF* (b2930), *nudF* (b3034), and *yieG* (b3714). These genes had not been found in previous model update procedures. Due to the manual literature searches performed and the large scope of the model, it is always possible that some known genes are missed, illustrating the value of quality automated

tools such as Model SEED. Of the remaining 133 genes not included in the final *i*JO1366 model, 68 were determined to have non-metabolic functions. The other 65 genes in the SEED model currently have partially or completely uncharacterized functions, and their predicted functions provide hypotheses that could lead to new metabolic discoveries and could help to fill gaps in *i*JO1366.

3.2.5 Comparison of *i*JO1366 to the EchoLocation database

The *i*JO1366 model contains metabolites in three cellular compartments: the cytoplasm, the periplasm and the extracellular space. The set of metabolites that participate in a reaction can indicate the location of the protein that catalyzes the reaction. For example, a reaction that includes only cytoplasmic metabolites must be catalyzed by a protein in the cytoplasm or attached to the inner membrane. A periplasmic or outer membrane protein cannot catalyze this reaction. To verify the accuracy of the compartment assignments of the reactions in *i*JO1366, a comparison was made to the EchoLocation database [34]. This database contains experimentally verified and computationally predicted subcellular locations for all *E. coli* K-12 proteins, sorted into 12 locations such as “cytoplasmic”, “inner membrane”, and “integral membrane protein.” The protein locations in this database were compared to the compartments of the metabolites associated with each gene in *i*JO1366 using a set of Boolean rules. These rules are listed in **Table 3.2**.

After testing all 1366 model genes, 170 were found to be inconsistent. The most common type of inconsistency was “cytoplasmic” or “periplasmic” proteins in EchoLocation that were associated with both cytoplasmic and periplasmic metabolites in

Table 3.2 Boolean rules used to compare the compartments of model reactions to EchoLocation Database protein locations.

EchoLocation Compartment	Rule
Cytoplasmic	[c], not [p] or [e]
Integral Membrane Protein	[c] or [p], not [e]
Membrane anchored	[c] or [p], not [e]
Inner membrane lipoprotein	[c] or [p], not [e]
Membrane associated	[c] or [p], not [e]
Outer Membrane β-barrel protein	[p] or [e], not [c]
Outer membrane Lipoprotein	[p], not [c] or [e]
Periplasmic	[p], not [c] or [e]
Periplasmic with N-terminal membrane anchor	[p], not [c] or [e]

*i*JO1366. There were 132 such inconsistencies. Most of these were transport reactions in *i*JO1366 with proteins that may be cytoplasmic or periplasmic subunits of multi-subunit complexes. The remaining 38 inconsistencies were investigated one at a time to determine whether *i*JO1366, EchoLocation, or both are correct. Through targeted literature searches, experimental evidence was found indicating that 12 of the locations are correct in *i*JO1366 but incorrect in EchoLocation. The remaining 26 proteins were found to be correct in both EchoLocation and *i*JO1366, and all involve multi-subunit transporters with individual proteins spanning multiple locations. After manually reconciling *i*JO1366 and EchoLocation, most genes were consistent. Interestingly, the locations that are based on experimental evidence in EchoLocation are more likely to be inconsistent with *i*JO1366 than computationally predicted locations. As most "periplasmic" proteins in EchoLocation are actually associated with cytoplasm to periplasm transport reactions in *i*JO1366, these discrepancies may simply be due to the definition of a "periplasmic" protein in EchoLocation. Still, the overall content of

*i*JO1366 is consistent with EchoLocation, indicating that the compartments of the metabolites in most reactions are correct.

3.2.6 Prediction of metabolic phenotypes

Constraint-based models such as *i*JO1366 can be used to predict the metabolic phenotypes of microbial organisms under different conditions. FBA can be used in combination with a biomass reaction to predict metabolic flux distributions, growth rates, substrate uptake rates, and product secretion rates. To demonstrate the accuracy of the *i*JO1366 model in making these phenotypic predictions, several model predictions were compared to actual experimental data collected from *E. coli*.

In the first example, the fluxes through central metabolic reactions predicted by FBA were compared to experimental metabolic flux data [35]. In the experimental dataset, *E. coli* was grown in a chemostat at a dilution rate of 0.1 h^{-1}. The technique of isotopic dynamic ^{13}C metabolic flux analysis was used to determine the fluxes through 20 metabolic reactions, including glucose uptake and acetate secretion. To match the simulation conditions with the *i*JO1366 model, the core biomass reaction was constrained to 0.1 h^{-1}, and FBA was used to predict the flux distribution with minimum glucose uptake as the objective. When normalized to the glucose uptake rate, the predicted flux distribution matches the experimentally determined fluxes very well (**Figure 3.3 a and b**). The mean difference between predicted and measured fluxes was 6.6 ± 1.6%, and they were well correlated (R^2 = 0.98). The experimental data predicts a flux split between glucose-6-phosphate isomerase (*PGI*, the first step in glycolysis) and glucose-6-phosphate dehydrogenase (*G6PDH2r*, the first step in the pentose phosphate pathway) of

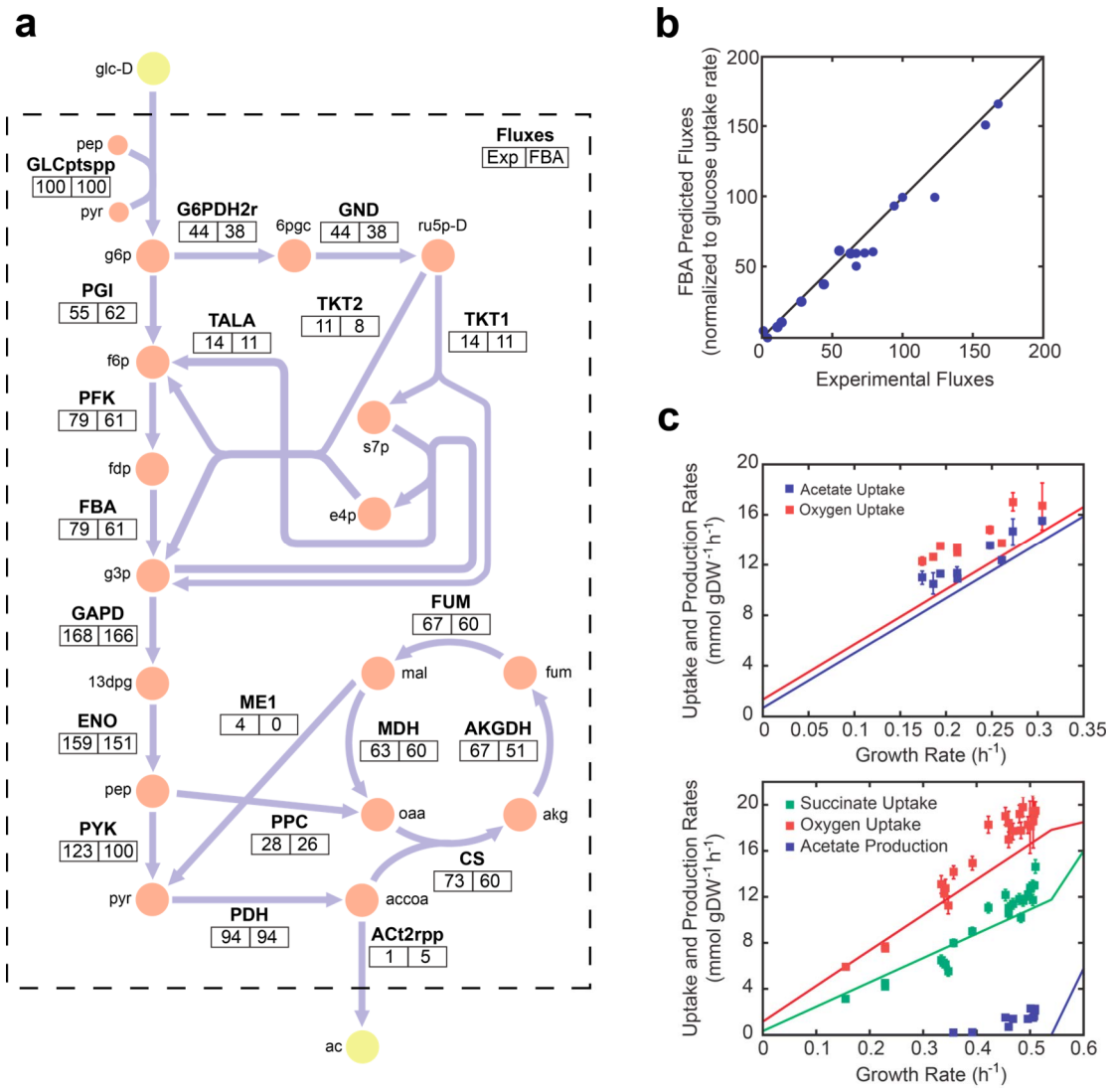

Figure 3.3 Comparison of model phenotype predictions by FBA to experimental data. (**a**) Experimental metabolic flux data for 20 central metabolic reactions was collected during growth on glucose minimal media at a growth rate of 0.1 h^{-1} [35]. Reactions and metabolites are listed by their model abbreviations. Experimental fluxes through each reaction (left side boxes under each reaction abbreviation) and model predicted fluxes (right side boxes) were calculated relative to the glucose uptake rates. (**b**) The model predicted fluxes plotted against the experimentally determined fluxes. (**c**) Model predicted substrate and oxygen uptake rates for growth on acetate (top) and succinate (bottom) compared to experimental chemostat growth data [36].

55/44, while the model predicts a flux split of 62/38. The model predicted and experimental flux distributions include similar fluxes through glycolysis and the TCA

cycle, and both have very small flux through phosphoenolpyruvate carboxylase (*PPC*). All predicted reaction fluxes match the directionality of the experimental fluxes. The model correctly predicted the minor production of acetate and that no other by-products were produced in significant quantities.

Model based growth predictions were compared to two other experimental datasets. In the first dataset, *E. coli* was grown in an acetate minimal medium chemostat [36]. The acetate and oxygen uptake rates were measured while the dilution rate was varied. In the second experimental dataset, succinate and oxygen uptake rates along with acetate secretion rates were measured in a succinate minimal media chemostat [36]. FBA was used to simulate optimal *E. coli* growth under both acetate and succinate minimal media conditions, with varying growth rates. In the simulations of these experiments, the core biomass reaction was constrained and varied while the minimum substrate uptake rates were predicted. The predicted uptake and growth rates were then compared to the measured values (**Figure 3.3 c**). Overall, the computational predictions using *i*JO1366 compared well to the experimental measurements. Just as the experimental measurements indicated, the model predicts acetate secretion for growth on succinate at higher growth rates, once oxygen becomes limiting.

3.2.7 Prediction of all growth-supporting carbon, nitrogen, phosphorus, and sulfur sources

The *i*JO1366 computational model contains exchange reactions for 324 different compounds. 285 of these compounds contain at least one carbon atom, 178 contain nitrogen, 64 contain phosphorus, and 28 contain sulfur. It is therefore possible to use

*i*JO1366 to predict the growth capabilities of *E. coli* on a very wide range of media conditions. As a demonstration of the prediction of growth capabilities, FBA was used to predict growth on every possible carbon, nitrogen, phosphorus, and sulfur source, one at a time, under aerobic conditions (**Table 3.3**). For each prediction, only one of the four element source reactions was changed, and the default sources of the other three elements were used. The default carbon, nitrogen, phosphorus, and sulfur sources are glucose, ammonium, inorganic phosphate, and inorganic sulfate, respectively. If a growth rate above zero was predicted by FBA using the core biomass reaction as the objective, then a source was designated as growth supporting.

A total of 180 of the 285 possible carbon sources were found to be growth supporting. There are several reasons why a carbon containing metabolite cannot serve as a carbon source. First, not all extracellular compounds have transport reactions that allow them to enter the cell. Some may only have efflux reactions that allow them to be excreted. Second, some compounds are not connected to the central reactions of metabolism from which all essential biomass components are constructed. For example, cob(I)alamin can be converted only to vitamin B_{12}, but not to any other biomass

Table 3.3 Growth supporting carbon, nitrogen, phosphorus, and sulfur sources.

	***i*JO1366**		***i*AF1260**	
Source	**Potential Substrates**	**Growth Supporting**	**Potential Substrates**	**Growth Supporting**
Carbon	285	180	262	174
Nitrogen	178	94	163	78
Phosphorus	64	49	63	49
Sulfur	28	11	25	11

components. Third, carbon sources must also generally serve as energy sources for *E. coli*, so a highly oxidized compound such as CO_2 cannot be growth supporting. Not all compounds can serve as nitrogen, phosphorus, and sulfur sources for similar reasons. Some compounds may serve as a source of more than one essential element, such as L-alanine, which can provide both carbon and nitrogen simultaneously. The potential growth supporting carbon, nitrogen, phosphorus, and sulfur sources were also predicted using the *i*AF1260 *E. coli* model. *i*JO1366 contains the same number of growth supporting phosphorus and sulfur sources, but has new sources for carbon and nitrogen. Thus, the scope of the environmental conditions that can be analyzed through modeling has now been increased.

3.2.8 Prediction of gene essentiality

The GPR associations of every reaction in *i*JO1366 allow this model to predict the effects of gene knockouts. We used FBA to predict the optimal growth rate of *E. coli* growing on both glucose and glycerol with all 1366 genes knocked out one at a time. These computational knockout screens were then compared to experimental screens of the entire Keio Collection (**Table 3.4**) [9, 37, 38].

There are four possible outcomes, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), when one compares computationally predicted to experimental gene essentiality. FP predictions can be made when a model contains some unrealistic capabilities, such as pathways that are normally not expressed during the particular growth conditions. Because *i*JO1366 is a metabolic network model that does not contain regulatory systems, FP predictions are possible. FN cases, on the other hand,

Table 3.4 Gene essentiality predictions on glucose and glycerol minimal media.

	Experimental	
	Essential	**Non-essential**
	Growth on glucose	
Computational		
Essential	168 (12.3%)	39 (2.8%)
Non-essential	80 (5.9%)	1079 (79.0%)
	Growth on glycerol	
Computational		
Essential	161 (11.8%)	45 (3.3%)
Non-essential	87 (6.4%)	1073 (78.5%)

indicate that some realistic content such as an essential transport or enzymatic reaction may be missing from the model. These predictions can be used to drive model-based biological discovery [37, 39] (**Chapter 4: Gap-filling of the *Escherichia coli* metabolic network for model improvement and discovery**). When compared to the experimental gene essentiality data, most of the predictions made by *i*JO1366 are correct, confirming its overall accuracy (91%). Still, there are 80 FPs and 39 FNs among the 1366 predictions for growth on glucose minimal media. Predictions of growth on glycerol minimal media achieved similar accuracy. *i*JO1366 is slightly less accurate at predicting overall gene essentiality than *i*AF1260, when compared to the same datasets. This is because the 107 new genes added to this model version are from less well-studied systems and pathways than the existing genes in *i*AF1260. Many of these new genes are associated with peripheral metabolic systems, while the well-studied central metabolic genes were already included in previous model versions. The overall accuracy of gene essentiality predictions for the 107 new genes is only 89%.

3.3 Discussion

Although the metabolism of *E. coli* has been the subject of active research for decades, the new content added to the *i*JO1366 reconstruction demonstrates that new discoveries continue to be made. In fact, most of the newly characterized genes in *i*JO1366 were actually characterized in 2010 (**Figure 3.2 a**). Although this is a small sample size, it appears that the pace of new discoveries is not slowing as more of the metabolic network is uncovered. There are still numerous uncharacterized *E. coli* K-12 MG1655 genes, many of which are predicted to have metabolic functions [40]. As more discoveries are made in the future, additional updates to the *E. coli* network reconstruction will need to be made. Future increases in the scope of the *E. coli* metabolic network reconstruction will likely include the integration of this network with reconstructions of other cellular systems. A reconstruction of the transcription and translation machinery of *E. coli* has been published [41], and the combination of this large network with metabolism is a very promising prospect. A large-scale transcriptional regulatory network has previously been integrated with the *E. coli* metabolic network [42], and new experimental studies are continuing to improve our knowledge of this network [43-45].

The experimental phenotypic screen of *E. coli* knockout strains grown in four different media conditions has helped to indicate areas in which the metabolic network is not yet fully understood. By comparing model predicted growth phenotypes to the measurements, errors in the reconstruction were found. These disagreements were investigated, and it was found that there are several possible explanations for both false negative and false positive predictions. Most false positives were found to be due to the

presence of isozymes or alternative pathways in the model. Under the actual experimental conditions, these alternative genes are likely not expressed. They may be latent pathways, which are known to exist but may require significant regulatory adjustments to be activated [46]. Gene expression data can be mapped to metabolic models, and can help to reduce false positive predictions in cases such as these [47]. False negative model predictions are most likely due to genes and pathways missing from the model, and thus represent knowledge gaps. These cases could be used to fill gaps in the network and add new content when algorithms such as SMILEY [39] are used to predict the missing reactions.

A significant number of the gaps in the *i*AF1260 network were filled during the update to *i*JO1366, and several blocked pathways were unblocked. Still, over 10% of the metabolic network remains blocked due to gaps. Many of these gaps, however, are scope gaps. The roles of these dead-end metabolites in *E. coli* are known, but are not included in *i*JO1366 because they are not considered metabolic. In order to restore connectivity to these metabolites, the metabolic network must be connected to models of other cellular systems. By joining the network of transcription and translation in *E. coli* [41] with *i*JO1366, the blocked tRNA charging reactions will become essential contributors to cell growth. On the other hand, knowledge gaps exist in the reconstruction because not all components of *E. coli* metabolism have been experimentally characterized. New discoveries must be made to fill these gaps. Fortunately, *i*JO1366 itself can serve as a tool to aid in the discovery of these missing genes and reactions. A number of constraint-based analysis methods exist that can predict the likely missing content by comparing model predictions to experimental phenotypes. The knockout strain phenotypic screen

performed in this study is one example of a dataset that can be used for gap-filling. The orphan reactions in models such as *i*JO1366 can also help to identify the functions of metabolic genes. In the original *i*JR904 study, published in 2003 [4], BLAST [48] was used to predict the likely *E. coli* genes that encode the enzymes for 56 orphan reactions. No follow up experiments were performed as part of this study, but since then, 14 of these predictions have been independently confirmed to be correct, and these genes are now included in *i*JO1366. Of the remaining predictions, about half were later confirmed to have a different metabolic function, although they may also perform the predicted function. As metabolic models become more complete and gap-filling methods become more sophisticated, the use of model-based predictions to drive biological discovery is likely to increase.

The *i*JO1366 genome-scale metabolic network reconstruction of *E. coli* is the latest update to one of the workhorse models of the microbial systems biology community. For the past decade, previous versions of this reconstruction have been used in a wide range of studies, from the discovery and characterization of new metabolic genes [39, 49] to the design of high-yield production strains for industrially valuable compounds [8, 50]. Like its predecessors, *i*JO1366 is expected to have many practical applications [7]. Some of the applications, such as prediction of growth phenotypes in different media and with gene knockouts, have been presented here. The accuracy of the model has been confirmed by comparisons to experimental data. It is also expected that this model will soon be integrated with genome-scale network reconstruction of other cellular systems such as transcriptional regulation and transcription and translation. *i*JO1366 is the most advanced and comprehensive metabolic reconstruction of any

microorganism to date, and can thus continue to serve as a basis for the metabolic reconstruction of other bacteria. Based on the success of it its predecessors, we expect that *i*JO1366 will be an important tool in microbial systems biology for years to come.

3.4 Methods

3.4.1 Metabolic network reconstruction procedure

The *i*JO1366 reconstruction was assembled by updating the *i*AF1260b *E. coli* metabolic reconstruction [8], an updated version of the *i*AF1260 reconstruction [5]. *i*AF1260b contains six additional reactions that were not in *i*AF1260 (*ALAt2rpp*, *ASPt2rpp*, *CITt3pp*, *DHORDfum*, *GLYt2rpp*, and *MALt3pp*). A 96-step procedure for metabolic network reconstruction was recently published [1], and the appropriate steps were followed when adding new genes, reactions, and metabolites to form *i*JO1366. The reconstruction was assembled using the SimPheny (Genomatica Inc., San Diego, CA) software platform. All new metabolites were checked against public databases (KEGG, PubChem) for correct structure and charge at a pH of 7.2. New reactions were mass and charge balanced and reversibility was assigned based on experimental studies, thermodynamic information, or the heuristic rules in the standard reconstruction protocol [1]. Reactions were associated with genes and functional proteins to form GPRs. The *i*JO1366 model was exported from SimPheny as an SBML file and the COBRA Toolbox [51], a Matlab (The MathWorks Inc., Natick, MA) Toolbox, was used for additional model testing. The Tomlab (Tomlab Optimization Inc., Seattle, WA) CPLEX linear programming solver was used for all optimization procedures.

The GapFind MILP algorithm [13] was encoded in the COBRA Toolbox and used to identify all blocked metabolites in the *i*AF1260 and *i*JO1366 models. This algorithm was modified from the published version. Specifically, an option was included to change the mass balance constraint $\sum_j S_{ij} v_j \geq 0$ to $\sum_j S_{ij} v_j = 0$, allowing for metabolites without consuming reactions to be identified as gaps. For each GapFind run, the lower bounds of all exchange reactions were set to -1000 mmol/gDW/h and the upper bounds of all model reactions were set to 10^9 mmol/gDW/h. The GapFind algorithm was then run twice, once with each mass balance constraint option. Root no-production and no-consumption metabolites were identified from the model **S** matrix by searching for rows containing only negative or positive coefficients, respectively. Downstream no-production and upstream no-consumption gaps were identified by removing the root gaps from the GapFind outputs. The root gaps of each downstream gap were identified through targeted computational experiments in which metabolite source reactions were added to the network to restore connectivity. Orphan reactions were identified as all reactions without associated GPRs.

The core and wild-type biomass reactions were modified from the *i*AF1260 biomass reactions. Biotin was added with a coefficient based on a published biotin concentration of 250 molecules/cell [52], while the related cofactor lipoate was added with a similar number of molecules/cell assumed. Iron-sulfur clusters were added with coefficients based on the predictions that 5% of all *E. coli* proteins contain these clusters [53], and that the majority (90%) of these clusters are of the [4Fe-4S] type. The coefficient of Fe^{2+} was decreased to account for Fe used in iron-sulfur clusters.

Molybdenum cofactors were added with coefficients based on the measurement that the inorganic ion content of *E. coli* is 0.80% Mo [28], and the fact that the majority of this Mo is in the bis-molybdopterin guanine dinucleotide form (85%). The coefficients of Cu, Mn, Zn, Ni, and Co were also adjusted based on recent *in vivo* measurements [28]. All other biomass components remain the same as in *i*AF1260. Growth-associated and non-growth-associated maintenance values were recalculated based on recent *E. coli* K-12 MG1655 chemostat data for growth on glucose minimal media [29]. The slope and intercept of this experimental data were identified by linear regression. For these calculations, the electron transport system NADH dehydrogenase reactions *NADH16pp* (*nuo*) and *NADH5* (*ndh*) were constrained to carry identical fluxes by replacing these reactions with an equivalent "flux split" reaction, in order to constrain the model to a realistic P/O ratio of 1.375 [30]. The NGAM of 3.15 mmol ATP/gDW/h was identified by FBA as the maximum amount of ATP produced at a glucose uptake rate of 0.17 mmol/gDW/h, the intercept of the experimental data. The GAM of 53.95 mmol ATP/gDW was identified by FBA as the value that would give the correct experimentally determined slope of 10.83 mmol glucose/gDW/h/ (μ) h^{-1} when using the core biomass reaction.

The *i*JO1366 model is available in SBML format at BioModels (accession: MODEL1108160000).

3.4.2 Comparison of *i*JO1366 to the Model SEED *E. coli* reconstruction

The *E. coli* K-12 MG1655 model Seed83333.1 V20.21 was downloaded from the Model SEED database [32] in SBML format. The set of 1139 genes in this model was

compared by gene ID (b-number) to the 1366 genes in *i*JO1366 to identify common genes and the unique genes in each model. The genes in the Model SEED model that were not in *i*JO1366 were then investigated one at a time. EcoCyc was used to identify gene functions, and several genes with verified metabolic functions were then added to *i*JO1366.

3.4.3 Comparison of *i*JO1366 to the EchoLocation database

Predicted and experimentally determined protein location data was obtained for all *E. coli* K-12 genes from the EchoLocation database [34]. The cellular locations of the proteins associated with the 1366 model genes from this database were then compared, one at a time, to the compartments of the metabolites in the reactions associated with each gene in *i*JO1366. For each location in the EchoLocation database, a Boolean rule was written to determine if the location is consistent with the associated model metabolites. For example, "Cytoplasmic" proteins are consistent with genes only associated with cytoplasmic metabolites, and are inconsistent with genes associated with any periplasmic or extracellular metabolites. The genes whose locations were inconsistent with model metabolites were then investigated individually, except for "Cytoplasmic" and "Periplasmic" genes with both cytoplasmic and periplasmic metabolites in the model, because there were too many of these inconsistencies to investigate manually. Literature and database information was used to determine whether the EchoLocation database, the *i*JO1366 reconstruction, or both were correct.

3.4.4 Constraint-based modeling

The *i*JO1366 model, constructed in SimPheny, was exported as an SBML file and used to perform simulations and constraint-based analyses using the COBRA Toolbox and Tomlab CPLEX linear programming solver. The constraint-based model consists of a stoichiometric matrix (**S**) with 1805 rows and 2583 columns, where 1805 is the number of distinct metabolites (in all three compartments) and 2583 is the number of reactions including exchange and biomass reactions. Each of the reactions has an upper and lower bound on the flux it can carry. Reversible reactions have an upper bound of 1000 mmol/gDW/h and a lower bound of -1000 mmol/gDW/h, making them practically unconstrained, while irreversible reactions have a lower bound of zero.

By default, the core biomass reaction is set as the objective to be maximized. Certain reactions are by default constrained to carry zero flux to avoid unrealistic behaviors. These reactions are *CAT*, *DHPTDNR*, *DHPTDNRN*, *FHL*, *SPODM*, *SPODMpp*, *SUCASPtpp*, *SUCFUMtpp*, *SUCMALtpp*, and *SUCTARTtpp*. *CAT*, *SPODM*, and *SPODMpp* are hydrogen peroxide producing and consuming reactions that can carry flux in unrealistic energy generating loops. *DHPTDNR* and *DHPTDNRN* form a closed loop that can carry an arbitrarily high flux. The succinate antiporters *SUCASPtpp*, *SUCFUMtpp*, *SUCMALtpp*, and *SUCTARTtpp* can form unrealistic flux loops with other transporters for aspartate, fumarate, malate, and tartrate. The genes encoding formate hydrogen lyase (*FHL*) are known to be active under anaerobic conditions, but this reaction is constrained to zero to avoid unrealistic aerobic hydrogen production. The NGAM constraint is imposed by a lower bound of 3.15 mmol/gDW/h on the reaction *ATPM*. The exchange reactions that allow for extracellular metabolites to pass in and out

of the system are defined such that a positive flux indicates flow out. All exchange reactions have a lower bound of zero except for glucose (-10 mmol/gDW/h), the vitamin B_{12} precursor cob(I)alamin (-0.01 mmol/gDW/h), and oxygen and all inorganic ions required by the biomass reaction (-1000 mmol/gDW/h). The default lower bound on glucose uptake is based on typical glucose uptake rates. Because only a very small amount of B_{12} is required for growth, the lower bound on cob(I)alamin uptake is arbitrary and never actually constraining in practice. The *i*JO1366 computational model also includes drain reactions for six cytoplasmic metabolites without known consuming reactions that must be drained from the system to allow simulation of steady-state cell growth. These metabolites are p-cresol, 5'-deoxyribose, aminoacetaldehyde, S-adenosyl-4-methylthio-2-oxobutanoate, (2R,4S)-2-methyl-2,3,3,4-tetrahydroxytetrahydrofuran, and oxamate.

3.4.5 Prediction of all growth-supporting carbon, nitrogen, phosphorus, and sulfur sources

The possible growth-supporting carbon, nitrogen, phosphorus, and sulfur sources of *E. coli* were identified using FBA. First, all exchange reactions for extracellular metabolites containing the four elements were identified from the metabolite formulas. Every extracellular compound containing carbon was considered a potential carbon source, for example. Next, to determine possible growth supporting carbon sources, the lower bound of the glucose exchange reaction was constrained to zero. Then the lower bound of each carbon exchange reaction was set, one at a time, to -10 mmol/gDW/h (a typical uptake rate for growth supporting substrates), and growth was maximized by FBA

using the core biomass reaction. The target substrate was considered growth supporting if the predicted growth rate was above zero. While identifying carbon sources, the default nitrogen, phosphorus, and sulfur sources were ammonium (nh4), inorganic phosphate (pi), and inorganic sulfate (so4), respectively. Prediction of growth supporting sources of these other three elements was performed in the same manner as growth on carbon, with glucose as the default carbon source.

3.4.6 Prediction of gene essentiality

To simulate the effects of gene knockouts, the *i*JO1366 model with its default constraints and core biomass reaction objective was modified to match the genotype of *E. coli* BW25113. For growth on glucose, the lower bound of the glucose exchange reaction was set to -10 mmol/gDW/h. For growth on glycerol, the lower bound of the glucose exchange reaction was set to zero while the lower bound of the glycerol exchange reaction was set to -10 mmol/gDW/h. All 1366 genes in the model were knocked out one a time and growth was simulated by FBA using the singleGeneDeletion COBRA Toolbox function. Gene knockout strains with a growth rate above zero were considered non-essential. Experimental gene essentiality data for growth on glucose [9] and glycerol [37] was then obtained and adjusted based on an updated analysis of the Keio Collection strains [38]. The newly identified essential genes were added to the lists of essential genes under both conditions, while the genes whose essentiality was identified as uncertain were not changed from their original designations.

Acknowledgements

Chapter 3 is adapted from a paper that appeared in Molecular Systems Biology, Volume 7, Article Number 535, October 11, 2011. The dissertation author was the primary author of this paper, which was coauthored by Tom M. Conrad, Jessica Na, Joshua A. Lerman, Hojung Nam, Adam M. Feist, and Bernhard Ø. Palsson.

We would like to thank Harish Nagarajan, Vasiliy Portnoy, Jennie Reed, and Ines Thiele for their helpful comments and insights.

References

1. Thiele, I. and B.Ø. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.

2. Palsson, B.Ø., *Metabolic systems biology.* FEBS Lett, 2009. **583**(24): p. 3900-4.

3. Edwards, J.S. and B.Ø. Palsson, *The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities.* Proc Natl Acad Sci U S A., 2000. **97**(10): p. 5528-5533.

4. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biology, 2003. **4**(9): p. R54.1-R54.12.

5. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**(121).

6. Orth, J.D., R.M. Fleming, and B.Ø. Palsson, *10.2.1 - Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide*, in *EcoSal - Escherichia coli and Salmonella Cellular and Molecular Biology*, P.D. Karp, Editor 2010, ASM Press: Washington D.C.

7. Feist, A.M. and B.Ø. Palsson, *The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli.* Nat Biotech, 2008. **26**(6): p. 659-667.

8. Feist, A.M., et al., *Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli.* Metab Eng, 2010. **12**(3): p. 173-86.

9. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol Syst Biol, 2006. **2**: p. 2006.0008.

10. Yang, Y., et al., *Involvement of the gapA- and epd (gapB)-encoded dehydrogenases in pyridoxal 5'-phosphate coenzyme biosynthesis in Escherichia coli K-12.* J Bacteriol, 1998. **180**(16): p. 4294-9.

11. Keseler, I.M., et al., *EcoCyc: a comprehensive view of Escherichia coli biology.* Nucleic Acids Res, 2009. **37**(Database issue): p. D464-70.

12. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010. **38**(Database issue): p. D355-60.

13. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.

14. Brock, M., et al., *Oxidation of propionate to pyruvate in Escherichia coli. Involvement of methylcitrate dehydratase and aconitase.* Eur J Biochem, 2002. **269**(24): p. 6184-94.

15. Ferrandez, A., et al., *Catabolism of phenylacetic acid in Escherichia coli. Characterization of a new aerobic hybrid pathway.* J Biol Chem, 1998. **273**(40): p. 25974-86.

16. Teufel, R., et al., *Bacterial phenylalanine and phenylacetate catabolic pathway revealed.* Proc Natl Acad Sci U S A, 2010. **107**(32): p. 14390-5.

17. Kim, K.S., et al., *The Rut pathway for pyrimidine degradation: novel chemistry and toxicity problems.* J Bacteriol, 2010. **192**(16): p. 4089-102.

18. McCarty, R.M., et al., *The deazapurine biosynthetic pathway revealed: in vitro enzymatic synthesis of PreQ(0) from guanosine 5'-triphosphate in four steps.* Biochemistry, 2009. **48**(18): p. 3847-52.

19. Lin, S., R.E. Hanson, and J.E. Cronan, *Biotin synthesis begins by hijacking the fatty acid synthetic pathway.* Nat Chem Biol, 2010. **6**(9): p. 682-8.

20. Bandyopadhyay, S., K. Chandramouli, and M.K. Johnson, *Iron-sulfur cluster biosynthesis.* Biochem Soc Trans, 2008. **36**(Pt 6): p. 1112-9.

21. Wollers, S., et al., *Iron-sulfur (Fe-S) cluster assembly: the SufBCD complex is a new type of Fe-S scaffold with a flavin redox cofactor.* J Biol Chem, 2010. **285**(30): p. 23331-41.

22. Saini, A., et al., *SufD and SufC ATPase activity are required for iron acquisition during in vivo Fe-S cluster formation on SufB.* Biochemistry, 2010. **49**(43): p. 9402-12.

23. Schwarz, G. and R.R. Mendel, *Molybdenum cofactor biosynthesis and molybdenum enzymes.* Annu Rev Plant Biol, 2006. **57**: p. 623-47.

24. Schwarz, G., R.R. Mendel, and M.W. Ribbe, *Molybdenum cofactors, enzymes and pathways.* Nature, 2009. **460**(7257): p. 839-47.

25. Turner, R.J., J.H. Weiner, and D.E. Taylor, *Selenium metabolism in Escherichia coli.* Biometals, 1998. **11**(3): p. 223-7.

26. Varma, A. and B.Ø. Palsson, *Metabolic capabilities of Escherichia coli: II. Optimal growth patterns.* Journal of Theoretical Biology, 1993. **165**(4): p. 503-522.

27. Feist, A.M. and B.Ø. Palsson, *The biomass objective function.* Curr Opin Microbiol, 2010. **13**(3): p. 344-9.

28. Cvetkovic, A., et al., *Microbial metalloproteomes are largely uncharacterized.* Nature, 2010. **466**(7307): p. 779-82.

29. Taymaz-Nikerel, H., et al., *Genome-derived minimal metabolic models for Escherichia coli MG1655 with estimated in vivo respiratory ATP stoichiometry.* Biotechnol Bioeng, 2010. **107**(2): p. 369-81.

30. Noguchi, Y., et al., *The energetic conversion competence of Escherichia coli during aerobic respiration studied by 31P NMR using a circulating fermentation system.* J Biochem (Tokyo), 2004. **136**(4): p. 509-15.

31. Orth, J.D., I. Thiele, and B.Ø. Palsson, *What is flux balance analysis?* Nat Biotechnol, 2010. **28**(3): p. 245-8.

32. Henry, C.S., et al., *High-throughput generation, optimization and analysis of genome-scale metabolic models.* Nat Biotechnol, 2010. **28**(9): p. 977-82.

33. Aziz, R.K., et al., *The RAST Server: Rapid Annotations using Subsystems Technology.* BMC Genomics, 2008. **9**(1): p. 75.

34. Horler, R.S., et al., *EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations.* Bioinformatics, 2009. **25**(2): p. 163-6.

35. Schaub, J., K. Mauch, and M. Reuss, *Metabolic flux analysis in Escherichia coli by integrating isotopic dynamic and isotopic stationary 13C labeling data.* Biotechnol Bioeng, 2008. **99**(5): p. 1170-85.

36. Edwards, J.S., R.U. Ibarra, and B.Ø. Palsson, *In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data.* Nat Biotechnol, 2001. **19**(2): p. 125-130.

37. Joyce, A.R., et al., *Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli.* J Bacteriol, 2006. **188**(23): p. 8259-8271.

38. Yamamoto, N., et al., *Update on the Keio collection of Escherichia coli single-gene deletion mutants.* Mol Syst Biol, 2009. **5**: p. 335.

39. Reed, J.L., et al., *Systems approach to refining genome annotation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17480-4.

40. Riley, M., et al., *Escherichia coli K-12: a cooperatively developed annotation snapshot-2005.* Nucleic Acids Res, 2006. **34**(1): p. 1-9.

41. Thiele, I., et al., *Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization.* PLoS Comput Biol, 2009. **5**(3): p. e1000312.

42. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.

43. Cho, B.K., et al., *Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19462-7.

44. Cho, B.K., et al., *Microbial regulatory and metabolic networks.* Curr Opin Biotechnol, 2007. **18**(4): p. 360-4.

45. Cho, B.K., E.M. Knight, and B.Ø. Palsson, *Genomewide identification of protein binding locations using chromatin immunoprecipitation coupled with microarray.* Methods Mol Biol, 2008. **439**: p. 131-45.

46. Fong, S.S., et al., *Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes.* J Biol Chem, 2006. **281**(12): p. 8024-33.

47. Lewis, N.E., et al., *Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models.* Mol Syst Biol, 2010. **6**: p. 390.

48. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

49. Orth, J.D. and B.Ø. Palsson, *Systematizing the generation of missing metabolic knowledge.* Biotechnol Bioeng, 2010. **107**(3): p. 403-12.

50. Kim, H.U., T.Y. Kim, and S.Y. Lee, *Metabolic flux analysis and metabolic engineering of microorganisms.* Molecular BioSystems, 2008. **4**(2): p. 113-120.

51. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox.* Nat. Protocols, 2007. **2**(3): p. 727-738.

52. Delli-Bovi, T.A., M.D. Spalding, and S.T. Prigge, *Overexpression of biotin synthase and biotin ligase is required for efficient generation of sulfur-35 labeled biotin in E. coli.* BMC Biotechnol, 2010. **10**: p. 73.

53. Fontecave, M., *Iron-sulfur clusters: ever-expanding roles.* Nat Chem Biol, 2006. **2**(4): p. 171-4.

Chapter 4: Gap-filling of the *Escherichia coli* metabolic network for model improvement and discovery

The *i*JO1366 reconstruction of the metabolic network of *Escherichia coli* is one of the most complete and accurate metabolic reconstructions available for any organism. Still, because our knowledge of even well-studied model organisms such as this one is incomplete, this network reconstruction contains gaps and possible errors. There are a total of 208 blocked metabolites and 127 orphan reactions in *i*JO1366. A new model improvement workflow was developed to compare model-based phenotypic predictions to experimental data to fill gaps and correct errors. A Keio Collection based dataset of *E. coli* gene essentiality was obtained from literature data and compared to model predictions. The SMILEY algorithm was then used to predict the most likely missing reactions in the reconstructed network, adding reactions from a KEGG based universal set of metabolic reactions. The feasibility of these putative reactions was determined by comparing updated versions of the model to the experimental dataset, and genes were predicted for the most feasible reactions. Numerous improvements to the *i*JO1366 reconstruction were suggested by these analyses. Finally, experiments were performed to verify several computational predictions, including a new function for the *leuABCD* operon and a new mechanism for growth on myo-inositol.

4.1 Introduction

Constraint-based modeling is a widely used systems biology method, and is particularly well suited for predicting the phenotypes of microbial organisms after gene knockouts or when grown on different substrates [1-3]. These variable conditions are simply represented as additional constraints on a model, and growth can be predicted by flux balance analysis (FBA) [9]. Because not every realistic constraint is represented in a typical metabolic model, it is quite possible for such a model to predict growth under conditions where growth does not really occur. The actual organism may not express a required gene, or fluxes may be limited by kinetic constraints, for example. On the other hand, the nature of these constraints causes false predictions of no growth to be indications that the model is missing an essential reaction [13]. No current metabolic network reconstruction is entirely complete and realistic because our knowledge of the metabolism of no organism is complete. Even in very well-studied model organisms such as *Escherichia coli* there are still many genes with unknown functions [14, 15]. The result of this is that there are gaps in metabolic network reconstructions. These gaps take the form of dead-end metabolites, which have either no producing or no consuming reactions.

The comparison of model predictions to experimental data can be a useful way to fill network gaps and discover new genes and reactions [16]. There are four possible outcomes when comparing computationally predicted to experimentally measured growth phenotypes: true positives, when the model correctly predicts growth; true negatives, when the model correctly predicts that no growth is possible; false positives, when the model predicts growth on a condition where growth was not observed; and false

negatives, when the model fails to predict growth where growth was actually observed. Both false positive and false negative results can be useful for refining model content, but it is the false negative cases that can help fill gaps. Several methods have been developed to predict the correct gap-filling reactions based on comparisons to experimental data.

The first such method to be published was called SMILEY [13]. The SMILEY algorithm was first developed and used to predict reactions missing from the *i*JR904 *E. coli* reconstruction [17] by comparing model growth predictions to Biolog growth data [18]. Several results were experimentally verified and new genes were characterized [13]. SMILEY was also recently used to predict gap-filling reactions in the RECON1 human metabolic reconstruction [19, 20]. The algorithms GapFill [21] and GrowMatch [22] were later developed, and could predict missing reactions by connecting model gaps and by comparing model predictions to gene essentiality data, respectively. To date, these methods have been used to make predictions for the *E. coli* and yeast metabolic networks [20, 22], but these predictions have not yet been experimentally verified.

The present study builds on these methods with a new workflow that includes use of the SMILEY algorithm. A large dataset of *E. coli* gene essentiality data from the Keio Collection [23], combined from four published datasets [12, 23-25], was assembled. Model growth predictions made using the *i*JO1366 *E. coli* metabolic network reconstruction were compared to this dataset, and both false positive and false negative comparisons were analyzed to identify potential errors in the model and in the experimental datasets. The SMILEY algorithm was then used to predict gap-filling reactions and putative reactions that correct false negative predictions. The feasibility of these reactions was then assessed by comparing augmented model predictions to the

experimental dataset. Finally, genes were predicted for the most feasible putative reactions. Several sets of gene function predictions are presented, and provide plausible hypotheses for experimental validation. These predictions have the potential to improve the metabolic reconstruction and lead to new metabolic gene discoveries [16]. Two sets of experiments were performed to demonstrate the utility of these types of predictions. Knockout strain growth phenotyping experiments were performed to identify a gene involved in myo-inositol metabolism, and *in vitro* enzyme assays were performed to identify alternate reactions for the enzymes of the LeuABCD operon.

4.2 Results

4.2.1 Remaining gaps in the *i*JO1366 network

The modified GapFind algorithm used in construction of the *i*JO1366 network was used to identify all gaps in the final version of the *i*JO1366 reconstruction [12]. Several different types of gaps in metabolic networks are possible. Root no-production gaps are metabolites with consuming reactions but no producing reactions. Root no-consumption gaps are metabolites with producing reactions but no consuming reactions. Downstream gaps are metabolites with producing and consuming reactions but which are unable to be produced at steady state because they are downstream of a root no-production gap. Similarly, upstream gaps are upstream of root no-consumption gaps. The final *i*JO1366 reconstruction contains 48 root no-production gaps, 63 root no-consumption gaps, 52 downstream gaps, and 69 upstream gaps (**Figure 4.1 a**). The total number of blocked metabolites in *i*JO1366 is 208, with some metabolites occurring as

more than one type of gap. In total, 11.5% of the metabolites in *i*JO1366 are blocked under all conditions due to gaps.

All gaps were manually sorted into scope and knowledge gaps. Scope gaps are metabolites that are blocked in a model due to the limited scope of the network reconstruction, but have actual known producing and consuming reactions. Knowledge gaps exist because our knowledge of any metabolic network is incomplete. More than half of the total blocked metabolites in *i*JO1366 are due to scope gaps. The two main classes of scope gaps in the model are tRNA related and metal ion related. Like its predecessor, *i*AF1260, *i*JO1366 contains charging reactions for all 20 standard amino acids as well as several non-standard amino acids such as N-formylmethionine and L-selenocysteine. These reactions are blocked because *i*JO1366 does not contain producing reactions for the uncharged tRNAs or consuming reactions for the charged tRNAs. These reactions could be used if the metabolic network is connected to a transcription and translation network [26], and thus, they are included for ease of integration and completeness of the reconstruction. *i*JO1366 also contains many reactions involving metal ions. Some metal ions, such as Fe^{2+}, Mg^{2+}, Ca^{2+}, and Na^{+} are included in the core and wild-type biomass reactions, providing a consuming reaction for these metabolites. Others, such as Ag^{+}, Hg^{2+}, and WO_4^{2-}, may be toxic to cells or may not serve any essential biological purpose. *E. coli* contains efflux transporters for such metals, but their exact uptake mechanisms are not known. Other scope gaps are due to metabolites that, like tRNAs, serve non-metabolic functions once they are produced. For example, *E. coli* resists osmotic stress by producing glycine betaine. [27].

Most root gaps have only one or no downstream or upstream gaps (**Figure 4.1 b**). This indicates that few long pathways in *i*JO1366 are blocked, and that most gaps have very small effects on the network as a whole. There are a few pathways blocked by gaps, however. For example, a set of nine metabolites including carnitine and carnitinyl-CoA are blocked by their downstream product γ-butyrobetainyl-CoA. This compound is not well characterized, but has been shown to be converted to crotobetainyl-CoA by *caiA* (b0039), although the mechanism and electron acceptor for this reaction are unknown [28].

A biologically realistic gap in *E. coli* K-12 metabolism occurs in the O-antigen synthesis pathway. An IS5 insertion in the *rfb* operon has left these *E. coli* strains without a functional rhamnosyltransferase, leaving rhamnosyl-N-acetylglucosamyl-undecaprenyl diphosphate without a producing reaction [29, 30]. Eleven downstream metabolites are also blocked by this gap, which is listed here as a scope gap. This is a real gap in the *E. coli* metabolic network, and is not due to limited knowledge.

Most downstream or upstream blocked metabolites are blocked by only one root gap (**Figure 4.1 c**). If the missing producing or consuming reaction is identified, the downstream or upstream metabolite would be unblocked. There are a few cases, however, in which a metabolite has both upstream and downstream gaps. Mercaptopyruvate, for example, has no known producing reaction in *E. coli*. This compound is consumed in a reaction catalyzed by *sseA* (b2521) that produces thiocyanate [31]. This product has no known consuming reactions [32]. In this unusual situation, neither of these compounds can be produced until both of these gaps are filled.

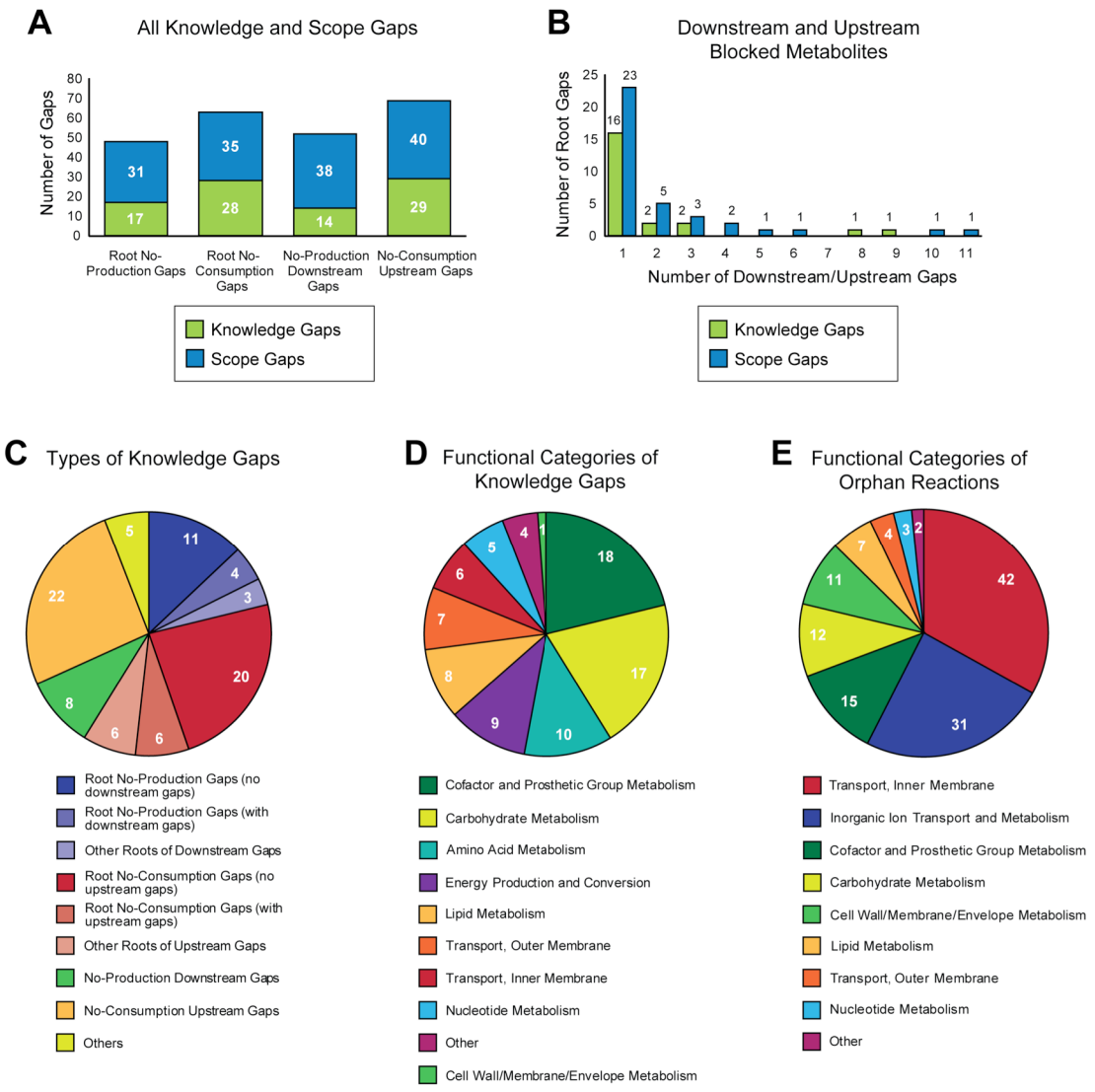

Figure 4.1 Properties of the gaps and orphan reactions in *i*JO1366. (**a**) Numbers of root no-production, root no-consumption, no-production downstream, and no-consumption upstream gaps in the network. (**b**) Histogram of the number of downstream or upstream blocked metabolites for each root gap. Most root gaps only result in one downstream gap. (**c**) The 85 knowledge gaps in *i*JO1366 sorted by type of gap. "Others" includes special cases such as metabolites that are both root and downstream gaps. (**d**) The 85 knowledge gaps by primary metabolic functional category (see **Figure 3.1**) of the reactions in which the blocked metabolites participate. (**e**) The 127 orphan reactions (excluding the artificial reaction *ATPM*) by functional category.

Knowledge gaps occur throughout the *i*JO1366 metabolic network, with the largest number occurring in cofactor and prosthetic group metabolism (**Figure 4.1 d**).

In addition to gaps, *i*JO1366 contains 128 orphan reactions. One of these, the ATP maintenance reaction, is not a real biological reaction, and is used for modeling purposes to simulate the non-growth associated maintenance requirement of *E. coli*. The other orphan reactions are due to incomplete knowledge of *E. coli* metabolism. Orphan reactions occur in all of the metabolic systems of *i*JO1366 except for "energy production and conversion" and "amino acid metabolism" (**Figure 4.1 e**). Most orphan reactions are inner membrane and inorganic ion transport reactions. One possible reason for this is that transport proteins tend to be more difficult to purify and assay than soluble enzymes.

Another notable feature of orphan reactions is that they are often grouped adjacent to each other in the *i*JO1366 network. Two reactions are considered adjacent if they have a common metabolite. Orphan reactions are adjacent to an average of 3.05 other orphans, while the average for all reactions in *i*JO1366 (excluding biomass, demand, and exchange reactions) is 1.53 adjacent orphans, a significant difference (p = 0.0005, t-test). This characteristic indicates that orphans are more common in certain poorly studied pathways and subsystems than in well studied pathways.

4.2.2 Comparison of model predictions to experimental data

By applying a newly developed workflow to analyze the *i*JO1366 model gaps and compare model predicted phenotypes to experimental data, new biological hypotheses were generated (**Figure 4.2**). First, the experimental datasets were assembled and combined. Each dataset consisted of a large set of *E. coli* gene knockout strains grown on different types of media. All of these gene knockout strains were from the Keio Collection of *E. coli* BW25113 single gene knockouts, allowing them to be analyzed

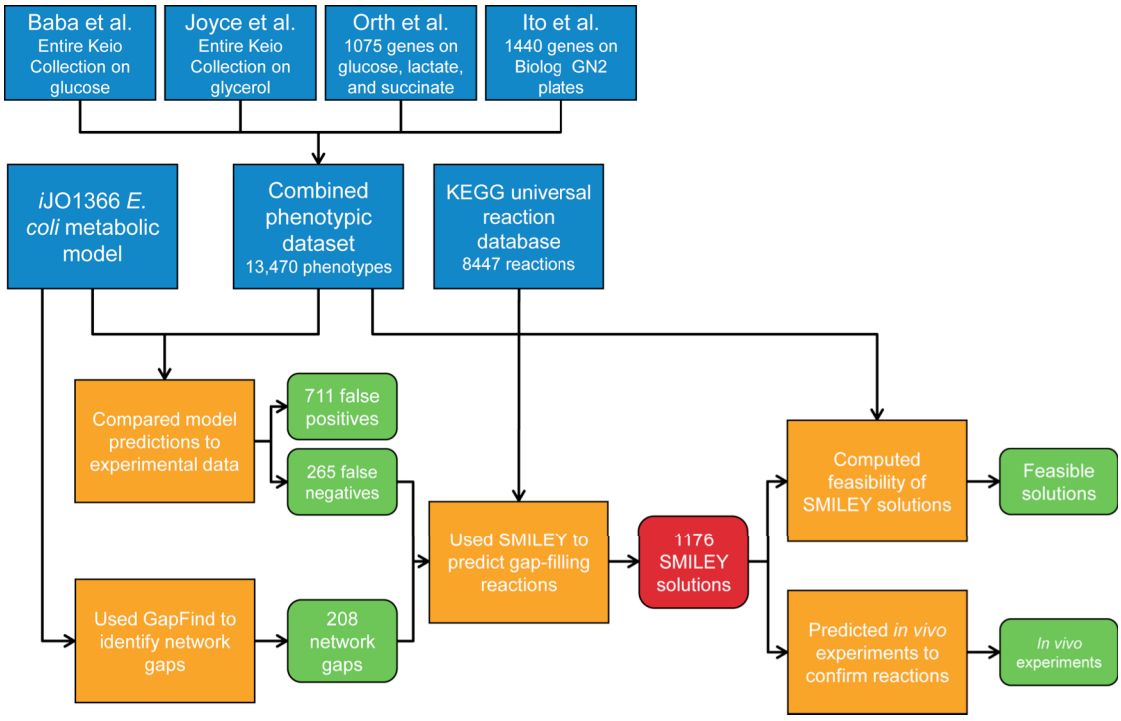

Figure 4.2 Workflow for predicting FN-correcting and gap-filling reactions using SMILEY. Input datasets are shown in blue, computational prediction steps are orange, and analyzed outputs are green.

together. The first dataset, from Baba et al [23], contained phenotypes from the entire Keio Collection grown on glucose MOPS minimal media. This defined media contains the buffer MOPS (3-(N-morpholino)propanesulfonic acid), a potential sulfur source. The second dataset was another growth screen of the entire Keio Collection, but on glycerol M9 minimal media [25]. The third dataset was a screen of 1075 Keio Collection strains, all included in the *i*AF1260 *E. coli* metabolic reconstruction, grown on four different conditions [12]. The strains were grown on glucose M9 media under both aerobic and anaerobic conditions, on lactate M9 aerobically, and on succinate M9 aerobically. The fourth dataset consisted of phenotypes from 1440 Keio Collection strains grown on

Biolog GN2 plates [24]. It was found that wild-type *E. coli* can grow on 38 different carbon sources on this Biolog plate, so the dataset only included these 38 substrates.

The four datasets were combined together into one large phenotypic dataset. From the screens of the entire Keio Collection on glucose and glycerol, a growth phenotype was included for each of the 1366 genes in *i*JO1366. For the screen on four conditions, phenotypes were available for 1075 of the 1366 genes. For the screen on Biolog plates, only 259 of the 1440 genes were also in *i*JO1366, so only these genes were included. Five of the 38 substrates were not included in the *i*JO1366 model, so these were not included. The phenotypes in this combined dataset were adjusted slightly from their original publications, based on a more recent analysis of the Keio Collection genotypes [33]. Several new genes were classified as essential, and these were added to the essential genes on glucose and glycerol. One gene, b0103, was removed from the Biolog screen data based on this analysis. The screen of 1075 strains on four conditions was published after the Keio Collection update, and thus already accounted for these changes. Some of the datasets contained phenotypes on the same substrates. For example, the Biolog data contained strains grown on glycerol, succinate, and lactate. In these cases, only one data point was included for each gene knockout strain grown on each substrate. If any one of the datasets included a "growth" phenotype, then it was set to "growth" in the combined dataset. Only if a strain had a "no growth" phenotype on all conditions was it classified as essential in the combined dataset. After making these adjustments, the final combined dataset contained 13,470 experimental phenotypes. There were 12,120 "growth" phenotypes and 1350 "no growth" phenotypes.

The *i*JO1366 *E. coli* metabolic network model was then used to predict growth phenotypes for these 13,470 conditions. The model of *E. coli* K-12 MG1655 metabolism was first modified slightly to match the genotype of *E. coli* BW25113, the parent strain of the Keio Collection. FBA [9] was then used to predict growth rates using the *i*JO1366 core biomass objective with each gene knockout and on every substrate in the experimental dataset. Any growth rate above zero was classified as a computational "growth" phenotype, while a growth rate of zero was classified as "no growth". An *in silico* dataset of 13,470 phenotypes was thus generated, and was compared to the *in vivo* dataset. Each model prediction was classified as either a true positive, true negative, false positive, or false negative. A total of 11,855 true positives, 639 true negatives, 711 false positives, and 265 false negatives were identified (**Figure 4.3 a**). Overall, the *in silico* screen predicted more growth phenotypes than were found in the experimental data (93.3% and 90.0%, respectively). This can largely be explained by the nature of constraint-based modeling and FBA. Because the *i*JO1366 model does not contain regulation, FBA may use any reaction in the network to produce biomass. In an *E. coli* cell, different levels of regulation may make certain enzymes unavailable under certain conditions, even if they may have allowed for growth. Other real constraints, such as kinetic constraints, may not be accounted for in the model and also may be the cause of false positive predictions.

The genes in the *i*JO1366 model have been classified into 11 functional categories, according to the metabolic functions they serve [12]. The different categories were found to contain genes with varying levels of predictive accuracy (**Figure 4.3 b**). Genes in the "Others" category, including mainly tRNA charging genes and genes that

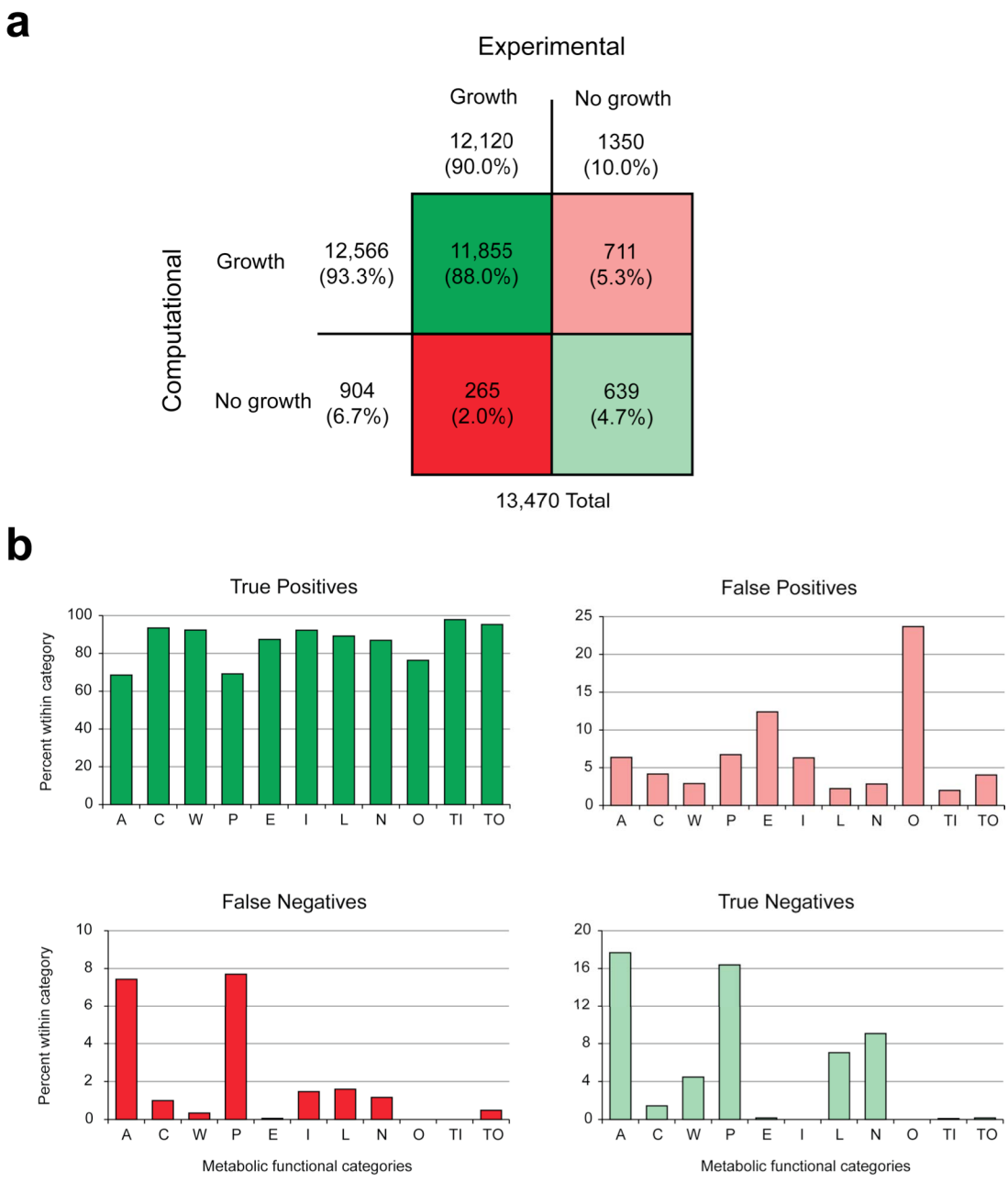

Figure 4.3 Comparison of model predicted growth phenotypes to experimental data. (**a**) The overall comparison, indicating numbers of true positives, true negatives, false positives, and false negatives. (**b**) The numbers of each type of prediction within 11 functional categories of metabolic reactions. The categories are: amino acid metabolism (A), carbohydrate metabolism (C), cell wall/membrane/envelope metabolism (W), cofactor and prosthetic group metabolism (P), energy production and conversion (E), inorganic ion transport and metabolism (I), lipid metabolism (L), nucleotide metabolism (N), other (O), inner membrane transport (TI), and outer membrane transport (TO).

could not be placed in the other categories, were found to lead to false positives in 23.7% of cases. This is due to the tRNA charging gaps in the *i*JO1366 model, discussed in **Section 4.2.1 Remaining gaps in the *i*JO1366 network**. These tRNA charging genes are essential *in vivo*. There were also many false positives among the "Energy Production and Conversion" genes (12.4 %). This may be due to the fact that disruptions to cellular energy generation may cause *E. coli* to grow very slowly, so that these strains would have been found to be essential in the experimental screens even though they were actually slowly growing. The computational screen classified all growing strains as non-essential, even if they grew slowly. False negatives were most common among the genes in "Amino Acid Metabolism" and "Cofactor and Prosthetic Group Metabolism," at 7.4% and 7.7% respectively. These false negative cases indicate the likely presence of currently unknown isozymes and alternative pathways.

False positive model predictions

The set of false positive predictions were investigated in more detail to determine why they occurred. Every gene that had a false positive prediction on at least one substrate and had no experimental growth on any substrate was tested. It was found that there are several possible reasons for a false positive prediction to be made by the *i*JO1366 model. First, it is possible that the model may contain an error such as an unrealistic reaction (**Table 4.1**). In the model, the reaction *CBPS* (carbamoyl phosphate synthase (glutamine-hydrolyzing)) converts L-glutamine to carbamoyl phosphate, an essential precursor of L-arginine. This gene is catalyzed by a complex of *carA* (b0032) and *carB* (b0033), which were experimentally found to be essential on glucose, glycerol,

succinate, and lactate minimal media. In the model, these genes are non-essential due to an alternate reaction that produces carbamoyl phosphate, *CBMKr* (carbamate kinase), catalyzed by the products of *yahI* (b0323), *arcC* (b0521), or *yqeA* (b2874). This putative reaction is included in *i*JO1366 based only on physiological data [34], and the functions of these genes are not well characterized. It is therefore likely that the *CBMKr* reaction is unrealistic. False positives may also be caused by errors in the *i*JO1366 core biomass reaction. The gene *pdxH* (b1638) catalyzes the reactions *PDX5POi* (pyridoxine 5'-phosphate oxidase) and *PYAM5PO* (pyridoxamine 5'-phosphate oxidase), required for the synthesis of pyridoxal 5'-phosphate (vitamin B_6). This vitamin is not included in the core biomass, so these reactions are not essential in the model. However, the essentiality of this gene on glucose and glycerol minimal media indicates that vitamin B_6 is in fact essential, and should be included in the model biomass reaction.

Some false positive predictions likely occurred because a gene was incorrectly identified as essential in one of the experimental screens (**Table 4.2**). This the case with several genes involved in energy production. The cytochrome oxidase gene *cydA* (b0733) knockout strain does not exist in the Keio Collection, and is presumed to be essential. A viable knockout strain for this gene has been produced, however, along with knockouts for other cytochrome oxidases [7]. The ATP synthase genes *atpCDGAHFEB* (b3731-8) were classified as essential on minimal media, but inspection of the actual growth measurements from these experiments [12, 23, 25] reveals that these knockout strains did actually grow, albeit slowly.

Many false positive cases occurred for gene knockout strains that have known isozymes or alternative pathways (**Table 4.3**). In the *i*JO1366 model, these knockouts are

Table 4.1 False positive model predictions that indicate model errors.

Gene	Error
carA (b0032)	alternate pathway (*CBMKr*) gene functions not confirmed
carB (b0033)	alternate pathway (*CBMKr*) gene functions not confirmed
proB (b0242)	alternate pathway (*NACODA*) gene function not confirmed
proA (b0243)	alternate pathway (*NACODA*) gene function not confirmed
folD (b0529)	5fthf[c] and methf[c] may be essential
entD (b0583)	enter[c] may be essential
pyrD (b0945)	alternate pathway (*DHORDfum*) is an orphan reaction
pdxH (b1638)	pydx5p[c] may be essential
pgsA (b1912)	pgp120[p] - pgp181[p] may be essential
nrdA (b2234)	alternate pathway (*RNDR1b – RNDR4b*) gene functions not confirmed
nrdB (b2235)	alternate pathway (*RNDR1b – RNDR4b*) gene functions not confirmed
ptsI (b2416)	alternate pathway (*GLCt2pp*) glucose transport not confirmed
waaK (b3623)	colipa[e] may be essential
wzyE (b3793)	eca4colipa[e] may be essential
ubiE (b3833)	reactions *AMMQLT8* and *OMBZLM* are blocked by gaps
ubiB (b3835)	alternate pathway (*OPXHH3*) is an orphan reaction
ppa (b4226)	isozymes, *ppx* (b2502) and *surE* (b2744), may be incorrect

Table 4.2 False positive model predictions that indicate incorrectly identified essential genes.

Gene	Reason for incorrect phenotype
cydA (b0733)	knocked out successfully by Portnoy et al. [7]
atpC (b3731)	ATP synthase knockout causes low growth rate
atpD (b3732)	ATP synthase knockout causes low growth rate
atpG (b3733)	ATP synthase knockout causes low growth rate
atpA (b3734)	ATP synthase knockout causes low growth rate
atpH (b3735)	ATP synthase knockout causes low growth rate
atpF (b3736)	ATP synthase knockout causes low growth rate
atpE (b3737)	ATP synthase knockout causes low growth rate
atpB (b3738)	ATP synthase knockout causes low growth rate

overcome by using the isozyme or alternative pathway to synthesize biomass components. *In vivo*, these genes may be essential because isozyme genes are not expressed under the experimental conditions, or they may not be capable of catalyzing the same reaction at a sufficient rate for growth to occur. These types of false positive

Table 4.3 False positive model predictions caused by isozymes or alternate pathways.

Gene	Isozyme or alternate pathway reactions
thrA (b0002)	*metL* (b3940) or *lysC* (b4024)
carA (b0032)	alternate reaction: *CBMKr*
carB (b0033)	alternate reaction: *CBMKr*
folA (b0048)	*folM* (b1606)
can (b0126)	*cynT* (b0339)
pyrH (b0171)	*cmk* (b0910)
int (b0657)	*lpp* (b1677)
fldA (b0684)	*fldB* (b2895)
fabA (b0954)	*fabZ* (b0180)
nrdA (b2234)	alternate reactions: *RNDR1b*, *RNDR2b*, *RNDR3b*, *RNDR4b*
nrdB (b2235)	alternate reactions: *RNDR1b*, *RNDR2b*, *RNDR3b*, *RNDR4b*
cysK (b2414)	*cysM* (b2421)
ptsI (b2416)	alternate reaction: *GLCt2pp*
cysA (b2422)	*modA* (b0763) + *modB* (b0764) + *modC* (b0765)
cysP (b2425)	*modA* (b0763) + *modB* (b0764) + *modC* (b0765)
guaB (b2508)	alternate reaction: *XPPT*
glyA (b2551)	alternate reaction: *GLYCL*
acpS (b2563)	*acpT* (b3475)
serA (b2913)	alternate reaction: *GHMT2r*
metC (b3008)	*tnaA* (b3708) or *malY* (b1622)
aroE (b3281)	*ydiB* (b1692)
ilvA (b3772)	*tdcB* (b3117)
metE (b3829)	*metH* (b4019)
ubiB (b3835)	alternate reaction: *OPHHX3*
glnA (b3870)	*ycjK* (b1297)
metL (b3940)	*thrL* (b0002) or *malY* (b1622)
ppa (b4226)	*ppx* (b2502) or *surE* (b2744)
serB (b4388)	alternate reaction: *GHMT2r*

model predictions cannot be overcome through standard FBA using a metabolic model. A model including regulation or other additional constraints is required. Many more false positives occur when tRNA charging genes are knocked out in the model (**Table 4.4**). Since the *i*JO1366 tRNA charging reactions are blocked by scope gaps, these important reactions cannot be used in the model. Finally, several false positives cannot be explained by the model alone (**Table 4.5**). For example, the gene *spoT* (b3650) is

Table 4.4 False positive model predictions caused by tRNA charging reactions.

Gene
ileS (b0026)
proS (b0194)
cysS (b0526)
leuS (b0642)
glnS (b0680)
serS (b0893)
asnS (b0930)
tyrS (b1637)
pheT (b1713)
pheS (b1714)
thrS (b1719)
aspS (b1866)
argS (b1876)
metG (b2114)
hisS (b2514)
alaS (b2697)
fmt (b3288)
trpS (b3384)
glyS (b3559)
glyQ (b3560)
valS (b4258)

Table 4.5 False positive model predictions that can't be explained by the *i*JO1366 model.

Gene
ftsI (b0084)
adk (b0474)
mrdA (b0635)
cydC (b0886)
gapA (b1779)
ligA (b2411)
suhB (b2533)
eno (b2779)
fbaA (b2925)
pgk (b2926)
dut (b3640)
spoT (b3650)
pslB (b4041)
alsK (b4084)

required to synthesize the signaling molecule guanosine tetraphosphate (ppGpp). Since the metabolic model does not require signaling, this gene is found to be non-essential. Experimentally, *spoT* is essential on rich media, and this is likely due to its non-metabolic function.

False negative model predictions

All genes with false negative predictions for at least one substrate and no computationally predicted growth on any substrate were investigated in more detail. If a constraint-based metabolic model fails to predict growth under a condition where growth was observed experimentally, it is an indication of missing metabolic reactions or pathways in the model. In the next section, use of the SMILEY algorithm to predict likely missing reactions is presented. There are several other possible explanations for false negative predictions. First, it is possible that the model biomass reaction being used as an objective is incorrect (**Table 4.6**). Several false negative cases occurred with knockouts of genes involved in molybdenum cofactor synthesis, including *mog* (b0009), *moaA* (b0781), *moaC* (b0783), *moaD* (b0784), *moaE* (b0785), *moeA* (b0826), *moeB* (b0827), and *mobA* (b3857). In the *i*JO1366 model, these genes are essential because they are required to produce bmocogdp[c] (bis-molybdopterin guanine dinucleotide), a component of the core biomass reaction. Because these gene knockout strains are experimentally viable on most conditions, it is likely that this cofactor is not essential for growth, and thus should not be included in the *i*JO1366 core biomass reaction.

Two false negative cases could be explained by incorrect gene-protein-reaction associations (GPRs) in *i*JO1366 (**Table 4.7**). In one, the gene *hisH* (b2023) is required

Table 4.6 False negative model predictions caused by incorrect core biomass composition.

Gene	Biomass component
mog (b0009)	bmocogdp[c]
moaA (b0781)	bmocogdp[c]
moaC (b0783)	bmocogdp[c]
moaD (b0784)	bmocogdp[c]
moaE (b0785)	bmocogdp[c]
moeA (b0826)	bmocogdp[c]
moeB (b0827)	bmocogdp[c]
ubiX (b2311)	2ohph[c]
iscS (b2530)	bmocogdp[c]
cysG (b3368)	sheme[c]
mobA (b3857)	bmocogdp[c]
ubiC (b4039)	2ohph[c]

Table 4.7 False negative model predictions that suggest changes to *i*JO1366 model GPRs.

Gene	GPR correction
hisH (b2023)	not essential for *IG3PS* [8]
cyaY (b3807)	not essential for *I2FE2SS*, *I2FE2SS2*, *S2FE2SS*, *S2FE2SS2*

for the reaction *IG3PS* (Imidazole-glycerol-3-phosphate synthase), along with *hisF* (b2025). This reaction is an essential part of the histidine synthesis pathway, and is thus essential on all minimal media in the model. In the *in vivo* datasets, however, *hisH* is not essential under any aerobic conditions. It is not essential because without HisH, HisF is still able to catalyze this reaction, using NH_3 instead of glutamine as an N donor [8]. *hisH* should therefore not be an essential component of the *IG3PS* GPR. The other GPR change suggested is for *cyaY* (b3807), a gene involved in transferring iron during [Fe-S] cluster synthesis. In the model, this gene is an essential component of two reactions in

Table 4.8 False negative model predictions due to misidentified experiment phenotypes or media compositions.

Gene	Explanation
mtn (b0159)	essential according to Choi-Rhee et al. [6]
thiI (b0423)	possibly thiamin in media due to incomplete washing
bioA (b0774)	possibly biotin in media due to incomplete washing
bioB (b0775)	possibly biotin in media due to incomplete washing
bioF (b0776)	possibly biotin in media due to incomplete washing
bioC (b0777)	possibly biotin in media due to incomplete washing
bioD (b0778)	possibly biotin in media due to incomplete washing
aroD (b1693)	only experimental growth under one condition, possible error
thiD (b2103)	essential according to Orth et al. [12]
cysD (b2752)	only experimental growth under one condition, possible error
argG (b3172)	only experimental growth under one condition, possible error
cysG (b3368)	only experimental growth under one condition, possible error
bioH (b3412)	possibly biotin in media due to incomplete washing
ilvE (b3770)	only experimental growth under one condition, possible error
thiH (b3990)	essential according to Orth et al. [12]
thiG (b3991)	essential according to Orth et al. [12]
thiF (b3992)	essential according to Orth et al. [12]
thiE (b3993)	essential according to Orth et al. [12]
thiC (b3994)	essential according to Orth et al. [12]
cysQ (b4214)	MOPS is a possible alternate S source

both the ISC and SUF [Fe-S] cluster synthesis pathways, and is essential under all conditions. This gene is still not well characterized, and since it is experimentally non-essential, it is likely not strictly required for the reactions *I2FE2SS*, *I2FE2SS2*, *S2FE2SS*, and *2FE2SS2*. Other false negative cases are likely due to experimental errors (**Table 4.8**). Several genes involved in the synthesis of the cofactors biotin and thiamin were experimentally classified as non-essential. These cofactors are known to be required in small quantities [35-37], so it is likely that there was residual biotin and thiamin in the media during growth experiments. In the experimental screen on four different conditions, more thorough washing procedures were used to prevent carryover of

Table 4.9 False negative model predictions caused by missing isozymes or alternate pathways.

Gene	Putative Isozyme	E-value
purK (b0522)	*purT* (b1849)	2.00E-12
gltA (b0720)	*prpC* (b0333)	1.00E-41
aspC (b0928)	*tyrB* (b4054)	4.00E-94
fabH (b1091)	none identified	
pabC (b1096)	*ilvE* (b3770)	7.00E-8
icd (b1136)	*dmlA* (b1800)	3.00E-19
aldA (b1415)	*gabD* (b2661)	1.00E-90
	prr (b1444)	3.00E-80
	feaB (b1385)	1.00E-69
	aldB (b3588)	2.00E-66
	betB (b0312)	4.00E-65
ubiX (b2311)	none identified	
luxS (b2687)	none identified	
thyA (b2827)	none identified	
zupT (b3040)	none identified	
folB (b3058)	*folX* (b2303)	1.00E-4
argG (b3172)	none identified	
folP (b3177)	none identified	
yrbG (b3196)	none identified	
kdsC (b3198)	none identified	
argD (b3359)	*astC* (b1748)	1.00E-146
	gabT (b2662)	3.00E-64
	puuE (b1302)	3.00E-54
	patA (b3073)	4.00E-52
	hemL (b0154)	3.00E-32
cysG (b3368)	none identified	
ilvE (b3770)	*pabC* (b1096)	8.00E-8
dapF (b3809)	none identified	
argC (b3958)	none identified	
argB (b3959)	none identified	
hemE (b3997)	none identified	
ubiC (b4039)	none identified	
purA (b4177)	none identified	

preculture media, and these genes were classified as essential. Finally, false negatives can be caused by currently unidentified isozymes (**Table 4.9**). For cases in which false negatives could not be explained by other means, BLASTp was used to identify possible

isozymes in the *E. coli* genome. One predicted isozyme has already been experimentally verified. *prpC* (b0333), which currently is associated with *MCITS* (2-methylcitrate synthase), has been confirmed to also be an isozyme of *gltA* (b0720), catalyzing *CS* (citrate synthase) [38, 39].

4.2.3 Computational prediction of gap-filling reactions

One cause of model gaps and false negative phenotypic predictions is that some realistic reactions may be missing from the *i*JO1366 model. The SMILEY algorithm was used to predict the most likely sets of reactions missing from the model. To predict false negative resolving reactions, the model was constrained to match each false negative condition, one at a time, and SMILEY was run. For gene knockout strains which lead to false negative predictions on all 34 tested substrates (or all but one or two), it is likely that the same set of missing reactions is the cause of all incorrect predictions for this strain. In these cases, SMILEY was run on the model with only glucose (both aerobic and anaerobic), glycerol, lactate, and succinate as substrates. To predict gap-filling reactions, a small lower bound was placed on the known producing or consuming reaction for each knowledge gap metabolite, and SMILEY was run. In order to actually carry a small flux through these reactions and satisfy all model constraints, a gap-filling reaction or set of reactions would need to be added. SMILEY was run on 166 false negative cases and 49 gap reactions. The algorithm was set to find up to 25 alternate solutions for each condition, and a time limit of 2 h was placed on each solution. The reactions added by SMILEY were from a universal set of reactions based on all reactions in KEGG Release 58.0 [40]. Unrealistic and incomplete reactions were removed from this set.

A total of 1176 optimal and suboptimal solutions were identified by SMILEY. Solutions were identified for 106 of the false negative cases and for 32 gaps. Multiple optimal solutions were found for many cases, and there were a total of 198 different optimal solutions and 983 different suboptimal solutions. Five solutions were found as both optimal and suboptimal solutions in different cases, and 385 solutions were found multiple times. Most of these were for gene knockout strains grown on multiple substrates or for genes that are required by the GPRs of the same reaction or for reactions in the same pathway. For most false negative cases and gaps, only a small number of optimal solutions were found (**Figure 4.4 a**). No solution was found for 77 cases, and only one or two solutions were found for 91 cases. In four cases, all 25 solutions were optimal. These cases were for the *iscS* (b2530) knockout strain grown on four different conditions. This gene is a part of the ISC [Fe-S] cluster generation system and in the model is essential due to its role in molybdenum cofactor synthesis. Each of these alternate solutions involves the import of this cofactor. The average number of optimal solutions found per SMILEY run is 2.56. Most optimal solutions involve only one reaction, and none involve more than five (**Figure 4.4 b**). The average number of reactions per optimal solution is 1.41.

As the molybdenum cofactor uptake reactions demonstrate, not all SMILEY solutions are realistic. It would be extremely time consuming to manually check all 1176 solutions, so a computational feasibility check was performed. Each of the solutions was added to the *i*JO1366 model one at a time, and the augmented model was then used to predict growth phenotypes by FBA on all 13,470 conditions from the experimental dataset. The false positive and false negative predictions were identified by comparison

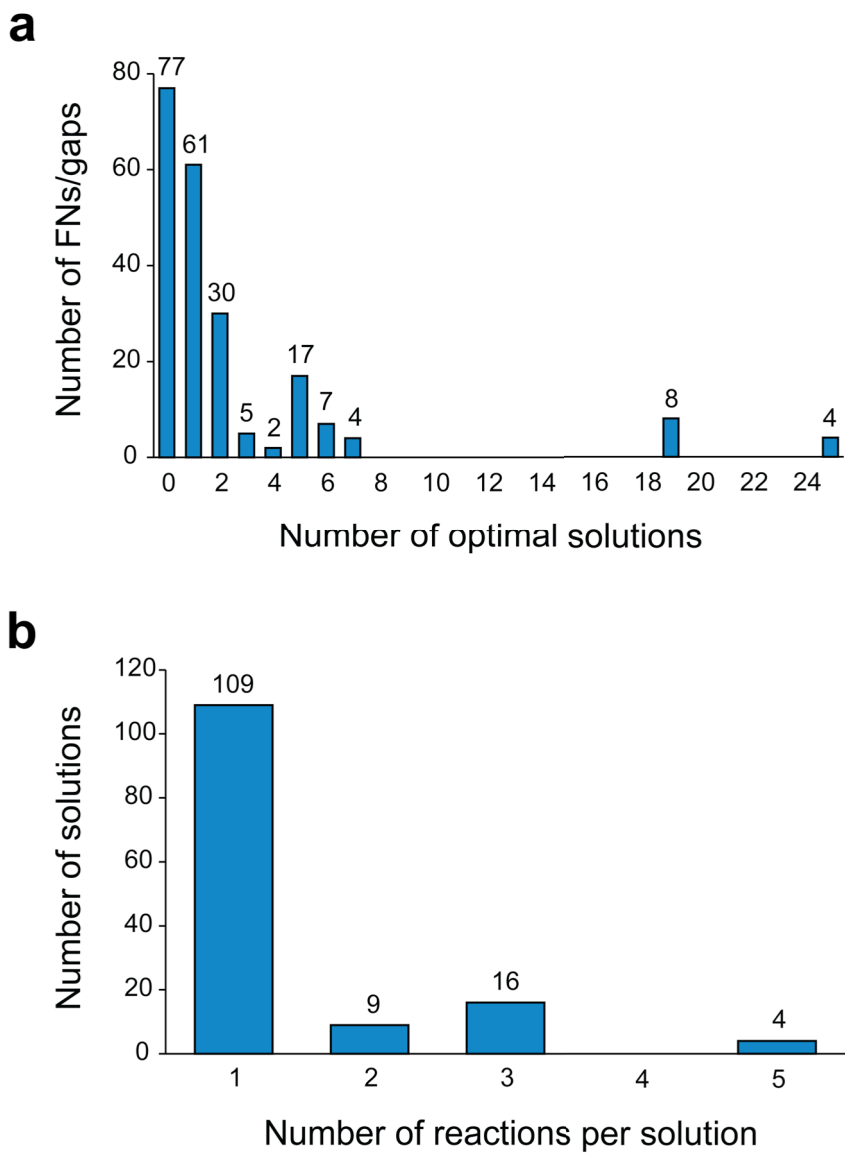

Figure 4.4 Properties of the 198 optimal SMILEY solutions. (**a**) Number of optimal solutions per SMILEY run. For most cases, between zero and two optimal solutions were found. (**b**) Number of reactions per optimal solution. Most optimal solutions consisted of only one reaction.

to the experimental dataset, and the number of false negatives eliminated and new false positives created by each SMILEY solution could be counted. The most feasible solutions would be those that fixed the most false negatives while introducing few false

positives. On average, each solution corrected 7.48 false negatives and created 7.07 new false positives. A total of 144 solutions (11 optimal and 133 suboptimal) were found that eliminate false negatives while producing no new false positives. GapFind was also run on the model with each solution added, to determine if any model gaps were eliminated. 74 solutions that fill at least one gap were found.

4.2.4 Predictions of genes for hypothesized reactions

The most feasible SMILEY solutions out of the complete set of 1176 solutions were investigated in more detail. For the most feasible solutions, BLASTp was used to try to identify candidate genes in the *E. coli* genome (**Table 4.10**). These solutions were divided into four categories. Category I solutions were optimal solutions that eliminated at least one false negative condition while creating no new false positives. Of the 11 category I solutions, five fixed false negatives by adding the deleted model reaction back in from the universal reaction list. This indicates that uncharacterized isozymes are possible for *aspC* (b0928), *pabC* (b1096), *aldA* (b1415), *argD* (b3359), and *hemE* (b3397). Another solution suggested that false negatives for Δ*aspC* strains could be corrected by adding the existing model reaction *ASP1DC* (aspartate 1-decarboxylase) in reverse. No literature evidence was found to support or refute the reversibility of this reaction. The other five solutions involve the addition of new reactions to the model. Four of these provide potential production routes for aspartate to compliment an *aspC* deletion. The other provides a new reaction to consume glycoaldehyde for Δ*aldA* strains. None of these five reactions have associated genes in the KEGG database, indicating that they are

global orphan reactions. Candidate genes for these reactions could not be identified with no reference sequences available.

The second category of SMILEY solutions to be investigated in detail was all optimal solutions that fixed more false negatives than the number of new false positives they created. There were 70 category II solutions (not include the category I solutions, which also fall within this definition). Most of these solutions involve the uptake of molybdenum cofactors or their precursors. As explained above, the most likely explanation for these false negatives is that the molybdenum cofactor is not strictly required for growth by *E. coli*. Several other solutions added deleted reactions back into the network, and two solutions added feasible new reactions. In one, a slightly different reaction for producing dTMP was added to compliment a *thyA* (b2827) deletion. A currently uncharacterized *E. coli* gene, *ybiU* (b0821), was identified by BLASTp as a candidate gene for this reaction, providing a testable hypothesis for the function of this gene. The other category II feasible solution added a new reaction to convert L-glutamate to α-ketoglutarate. Two candidate genes with high sequence homology to known genes from other organisms, *ydbL* (b0600) and *ydcR* (b1439), were found.

The third category to be investigated consisted of the suboptimal solutions that fixed at least one false negative while producing no new false positives. 133 category III solutions were found. Some of these solutions included unrealistic reactions, such as the oxygen consuming KEGG reaction R00357 in the reverse, oxygen producing direction. Others attempt to compensate for the loss of cofactor producing pathways by simply adding new uptake reactions for those cofactors. Still, several realistic reactions were suggested and candidate genes were identified by BLASTp. One solution consisted of the

Table 4.10 Predicted genes for most feasible FN-correcting SMILEY solutions.

Hypothesized changes in directionality			
Reaction	**Category**	**Support**	
ASP1DC	I		
ASPT	III	reversible (Karsten and Viola [4])	
ICL	IV	reversible (MacKintosh and Nimmo [11])	
AKGDH	IV	not reversible (EcoCyc)	
CITL	IV		
Hypothesized gap-filling reactions			
Reaction	**Category**	**Putative gene**	**E-value**
R00352 (R)	IV	*sucD* (b0729)	2.00E-20
R00373 (F)	I	global orphan	
R00400 (F)	I	global orphan	
R00507 (R)	IV	*yhfW* (b3380)	0.47
R00529 (F)	IV	*cysN* (b2751) and *cysD* (b2752) *	
R00530 (F)	IV	global orphan	
R00531 (R)	IV	global orphan	
R00695 (R)	I	global orphan	
R00709 (F)	IV	*dmlA* (b1800)	6.00E-26
		icd (b1136)	1.00E-26
		leuB (b0073)	2.00E-15
R00732 (R)	III	*aroA* (b0908)	5.00E-32
		murA (b3189)	7.00E-8
R00733 (R)	III	*tyrA* (b2600)	2.80E-2
R01393 (R)	I	global orphan	
R01618 (R)	IV	*glgP* (b3428)	2.10
R01713 (F)	I	global orphan	
R01731 (F)	IV	*tyrB* (b4054) *	
R01785 (R)	III	*rhaD* (b3902) *	
R01902 (R)	III	*rhaB* (b3904) *	
R02200 (F)	IV	global orphan	
R04209 (R)	IV	*purC* (b2476)	7.00E-16
R05717 (R)	IV	*cysH* (b2762)	3.00E-12
R06613 (F)	II	*ybiU* (b0821)	1.6
R07164 (R)	III	*ydiJ* (b1687)	0.9
R07165 (R)	III	*ydiJ* (b1687)	0.9
R07176 (R)	IV	global orphan	
R07463 (F)	IV	*dadA* (b1189)	2.00E-18
R07613 (R)	II	*ydbL* (b0600)	7.00E-26
		ydcR (b1439)	6.00E-15
R08553 (R)	IV	*ysaA* (b3573)	4.00E-5

* Genes already assigned to reactions for *E. coli* in KEGG according to their EC Numbers

addition of the current model reaction *ASPT* (L-aspartase) in reverse. Experimental evidence supports the reversibility of this reaction [4], which is currently listed as irreversible in *i*JO1366. The fourth and final category of SMILEY solutions to be examined was all other optimal solutions that were not in categories I and II. There were 62 solutions in this category. Most of these solutions were simply new uptake reactions for blocked essential biomass components, but 14 new realistic reactions were suggested, as well as three current model reactions running in their opposite directions. One of these new reversible reactions, *ICL* (isocitrate lyase), was confirmed in a published study [11], while another, *AKGDH* (2-Oxoglutarate dehydrogenase), is not actually reversible according to EcoCyc [41].

All 72 gap-filling SMILEY solutions were also investigated, and BLASTp was used to predict genes for the realistic reactions (**Table 4.11**). A total of 20 new realistic reactions were found, and candidate genes could be predicted for about half of them. The others were global orphan reactions. SMILEY also suggested that 15 existing model reactions could be made reversible to fill gaps. According to EcoCyc, most of these reactions are not reversible. However, evidence was found supporting the reversibility of two model reactions, *DKGLCNR1* (2,5-diketo-D-gluconate reductase) [5] and *DKGLCNR2y* (2,5-diketo-D-gluconate reductase (NADPH)) [10].

4.2.5 Experimental validation of predicted genes

SMILEY and other gap-filling algorithms are useful because they can use a model and existing experimental data to generate predictions. Without performing an experiment to verify these predictions, they are only hypotheses. The *i*JO1366 model was

Table 4.11 Predicted genes for gap-filling SMILEY solutions.

Hypothesized changes in directionality	
Reaction	**Support**
DOGULNR	not reversible (EcoCyc)
DKGLCNR1	reversible (Habrych et al. [5])
DKGLCNR2y	reversible (Yum et al. [10])
PGLYCP	not reversible (EcoCyc)
CYSSADS	
HMPK1	not reversible (EcoCyc)
4HTHRS	
HETZK	not reversible (EcoCyc)
NNDMBRT	not reversible (EcoCyc)
ACONMT	not reversible (EcoCyc)
CINNDO	not reversible (EcoCyc)
MCPST	
GPDDAS	not reversible (EcoCyc)
APCS	not reversible (EcoCyc)
SARCOX	

Hypothesized gap-filling reactions		
Reaction	**Putative gene**	**E-value**
R01742 (F)	*ydiS* (b1699)	0.003
R00893 (F)	*ygfM* (b0419)	0.58
R02133 (F)	*yhbO* (b3153)	0.069
R02721 (F)	global orphan	
R03472 (R)	global orphan	
R01297 (R)	global orphan	
R01299 (R)	global orphan	
R02252 (F)	*fadH* (b3081)	7.00E-71
	nemA (b1650)	1.00E-22
R00895 (R)	*aspC* (b0928)*	
R03530 (F)	*ndk* (bb2518)*	
R00012 (F)	global orphan	
R01232 (R)	*yjhG* (b4297)	0.028
	yagF (b0269)	0.057
R00838 (F)	*chbF* (b1734)	9.00E-42
	melA (b4119)	2.00E-21
R00655 (R)	global orphan	
R07300 (F)	global orphan	
R00683 (F)	global orphan	
R00367 (F)	global orphan	
R02559 (F)	global orphan	
R02560 (F)	global orphan	
R05623 (F)	*yjiN* (b4336)	0.16

* Genes already assigned to reactions for *E. coli* in KEGG according to their EC Numbers

used to design simple growth phenotype experiments to confirm some of these predictions. Each of the 1176 solutions was added to the model one at a time, and growth was simulated on all combinations of a set of 115 carbon sources and 62 nitrogen sources under both aerobic and anaerobic conditions. These substrates were selected for being readily available chemicals for use in the laboratory. For every substrate combination on which growth is predicted for the model with a SMILEY solution added, but not for the unmodified *i*JO1366 model, an *in vivo* experiment can be performed to determine if *E. coli* can actually grow with those substrates, giving supporting experimental evidence to the predicted reactions. For most solutions, no new growth conditions were identified. However, several realistic solutions were predicted to grow on a set of nitrogen containing compounds under anaerobic conditions. These reactions mostly involved conversions between TCA cycle intermediates and amino acids derived from TCA cycle compounds. For example, the reaction R00373 converts glycine and oxaloacetate to glyoxylate and L-aspartate. Five false negatives are fixed by adding this reaction, and according to the *i*JO1366 model this reaction enables growth on several combinations of adenine, glycine, ethanol, and hypoxanthine. Based on predictions such as these, eight substrates were selected for growth experiments: adenine, L-arginine, L-cysteine, glycine, hypoxanthine, ornithine, L-proline, putrescine. Wild-type *E. coli* K-12 MG1655 was grown on all pairwise combinations of these compounds in minimal media without any other nitrogen or carbon sources under anaerobic conditions. *E. coli* was also grown with glucose and ammonium chloride in combination with these compounds as a control. Three replicates of each condition were used, and OD_{600} was measured after 48 h each time. Significant growth was only observed for five conditions: glucose with ammonium

chloride, glucose with L-Arg, glucose with L-Cys, glucose with Gly, and glucose with ornithine. Since *E. coli* did not grow on any conditions without glucose, no evidence was provided for these predicted reactions. Still the negative result of this experiment does not disprove the presence of these reactions. The required genes might not be expressed under these conditions, or other constraints may be preventing growth.

Another reaction for which a growth experiment was predicted to be possible was R01184, myo-inositol:oxygen oxidoreductase. This reaction combines myo-inositol with oxygen to form D-glucuronate and water. Myo-inositol is a root-no consumption gap in the *i*JO1366 model, and this reaction fills this gap. With this reaction included, *E. coli* is predicted to grow with inositol as a substrate. An *in vivo* experiment was performed, and wild-type *E. coli* was inoculated into 2 g/L myo-inositol minimal media with no other carbon sources. Three replicates were performed, and after 72 h, the cultures had reached an OD_{600} of 0.017 ± 0.006. This indicates that *E. coli* can grow very slowly with myo-inositol as its only carbon source. Next, four candidate genes were predicted for this reaction using BLASTp. These genes were *ydeN* (b1498), *yfdE* (b2371), *yphC* (b2545), and *yhiJ* (b3488). The functions of these genes are currently unknown, and they are non-essential. The Keio Collection knockout strains for these four genes were then obtained and grown in both glucose and myo-inositol minimal media. All four strains grew to a similar OD_{600} as wild-type *E. coli* on glucose, but on myo-inositol the knockout strains did not grow as well (**Figure 4.5**). The *yhiJ* knockout strain did not grow at all, indicating this gene may code for a myo-inositol:oxygen oxidoreductase.

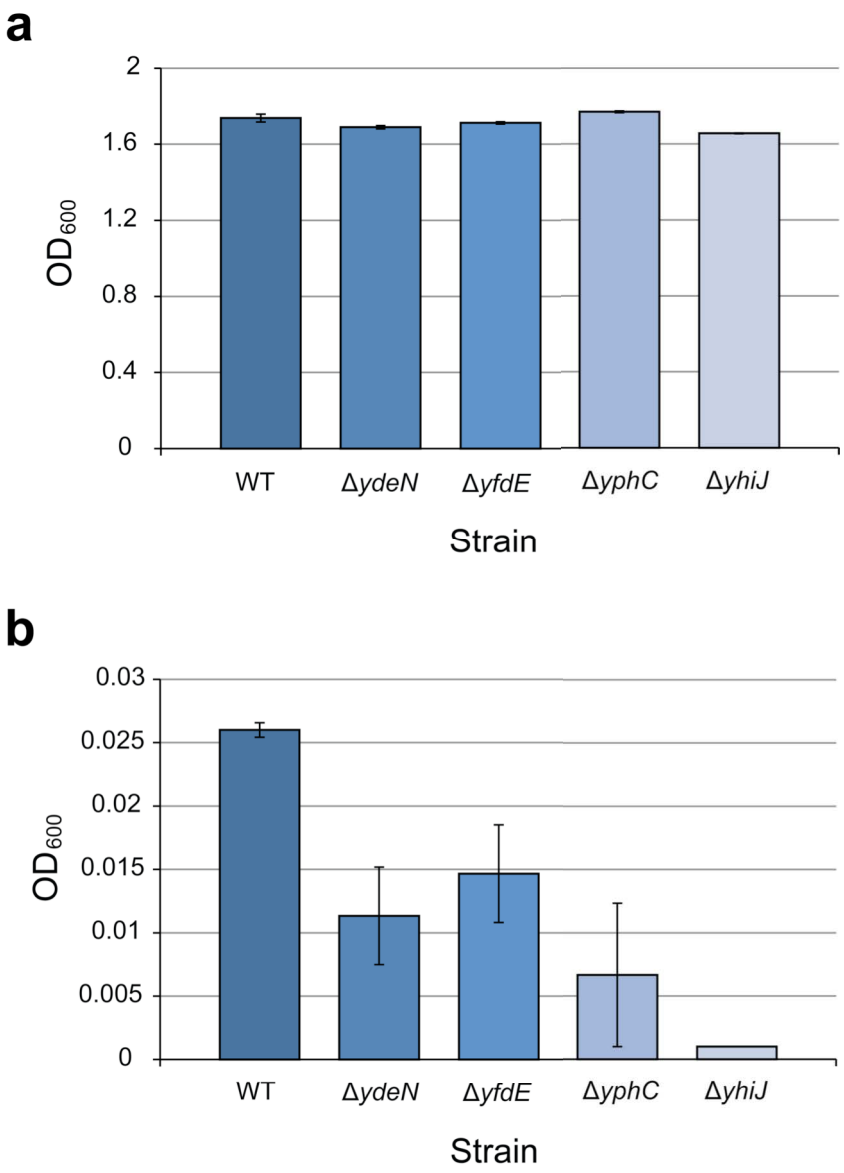

Figure 4.5 Growth of four Keio Collection gene knockout stains to identify a possible myo-inositol:oxygen oxidoreductase. (**a**) Growth of the four strains and WT on glucose M9 minimal media. (**b**) Growth of the four strains and WT on myo-inositol minimal media.

A different experimental approach was used to validate another predicted set of gene functions. In many microorganisms, the genes of the *leuABCD* operon, which code for enzymes in the leucine synthesis pathway, are also known to catalyze a second set of reactions [42, 43]. The LeuA enzyme converts pyruvate and acetyl-CoA to R-2-

methylmalate (citramalate). Next, the LeuCD protein complex catalyzes the conversion of R-2-methylmalate to D-erythro-3-methylmalate, with 2-methylmalate (citraconate) as an intermediate. In the final step, the LeuB enzyme converts D-erythro-3-methylmalate to 2-oxobutanoate and CO_2, while reducing one NAD^+ to NADH. To test the functions of these genes in *E. coli*, a series of *in vitro* enzyme assays were performed. Enzymes were expressed and purified from the ASKA Collection of His-tagged *E. coli* ORFs, and these were added to buffer solutions containing the predicted reaction substrates and cofactors. The *in vitro* reactions were run at 37°C overnight, and the metabolites in reaction and control mixtures (with no enzyme added) were measured my LC-MS. The purified LeuA enzyme produced a significant amount of R-2-methylmalate from pyruvate, confirming this predicted function (**Figure 4.6 a**). The complete LeuCD reaction is an isomerization, so heavy water (D_2O) was added to the reaction mixture to give the product a slightly higher mass than the substrate. The substrate of this reaction, R-2-methylmalate, was not available, so the LeuA reaction was combined with the LeuCD reaction *in vitro*. All three enzymes were added to a mixture containing pyruvate and acetyl-CoA, and products of both reactions were detected, confirming the predicted LeuCD functions as well (**Figure 4.6 b**). The LeuB reaction could not be confirmed *in vitro*, despite numerous attempts. The purified enzyme could also not be confirmed to catalyze the known *E. coli* LeuB reaction from the leucine synthesis pathway, so the enzyme was likely not active under the experimental conditions.

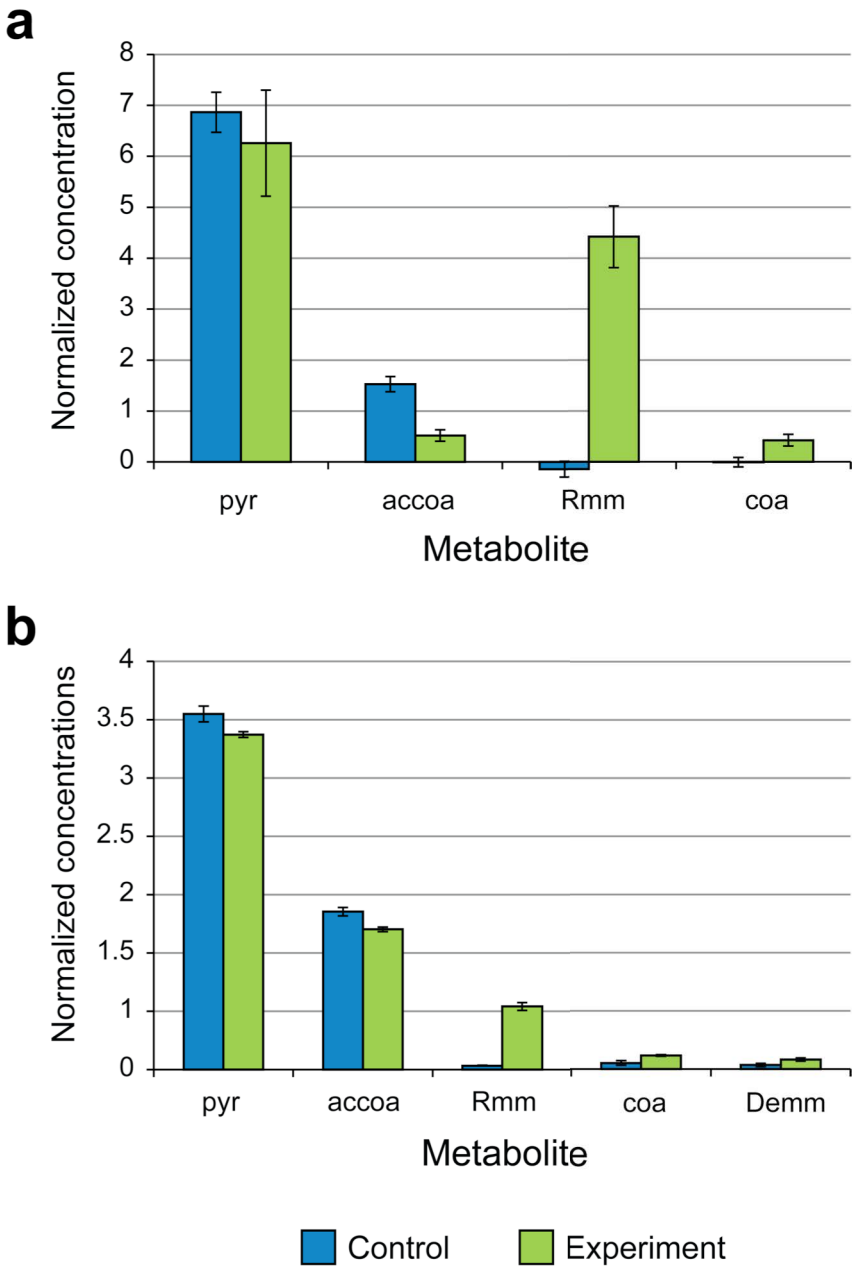

Figure 4.6 *In vitro* enzyme assays to identify alternate functions of the *E. coli* enzymes LeuA and LeuCD. Metabolite concentrations were measured by LC-MS both with the enzyme added to a reaction mixture (Experiment) and without (Control). (**a**) Metabolites with the LeuA enzyme and (**b**) metabolites with both LeuA and LeuCD added to reaction mixtures containing pyruvate (pyr) and acetyl-coA (accoa). The products of the reactions were R-2-methylmalate (Rmm), coenzyme A (coa), and D-erythro-3-methylmalate (Demm).

4.3 Discussion

In this study, the *i*JO1366 metabolic network model of *E. coli* was used as a discovery tool, leading to predictions for new metabolic gene functions. The most up to date model represents the current state of knowledge of *E. coli* metabolism in a structured format, and by comparing model predictions to experimental data, errors and gaps in this knowledge can be identified. A large dataset was assembled from Keio Collection gene knockout phenotypes grown on 34 different substrates. These phenotypes were compared to model predicted phenotypes, and the model false positive and false negative predictions were identified. When analyzed, the false positive predictions indicated several possible errors in the current model, including pathways thought to be catalyzed by poorly studied enzymes, and the uncertain requirements of *E. coli* biomass formation. The false negative cases also indicated several potential errors in the biomass as well as several likely experimental errors. Importantly, the false negative cases also indicate where the model is currently incomplete. The SMILEY algorithm was used to predict the most likely missing reactions, and in a novel procedure, the feasibility of these predictions was assessed through comparisons to the experimental data and through model gap analysis. Gene predictions were made for the most feasible predicted reactions, and experimental evidence was generated to support the predicted functions of *yhiJ* and *leuABCD*.

Through careful analysis of the false positive results, several possible model errors were identified. In some cases, the model predicted growth when pathways with poorly characterized genes were required. These genes need to be investigated in more detail. If future experimental evidence shows that they do not encode the enzymes they

are currently believed to encode, then the model can be updated by removing these reactions. False negative results, on the other hand, can indicate errors in experimental design, rather than in the model. Several cases were identified in which *E. coli* could grow on minimal media despite lacking genes for the synthesis of essential cofactors. Biotin and thiamin are known to be essential, but on several substrates growth was observed for knockout strains that should not be able to produce these cofactors. One possible explanation is that there are alternate synthesis routes for these compounds, but in these cases, it is more likely that trace amounts of biotin and thiamin were present in the media during the growth experiments [23, 25]. Both false positive and false negative results indicate potential errors in the *i*JO1366 core biomass reaction. False positives indicate a gene that helps produce an essential compound that is missing from the biomass reaction, while false negatives indicate a gene that produces a non-essential compound that is included. These biomass components may only be essential or non-essential under certain conditions, however, necessitating the use of condition-specific biomass reactions for specific model applications.

SMILEY was used to predict both gap-filling and false negative correcting reactions that could be added to the metabolic network. Some of these reactions were the same as existing model reactions but in the opposite direction. Literature data was searched to confirm or refute these predictions, and supporting evidence was found for several reactions. Other SMILEY solutions predicted the addition of completely new reactions to the network. All gap-filing solutions and the most feasible false negative correcting solutions were inspected manually, and for potentially realistic reactions, genes were predicted based on protein sequence homology. The most feasible predicted

reactions cover a wide range of metabolic functions. Many of them corrected false negative predictions for *aspC* knockout strains, which in the model are unable to produce L-aspartate. Several others predicted new reactions involving TCA cycle intermediates. New reactions were also predicted for the metabolism of adenosine 5'-phosphosulfate, dehydroglycine, dTMP, glycoaldehyde, L-isoleucine, and 5-phosphoribosyl-5-carboxyaminoimidazole.

Despite the number of potentially useful predictions made, SMILEY did not find solutions for nearly half of the cases on which it was run. Part of the reason for this is that the universal set of reactions used was based on KEGG [40]. This database only contains reactions that are already known to exist in at least one organism, so completely undiscovered reactions cannot be added. Also, not every metabolite in *i*JO1366 can be connected to KEGG reactions. Of the 1133 compartment-independent metabolites in *i*JO1366, 203 do not have KEGG compound IDs. A lager set of reactions including more model metabolites would allow for additional valid SMILEY solutions to be found. Many gene predictions were made based on sequence homology, but for some reactions, no gene could be predicted because there was no reference sequence available. These are global orphan reactions [44], which have no known gene in any organism. The proliferation of global orphans (estimated to be 30-40% of all known enzymatic functions [14]) makes gene function prediction difficult, and can account for the fact that even in a well-studied organism such as *E. coli*, there are still many uncharacterized genes.

This study utilized a genome-scale metabolic network reconstruction as a tool for the analysis of high-throughput experimental data. The ultimate result of this study is that a number of valuable predictions have been made. Some of these predictions are for

adjustments to the *i*JO1366 model, such as the predicted changes to the core biomass reaction and changes to the GPRs of the reactions *IG3PS*, *I2FE2SS*, *I2FE2SS2*, *S2FE2SS*, and *S2FE2SS2*. Corroborating literature evidence has been found for some of these predictions, so they should be incorporated into future model updates. The other predictions made through this study are for gene functions, both for isozymes and for reactions currently not known to occur in *E. coli*. These predictions provide hypotheses that can be experimentally tested. As an example, the prediction of a missing myo-inositol:oxygen oxidoreductase reaction led to the design of a simple experiment in which the previously uncharacterized gene *yhiJ* was found to be essential for growth on myo-inositol. We expect that many of the other predictions made in this study can likewise serve as hypotheses for experimental analysis.

4.4 Methods

4.4.1 Identifying model gaps with GapFind

The modified version of the GapFind MILP algorithm [21] encoded in the COBRA Toolbox [45, 46] was used in this study. The *i*JO1366 *E. coli* metabolic model [12] in SBML format [47] was loaded into the COBRA Toolbox, and the lower bounds of all exchange reactions were set to -1000 mmol/gDW/h and the upper bounds of all model reactions were set to 10^9 mmol/gDW/h. All other constraints were kept at their default values, as described in **Section 3.4.4 Constraint-based modeling**. The GapFind algorithm was then run twice, once with each mass balance constraint option. Root no-production and no-consumption metabolites were identified from the model **S** matrix by

searching for rows containing only negative or positive coefficients, respectively. Downstream no-production and upstream no-consumption gaps were identified by removing the root gaps from the GapFind outputs. The root gaps of each downstream gap were identified through targeted computational experiments in which metabolite source reactions were added to the network to restore connectivity. Metabolites requiring demand reactions in the *i*JO1366 model were also counted as gaps. The Tomlab (Tomlab Optimization Inc., Seattle, WA) mixed-integer linear programming solver was used with GapFind.

4.4.2 Comparison of model predictions to experimental data

The experimental gene essentiality data was obtained from four publications [12, 23-25], and the "essential" or "non-essential" designations assigned in the original studies were used. Several corrections to the essentiality assignments were made based on an updated analysis of the Keio Collection [33]. The newly identified essential genes were added to the lists of essential genes under all conditions, while the genes whose essentiality was identified as uncertain were not changed from their original designations.

The *i*JO1366 *E. coli* K-12 MG1655 metabolic network reconstruction was adjusted to match the phenotype of *E. coli* BW25113, which is missing several metabolic genes (Δ*araBAD*, Δ*rhaBAD*, Δ*lacZ*). Using the COBRA Toolbox, the associated reactions without isozymes (*ARAI*, *RBK_L1*, *RMPA*, *LYXI*, *RMI*, *RMK*, and *LACZ*) were constrained to carry zero flux. All other model reactions retained their default bounds [12]. Each knockout strain was modeled by using the deleteModelGenes function to constrain the correct reactions to zero. Model growth phenotypes were determined using

FBA with the core biomass reaction as the objective, one at a time on each condition. Strains with growth rates above zero were classified as non-essential, while strains with growth rates of zero were classified as essential.

4.4.3 Computational prediction of gap-filling reactions

The COBRA Toolbox implementation of the SMILEY algorithm (growthExpMatch) was used to predict sets of gap-filling reactions for each false negative model comparison. The universal database of reactions was obtained from KEGG Release 58.0 [40]. All reactions in this set listed as "incomplete reaction" were blacklisted, or excluded from possible SMILEY solutions. Any reaction with the same compound appearing as both a substrate and a product was also blacklisted, along with several reactions identified in initial tests (R00090, R00113, and R00274) as forming unrealistic energy generating reaction loops with existing *i*JO1366 model reactions. The minimum growth threshold required by the SMILEY algorithm was 0.05 h^{-1}. Up to 25 alternate solutions were allowed, with a single solution time limit of 2 h.

When SMILEY was run on gaps instead of false negative cases, each producing or consuming reaction for each gap metabolite was identified from the *i*JO1366 model. A lower bound of 0.01 mmol/gDW/h was applied to each reaction, one at a time, and SMILEY was used to predict gap-filling reactions. For gaps that have demand reactions in the model, the demand reactions were constrained to zero before running SMILEY.

4.4.4 Computational feasibility analysis of all predictions

After predicting sets of false negative correcting and gap-filling reactions, each of these sets of solution reactions was added to the *i*JO1366 model one at a time. The growth phenotype of each of these strains on all 13,470 experimental data conditions was then predicted using FBA, with a threshold of zero for determining growth or no-growth. The number of new false positives for each solution was determined from the number of conditions that were true negatives with the wild-type model but could grow when the new reactions were added. The number of corrected false negatives for each solution was the number of false negatives that became true positives when the new reactions were added. GapFind was also run on the *i*JO1366 model with each set of solution reactions added to it, one at a time. The set of network gaps was compared to the set of gaps in the original model to determine if any gaps were eliminated.

In order to determine which SMILEY solutions could be tested with simple *in vivo* experiments, FBA was used to test growth of the *i*JO1366 model with each set of solution reactions on a set of 115 carbon sources and 62 nitrogen sources under both aerobic and anaerobic conditions. The growth of the unmodified *i*JO1366 model was first tested on each condition using FBA. Next, the model with each set of solution reactions added, one at a time, was tested on all 3896 conditions on which the unmodified model predicted no growth. Conditions on which the modified versions of the model could grow were used to design experiments.

4.4.5 Experimental validation of predicted genes

Candidate genes for SMILEY predicted reaction sets were predicted using bi-directional protein BLAST (BLASTp) between a gene from another organism in KEGG and the *E. coli* K-12 MG1655 genome. Protein sequences from organisms that are phylogenically close to *E. coli* were used when possible. The gene with the highest BLAST expectation value (E-value) found was reported. When multiple genes were found with E-values below 10^{-13}, all were reported as candidate genes.

To identify potential reactions for growth on different combinations of nitrogen containing substrates, ten different types of minimal media were made. Each media contained 2 g/L of the main substrate along with M9 salts (6.8 g/L sodium phosphate dibasic, 3.0 g/L potassium phosphate monobasic, 0.5 g/L sodium chloride, 0.24 g/L magnesium sulfate, 0.011 g/L calcium chloride), trace elements (0.1 g/L iron (III) chloride, 0.02 g/L zinc sulfate, 0.004 g/L copper chloride, 0.01 g/L manganese sulfate, 0.006 g/L cobalt chloride, 0.006 g/L disodium EDTA), and Wolfe's Vitamin Solution. Unlike the standard M9 formulation, no ammonium chloride was added. The ten primary substrates were adenine, L-arginine, L-cysteine, glycine, hypoxanthine, ornithine, L-proline, putrescence, D-glucose, and ammonium chloride. Each media was made anaerobic by bubbling with N_2 gas for 30 min, and was filter sterilized. The anaerobic media was stored in an anaerobic chamber (Coy, Grass Lake, MI). 100 µL of each type of media was dispensed into the wells of a 96-well plate so that every combination of media types was added to a well, with a total of 200 µL of media in each well. 200 µL of each media alone was also dispensed into wells. Control wells contained M9 media with no substrate added.

A single colony of wild-type *E. coli* K-12 MG1655 was inoculated into 25 mL LB media and grown overnight at 37°C aerobically. The next day, 15 mL of the culture was centrifuged at 4000 rpm for 8 min, the supernatant was discarded, and the culture was resuspended in an M9 salt solution with no carbon or nitrogen sources. The culture was centrifuged and resuspended in new M9 four more times to completely wash out all LB. 1 μL of washed *E. coli* culture was then used to inoculate each well of the 96-well plate. The plate was sealed with a Microseal 'B' adhesive seal (Bio-Rad, Hercules, CA) and grown for 48 h at 37°C in the anaerobic chamber. Absorbance at 600 nm in each well was then measured with a VERSAmax microplate reader (Molecular Devices, Sunnyvale, CA).

To test the growth of *E. coli* with myo-inositol as a carbon and energy source, 2 g/L myo-inositol M9 media was made and filter sterilized. Wild-type *E. coli* along with four strains from the Keio Collection with *yphC* (JW5842), *yfdE* (JW2368), *ydeN* (JW5243), and *yhiJ* (JW3455) gene knockouts (supplied by Open Biosystems) were grown overnight in LB and washed four times in M9 as described above. 1 μL of washed *E. coli* cultures were then used to inoculate 10 mL aerobic myo-inositol M9 cultures, which were grown aerobically at 37°C. The optical density at 600 nm was measured at several points during growth.

Enzymes for *in vitro* assays were expressed and purified from the ASKA Library of histidine tagged *E. coli* ORF clones [48]. *E. coli* AG1 cells with the appropriate vector were grown in LB media and gene expression was induced using the Overnight Express Autoinduction System (EMD Biosciences, Darmstadt, Germany). The cells were then lysed and the His-tagged proteins were purified using the MagneHis Protein Purification

System (Promega, Fitchburg, WI), with the proteins eluted on MagneHis Ni-Particles. The purified proteins remained attached to the beads during subsequent assays.

Prior to each experiment, the beads were washed twice with a buffer consisting of 50 mM tricine, 20 mM potassium chloride, 5 mM magnesium chloride, 1 mM manganese chloride, and 0.1 mM calcium chloride. *In vitro* reactions were carried out in a buffer consisting of 50 mM tricine, 20 mM potassium chloride, 5 mM magnesium chloride, 1 mM manganese chloride, 0.1 mM calcium chloride, 0.05% Tween-20, 1 mM MES, and 0.1 mM MS. For some reactions, iron sulfate or zinc sulfate were added to the buffer. 2 μL beads with enzyme were added to each 500 μL reaction mixture. Potential substrates were added at concentrations of 100 mM, with cofactors such as NADH or acetyl-CoA added at concentrations of 10 mM. Reactions were run at 37°C overnight with vigorous shaking. After running the *in vitro* reactions, the metabolites were measured with an Agilent LC-MS (Agilent Technologies, Santa Clara, CA) in SIM mode, in both positive and negative modes as required by the known substrates and expected products. Each reaction sample was measured three times. LC-MS data was analyzed with the MathDAMP software package [49], and metabolite concentrations were normalized to concentrations of MS (in positive mode) or MES (in negative mode). Statistical significance of metabolite concentration differences was assessed with t-tests.

Acknowledgements

Chapter 4 is, in part, adapted from a paper that appeared in Molecular Systems Biology, Volume 7, Article Number 535, October 11, 2011. The dissertation author was

the primary author of this paper, which was coauthored by Tom M. Conrad, Jessica Na, Joshua A. Lerman, Hojung Nam, Adam M. Feist, and Bernhard Ø. Palsson.

Chapter 4 is also, in part, adapted from a paper that is being prepared for publication under the title "Gap-filling analysis of the *i*JO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions." The dissertation author was the primary author of this paper, which was coauthored by Bernhard Ø. Palsson.

We would like to thank Pep Charusanti, Tom Conrad, Adam Feist, Harish Nagarajan, Kenji Nakahigashi, Vasiliy Portnoy, Jennie Reed, and Ines Thiele for their helpful comments and insights. I would especially like to thank Professor Martin Robert for hosting me at the Institute for Advanced Biosciences in Tsuruoka, Japan and mentoring me while I performed *in vitro* enzyme assays and LC-MS analysis.

References

1. Edwards, J.S., and Palsson, B.O., *Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics*, 2000. **1**(1).

2. Edwards, J.S., M. Covert, and B. Palsson, *Metabolic modeling of microbes: the flux-balance approach.* Environmental Microbiology, 2002. **4**(3): p. 133-40.

3. Price, N.D., et al., *Genome-scale microbial in silico models: the constraints-based approach.* Trends in Biotechnology, 2003. **21**(4): p. 162-169.

4. Karsten, W.E. and R.E. Viola, *Kinetic studies of L-aspartase from Escherichia coli: pH-dependent activity changes.* Arch Biochem Biophys, 1991. **287**(1): p. 60-7.

5. Habrych, M., S. Rodriguez, and J.D. Stewart, *Purification and identification of an Escherichia coli beta-keto ester reductase as 2,5-diketo-D-gluconate reductase YqhE.* Biotechnol Prog, 2002. **18**(2): p. 257-61.

6. Choi-Rhee, E. and J.E. Cronan, *A nucleosidase required for in vivo function of the S-adenosyl-L-methionine radical enzyme, biotin synthase.* Chem Biol, 2005. **12**(5): p. 589-93.

7. Portnoy, V.A., M.J. Herrgard, and B.O. Palsson, *Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient Escherichia coli strain.* Appl Environ Microbiol, 2008. **74**(24): p. 7561-9.

8. Klem, T.J. and V.J. Davisson, *Imidazole glycerol phosphate synthase: the glutamine amidotransferase in histidine biosynthesis.* Biochemistry, 1993. **32**(19): p. 5177-86.

9. Orth, J.D., I. Thiele, and B.Ø. Palsson, *What is flux balance analysis?* Nat Biotechnol, 2010. **28**(3): p. 245-8.

10. Yum, D.Y., B.Y. Lee, and J.G. Pan, *Identification of the yqhE and yafB genes encoding two 2, 5-diketo-D-gluconate reductases in Escherichia coli.* Appl Environ Microbiol, 1999. **65**(8): p. 3341-6.

11. MacKintosh, C. and H.G. Nimmo, *Purification and regulatory properties of isocitrate lyase from Escherichia coli ML308.* Biochem J, 1988. **250**(1): p. 25-31.

12. Orth, J.D., et al., *A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011.* Mol Syst Biol, 2011. **7**: p. 535.

13. Reed, J.L., et al., *Systems approach to refining genome annotation.* Proc Natl Acad Sci U S A, 2006. **103**(46): p. 17480-4.

14. Karp, P.D., *Call for an enzyme genomics initiative.* Genome Biol, 2004. **5**(8): p. 401.

15. Pouliot, Y. and P.D. Karp, *A survey of orphan enzyme activities.* BMC Bioinformatics, 2007. **8**: p. 244.

16. Orth, J.D. and B.Ø. Palsson, *Systematizing the generation of missing metabolic knowledge.* Biotechnol Bioeng, 2010. **107**(3): p. 403-12.

17. Reed, J.L., et al., *An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR).* Genome Biology, 2003. **4**(9): p. R54.1-R54.12.

18. Bochner, B.R., P. Gadzinski, and E. Panomitros, *Phenotype microarrays for high-throughput phenotypic testing and assay of gene function.* Genome Res, 2001. **11**(7): p. 1246-55.

19. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.

20. Zomorrodi, A.R. and C.D. Maranas, *Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data.* BMC Syst Biol, 2010. **4**: p. 178.

21. Satish Kumar, V., M.S. Dasika, and C.D. Maranas, *Optimization based automated curation of metabolic reconstructions.* BMC Bioinformatics, 2007. **8**: p. 212.

22. Kumar, V.S. and C.D. Maranas, *GrowMatch: an automated method for reconciling in silico/in vivo growth predictions.* PLoS Comput Biol, 2009. **5**(3): p. e1000308.

23. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol Syst Biol, 2006. **2**: p. 2006.0008.

24. Ito, M., T. Baba, and H. Mori, *Functional analysis of 1440 Escherichia coli genes using the combination of knock-out library and phenotype microarrays.* Metab Eng, 2005. **7**(4): p. 318-27.

25. Joyce, A.R., et al., *Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli.* J Bacteriol, 2006. **188**(23): p. 8259-8271.

26. Thiele, I., et al., *Genome-scale reconstruction of Escherichia coli's transcriptional and translational machinery: a knowledge base, its mathematical formulation, and its functional characterization.* PLoS Comput Biol, 2009. **5**(3): p. e1000312.

27. Falkenberg, P. and A.R. Strom, *Purification and characterization of osmoregulatory betaine aldehyde dehydrogenase of Escherichia coli.* Biochim Biophys Acta, 1990. **1034**(3): p. 253-9.

28. Molina-Henares, M.A., et al., *Identification of conditionally essential genes for growth of Pseudomonas putida KT2440 on minimal medium through the screening of a genome-wide mutant library.* Environ Microbiol, 2010. **12**(6): p. 1468-85.

29. Liu, D. and P.R. Reeves, *Escherichia coli K12 regains its O antigen.* Microbiology, 1994. **140 (Pt 1)**: p. 49-57.

30. Rubires, X., et al., *A gene (wbbL) from Serratia marcescens N28b (O4) complements the rfb-50 mutation of Escherichia coli K-12 derivatives.* J Bacteriol, 1997. **179**(23): p. 7581-6.

31. Colnaghi, R., et al., *Properties of the Escherichia coli rhodanese-like protein SseA: contribution of the active-site residue Ser240 to sulfur donor recognition.* FEBS Lett, 2001. **500**(3): p. 153-6.

32. Adams, H., et al., *PspE (phage-shock protein E) of Escherichia coli is a rhodanese.* FEBS Lett, 2002. **518**(1-3): p. 173-6.

33. Yamamoto, N., et al., *Update on the Keio collection of Escherichia coli single-gene deletion mutants.* Mol Syst Biol, 2009. **5**: p. 335.

34. Cusa, E., et al., *Genetic analysis of a chromosomal region containing genes required for assimilation of allantoin nitrogen and linked glyoxylate metabolism in Escherichia coli.* J Bacteriol, 1999. **181**(24): p. 7479-84.

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

36. Lin, S., R.E. Hanson, and J.E. Cronan, *Biotin synthesis begins by hijacking the fatty acid synthetic pathway.* Nat Chem Biol, 2010. **6**(9): p. 682-8.

37. Bettendorff, L. and P. Wins, *Thiamin diphosphate in biological chemistry: new aspects of thiamin metabolism, especially triphosphate derivatives acting other than as cofactors.* FEBS J, 2009. **276**(11): p. 2917-25.

38. Patton, A.J., et al., *Does Escherichia coli possess a second citrate synthase gene?* Eur J Biochem, 1993. **214**(1): p. 75-81.

39. Gerike, U., et al., *Citrate synthase and 2-methylcitrate synthase: structural, functional and evolutionary relationships.* Microbiology, 1998. **144 (Pt 4)**: p. 929-35.

40. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010. **38**(Database issue): p. D355-60.

41. Keseler, I.M., et al., *EcoCyc: a comprehensive view of Escherichia coli biology.* Nucleic Acids Res, 2009. **37**(Database issue): p. D464-70.

42. Howell, D.M., H. Xu, and R.H. White, *(R)-citramalate synthase in methanogenic archaea.* J Bacteriol, 1999. **181**(1): p. 331-3.

43. Rao, M.R., et al., *Enzymatic hydration of citraconate to (minus)citramalate.* Biochem Biophys Res Commun, 1963. **12**: p. 78-82.

44. Osterman, A. and R. Overbeek, *Missing genes in metabolic pathways: a comparative genomics approach.* Curr Opin Chem Biol, 2003. **7**(2): p. 238-51.

45. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox.* Nat. Protocols, 2007. **2**(3): p. 727-738.

46. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox 2.0.* Nat Protoc, 2011.

47. Hucka, M., et al., *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.* Bioinformatics, 2003. **19**(4): p. 524-31.

48. Kitagawa, M., et al., *Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research.* DNA research : an international journal for rapid publication of reports on genes and genomes, 2005. **12**(5): p. 291-9.

49. Baran, R., et al., *MathDAMP: a package for differential analysis of metabolite profiles.* BMC Bioinformatics, 2006. **7**: p. 530.

Chapter 5: Metabolic engineering of *Escherichia coli* I: Design of growth-coupled production strains with constraint-based modeling

One of the most promising practical applications of constraint-based modeling of metabolic networks is in metabolic engineering. Models can be used to predict the systemic effects of gene or reaction knockouts, and can predict which fermentation products must be produced for a strain to achieve its optimal growth rate. In this chapter, the *i*AF1260b *Escherichia coli* metabolic model was used to design growth-coupled production strains that are predicted to produce high yields of useful chemicals. First, three substrates were chosen for analysis, along with 12 target compounds. The OptKnock algorithm was then used to predict strain designs for each combination of substrate condition and product, successfully finding growth-coupled designs for many combinations. Next, the OptGene algorithm was used to improve these growth-coupled designs and to identify designs for more combinations. To validate the growth-coupled design process, two strains were chosen for experimental analysis. One of these strains was predicted to produce lactate, while the other was predicted to produce R-1,2-propenediol.

5.1 Introduction

Metabolic engineering is the use of genetic engineering methods to alter the metabolic networks of organisms in beneficial ways, enabling the production of useful chemicals. As prices of petroleum-based chemicals continue to increase, the production of chemicals by microbial organisms is becoming an increasingly popular alternative [1, 2]. The potential use of renewable feedstocks such as sugars as substrates combined with efficient conversion are some of the environmentally friendly advantages of bioprocessing-based production of chemicals over traditional chemical synthesis strategies. Commodity chemicals [2], biofuels [3], and pharmaceuticals [4, 5] are some of the classes of chemicals that are being studied for production by engineered microbial strains. Engineering strategies commonly employed include the introduction of exogenous pathways to add new capabilities to an organism, upregulating genes for production pathways to increase flux through them, elimination of competing pathways, and methods to increase the robustness of strains in an industrial fermentation setting [6-10]. Typically, multiple strategies must be employed to develop strains capable of efficient production on a large scale [11, 12].

In recent years, constraint-based modeling of genome-scale metabolic networks has become an important tool for metabolic engineering [7, 13, 14]. Computational modeling is useful because it can allow for system-wide effects of gene knockouts or overexpression of particular pathways to be predicted [15]. It can also be useful for predicting the interacting effects of multiple simultaneous interventions. Depending on the modeling methodology used, potentially thousands of possible strain designs can be screened *in silico*, with only the most promising designs to be investigated

experimentally. To date, constraint-based computational modeling has mostly been used in combination with other metabolic engineering design strategies. Studies have been published in which modeling has been used to design strains for the production of lactate [16, 17], ethanol [18], succinate [19, 20], lycopene [21, 22], L-threonine [23], L-valine [12], and 1,4-butanediol [11], among many others.

One of the most promising and innovative uses of modeling in designing metabolic engineering strains is in the design of growth-coupled strains. In a standard strain design, the production of the target compound comes at the cost of production of biomass. If mutations accumulate that increase the growth rate of the strain, they must do so by decreasing yield of the desired product. Unless very stringent quality control measures are employed, the yields of these strains will decrease over time as they evolve to higher growth rates. In a growth-coupled strain, on the other hand, evolution under selection for highest growth rate will have the opposite effect. A growth-coupled product is one that must be produced for a cell to grow at its maximum rate, so growth increasing mutations will need to increase production of the targeted chemical as well.

Many different computational algorithms and procedures have been developed to design growth-coupled production strains. In this study, the algorithms OptKnock [24] and OptGene [25] were used to design growth-coupled *Escherichia coli* strains. OptKnock is a bi-level optimization algorithm. It searches for sets of reaction knockouts that lead to a maximum production rate of a defined target compound at the maximum growth rate of the strain. OptGene uses a genetic algorithm to identify and evolve strain designs with a user defined objective function. It is possible to use more complicated

nonlinear objectives with OptGene, but it is not guaranteed to identify a global optimal solution.

To characterize the metabolic engineering potential of *E. coli*, both of these algorithms were used to assemble a large set of growth-coupled strain designs using the *i*AF1260b constraint-based model [26]. First, random knockout simulations were used to identify some of the common fermentation products that can be produced by *E. coli*. Based partly on these results, a set of 12 target compounds were selected. OptKnock was used to predict three and five knockout strain designs for these twelve products on five different substrate conditions. These strain designs were then used as starting strains for OptGene, which was run with three different objective functions to identify many different strain designs. Finally, two strains were chosen for experimental validation of model predicted phenotypes. The actual construction and experimental analysis of these strains is presented in **Chapter 6: Metabolic Engineering of *Escherichia coli* II: Adaptive Evolution and Phenotypic Characterization of Computationally Designed Strains**.

5.2 Results

5.2.1 Selection of targeted substrates and products

In order to design growth-coupled *E. coli* production strains, it was first necessary to determine which compounds to target for production and which substrates to produce them from. Hexoses and pentoses are considered good potential substrates for industrial bioprocessing because they are widely available from renewable plant by-products.

Glucose is a commonly used hexose for bacterial growth media, and xylose is a widely-available pentose that is also commonly used in metabolic engineering studies [27]. The third substrate selected for study was glycerol, a three-carbon compound that is a common by-product of industrial processes such as biodiesel production.

E. coli can potentially produce hundreds of different chemicals, and the *i*AF1260b model contains 1039 different metabolites [26]. To identify the most feasible metabolites for growth-coupled production, a new constraint-based algorithm was developed. This algorithm, called RandKnock, uses flux balance analysis (FBA) to predict the phenotypes of strains with random sets of gene knockouts, and identifies all fermentation products produced at the maximum growth rate. RandKnock was run on the *i*AF1260b *E. coli* model constrained for growth on glucose under anaerobic conditions for random sets of three reactions knocked out at a time. RandKnock was run 1,000,000 times under these conditions, and the resulting growth rates and optimal fermentation products were recorded. These random genotype predictions were repeated for sets of four through ten simultaneous knockouts, also 1,000,000 times each. RandKnock was also run on *E. coli* for growth on xylose under anaerobic conditions. For each condition and number of simultaneous knockouts, the number of times each compound occurred as a fermentation product was counted, and the production rate and growth rate of the strain that gave the highest substrate specific productivity (SSP, production rate multiplied by growth rate) were reported. The results of the five and ten reaction RandKnock runs for growth on glucose are reported in **Table 5.1**.

In general, the number of different products that could be produced increased as the number of simultaneous knockouts increased. With only three knockouts at a time,

Table 5.1 Results of 1,000,000 RandKnock screens for five and ten reaction knockouts.

5 reaction KOs			
Product name	**Number of occurrences**	**Production rate (mmol/gDW/h)**	**Growth rate (h^{-1})**
Succinate	987331	18.27	0.28
Glycolate	985969	0.00	0.46
Ethanol	956770	16.33	0.46
Formate	954783	34.90	0.46
Acetate	953311	16.63	0.46
D-Lactate	45092	8.39	0.00
KDO-2-lipid-IV-A	32845	0.28	0.25
R-1,2-Propanediol	7588	2.04	0.21
Dihydroxyacetone	2826	19.81	0.19
Glycine	1005	1.26	0.28
Pyruvate	576	14.22	0.35
Fumarate	27	0.32	0.45
L-Valine	22	18.88	0.12
Acetaldehyde	17	36.24	0.24
Hydrogen sulfide	9	2.12	0.19
L-Alanine	6	17.58	0.13
D-Alanine	1	0.00	0.12

10 reaction KOs			
Product name	**Number of occurrences**	**Production rate (mmol/gDW/h)**	**Growth rate (h^{-1})**
Succinate	940158	18.27	0.28
Glycolate	938072	0.00	0.46
Ethanol	893735	16.33	0.46
Formate	881235	34.90	0.46
Acetate	876934	16.63	0.46
D-Lactate	115009	34.58	0.32
KDO-2-lipid-IV-A	60000	0.28	0.25
R-1,2-Propanediol	28777	11.90	0.27
Dihydroxyacetone	17984	6.10	0.00
Glycine	4086	1.26	0.28
Pyruvate	3381	17.33	0.33
L-Alanine	393	37.57	0.13
Acetaldehyde	332	36.23	0.24
Hydrogen sulfide	268	1.07	0.07
Fumarate	240	0.32	0.45
L-Valine	212	4.90	0.18
D-Alanine	7	0.01	0.36
Glycerol	3	5.59	0.00
N-Acetyl-D glucosamine(anhydrous)N-Acetylmuramic acid	2	0.03	0.25
L-Malate	1	0.30	0.42

there were 13 different compounds produced, not including non-organic compounds such as CO_2, H_2O, and H^+. With five knockouts, there were 17 different products, and with ten knockouts, there were 20 different products. The most commonly produced products were the natural products of glucose fermentation by wild-type *E. coli*: ethanol, acetate, formate, and succinate. These compounds, along with glycolate production at a very low level, were produced in over 95% of random strains with five reaction knockouts. They were less common in strains with ten knockouts because lethal sets of knockouts were more common in these strains. The most common non-wild-type products include D-lactate, R-1,2-propanediol, dihydroxyacetone, and pyruvate. Several different amino acids appeared among the RandKnock results, including glycine, L-alanine, and L-valine. Compounds not found among the RandKnock results are likely more difficult to design growth-coupled production strains for. In other words, there are likely to be fewer possible genotypes that lead to secretion of these compounds. RandKnock results for growth on xylose predicted similar sets of fermentation products.

5.2.2 Computational methods for predicting strain designs

Based partly on the results of the RandKnock analysis, a set of 12 target compounds was selected for investigation. These compounds were chosen to represent different metabolic pathways and types of products, thus illuminating the capabilities of *E. coli* for producing different types of potentially useful compounds. The 12 compounds investigated were R-1,2-propanediol, L-alanine, ethanol, fumarate, L-glutamate, glycerol, D-lactate, L-malate, 2-oxoglutarate, pyruvate, L-serine, and succinate. Knockout strain designs were predicted for these targets under five different conditions: growth on

glucose both aerobically and anaerobically, growth on xylose both aerobically and anaerobically, and growth on glycerol aerobically. Analysis of maximum potential production of compounds from glycerol anaerobically indicates that most products have only very low possible yields, so no designs under these conditions were attempted.

Two algorithms were used for design of growth-coupled knockout strains: OptKnock and OptGene. OptKnock, a bi-level optimization algorithm that seeks the set of reaction knockouts that simultaneously maximizes both product secretion and growth rate, was used first. The growth objective function used by OptKnock was modified to avoid non-unique growth-coupled solutions. As illustrated in **Figure 5.1 a**, there are three possible couplings between a compound that can be secreted and cell growth rate. Completely non-growth-coupled products reduce growth rate when they are secreted, and thus have a secretion rate of zero at the maximum growth rate of the strain. On the other hand, growth-coupled products tend to increase growth rate as they are secreted at a higher rate, and thus must be produced at a nonzero rate at the maximum growth rate. The third type of coupling is the case of non-uniquely growth-coupled products. These products have a variable secretion rate at the maximum growth rate, and can be secreted at a positive rate or at a rate of zero. This can occur when two or more different products can contribute equally to optimal growth, and any of them can be produced alternatively at the maximum growth rate. These non-uniquely growth-coupled strains are not desirable because there is no way to predict which product would actually be produced by the strain *in vivo*. To avoid these solutions, the OptKnock growth objective function was "tilted," including minimization of the target compound outer membrane transport reaction as a small component. This causes the alternate solution with the lowest possible

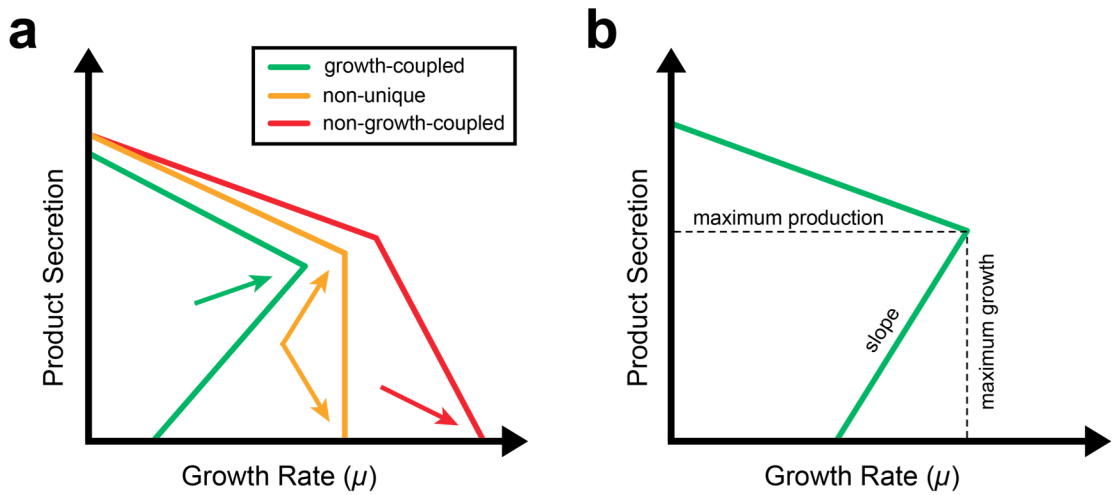

Figure 5.1 (**a**) The three different types of growth-coupling solutions that are possible. (**b**) Parameters used in the alternative objective functions of OptGene and AnalyzeGCdesign.

secretion rate at the maximum growth rate to be identified by OptKnock as the phenotype of a particular strain. Thus, non-uniquely growth-coupled strains appear to have a production rate of zero, and will not be identified as optimal solutions by OptKnock.

Before running OptKnock, the *i*AF1260b model and set of candidate knockout reactions were modified to increase the numerical stability of the optimization calculations, to reduce computation time, and to eliminate potential unrealistic solutions. A condition specific reduced metabolic model was made for each of the five substrate conditions. Specifically, flux variability analysis (FVA) [28] was used to determine which reactions were blocked under each set of conditions, and these reactions were removed from the reduced models. Then the upper and lower bounds on all reactions were constrained to their actual maximum and minimum values, as constrained by the substrate uptake rates. Next, the set of selected reaction knockout candidates was formed from the complete list of reactions in *i*AF1260b. All essential reactions, both experimentally and computationally determined, were removed from this set, along with

all orphan and spontaneous reactions. These reactions are not good candidates because they cannot actually be knocked out *in vivo*. Reactions for pathways not likely to be involved in production of the twelve target compounds, such as those for the synthesis of lipids and cell wall components, were also excluded, as were all reactions involving compounds with more than seven carbon atoms. Transport reactions were also excluded due to the low specificity of many transport proteins. Finally, the coupled reaction sets in each reduced model were identified, and only one reaction from each set was included in the final set of selected reactions. Each reaction knockout in a co-set would have the same effect on the network.

After forming the reduced models and selected reactions, OptKnock was used to identify strain designs for all 12 products on all five substrate conditions. Both three and five knockout design strategies were pursued. Intermediate solutions were also generated for ten simultaneous knockouts. Growth-coupled strain designs were successfully identified for L-alanine, ethanol, fumarate, glycerol, D-lactate, 2-oxoglutarate, pyruvate, and succinate. Strain designs for most products were found for both glucose and xylose, aerobically and anaerobically. Few OptKnock designs were found for growth on glycerol. Designs for glycerol production could only be found for glucose aerobic conditions, and 2-oxoglutarate designs could only be found for anaerobic conditions. A total of 679 different growth-coupled OptKnock strain designs were identified.

Next, the OptGene genetic algorithm was used to design more growth-coupled strains. OptKnock solutions were used as starting strains for OptGene, which was set to identify strain designs with a maximum of ten reaction knockouts. OptGene was run using three different objective functions, two of which were nonlinear. The product yield

objective function was defined as the target production rate divided by the substrate uptake rate (a constant in these computations). Substrate specific productivity (SSP) was defined as the production rate multiplied by the growth rate of a strain. Strength of coupling (SOC) was defined as the product yield squared divided by the slope between the point at the maximum growth rate with the lowest production rate and the point at the minimum production rate with the highest growth rate (**Figure 5.1 b**). A steeper slope indicates a lower degree of growth-coupling, and is less desirable. For all three objective functions used, improvements to the OptKnock designed strains could be found. By starting OptGene from random knockout sets, new growth-coupled designs were also found for several substrate/product combinations for which OptKnock solutions could not be found. Ultimately, growth-coupled designs were identified for 36 substrate/product combinations.

5.2.3 Selection of strains for *in vivo* construction

Based on the results of the OptKnock and OptGene screens, two strains were selected for *in vivo* construction to verify the utility of model-based growth-coupled strain design predictions. Several properties of the designed strains were considered in selecting strains for further analysis. High yield of the target compound is clearly desirable, as is a low level or even no secretion of by-product compounds. Production of additional compounds will necessarily reduce the total yield of the desired product, as substrate mass would need to be diverted to form by-products. In an industrial production strain, secretion of a single product will also be advantageous because it reduces the number of fermentation products which must be separated. A high optimal growth rate

compared to the substrate uptake rate is also desired. Because *in vivo* constructed growth-coupled strains must undergo adaptive laboratory evolution (ALE) to reach their optimal phenotype, a high growth rate will allow this phenotype to be reached faster. In addition, strength of growth-coupling can also affect the ability of a strain to evolve to its optimal phenotype. It is also desirable for a constructed strain to be able to utilize a variety of substrates, for example, both glucose and xylose. Finally, because the *i*AF1260b model was constructed using a variety of data sources, the confidence levels for different reactions and pathways can vary [26]. It is thus desirable for a production strain to utilize well-studied pathways in which the evidence for the necessary reactions is strong. If use of a poorly characterized pathway is required, the actual enzymes may not behave as predicted in the model, or may not function at all, and the knockout strain would not perform as predicted. All of these properties were assessed for potential strains for *in vivo* construction, though no single strain fully satisfied every condition.

To help analyze and improve potential designs for construction, a new algorithm called AnalyzeGCdesign was developed. It was used to improve strain designs by iteratively changing their sets of gene knockouts, while also adding or removing knockouts. One change at a time was made, and the change that increased the value of one of eight possible objective functions was selected. AnalyzeGCdesign is a recursive function, and continued to update the strain designs until no further improvements could be made. The eight objective functions included different combinations of yield, SSP, and SOC. Some of the objectives also included penalties for extra reaction knockouts. In these cases, a higher total number of knockout reactions led to a lower objective function value unless the knockouts actually contributed to the production phenotype. This

ensured that unnecessary knockouts were identified and removed from the strain designs. AnalyzeGCdesign was run on several dozen OptKnock and OptGene strain designs, and ultimately two designs were selected for experimental validation (**Figure 5.2**).

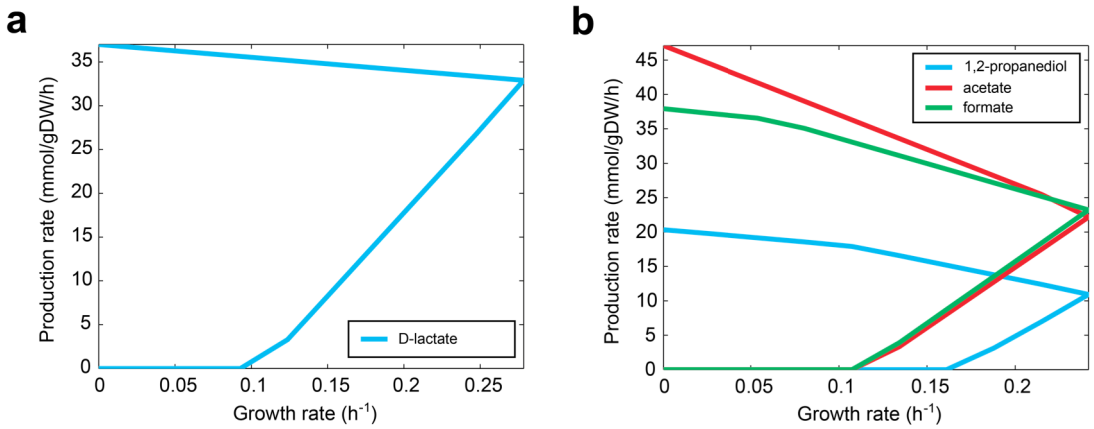

Figure 5.2 The predicted production envelopes with a glucose uptake rate of 20 mmol/gDW/h of (**a**) a growth-coupled lactate producing strain and (**b**) a growth-coupled 1,2-propanediol producing strain.

5.2.4 Predicted properties of lactate production strain

The first strain selected for construction was for the production of D-lactate from glucose under anaerobic conditions. 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 is a biodegradable plastic [29]. This design calls for knockouts of the reactions pyruvate dehydrogenase (*PDH*) and pyruvate formate lyase (*PFL*). Pyruvate dehydrogenase is catalyzed by a complex of *aceE* (b0114), *aceF* (b0115), and *lpdA* (b0116) [30]. Pyruvate formate lyase can be catalyzed either by a complex of *pflA* (b0902) and *pflB* (b0903) or by a complex of *pflC* (b3952) and *pflD* (b3951) [31, 32]. Pyruvate dehydrogenase and pyruvate formate lyase are the two central metabolic

reactions that convert pyruvate to acetyl-CoA. This unintuitive set of knockouts may appear to be lethal as it cuts off production of the essential precursor to fatty acids, acetyl-CoA. However, use of FBA with the *i*AF1260b model predicts that alternative pathways involving nucleotide recycling reactions can be used as a source of acetyl-CoA. The pyruvate dehydrogenase and pyruvate formate lyase knockouts block the synthesis of the wild-type fermentation products ethanol, formate, and acetate, leaving D-lactate as the optimal product (**Figure 5.3**). Lactate is produced by the well-characterized reaction D-lactate dehydrogenase (*LDH_D*), catalyzed by *ldhA* (b1380) [33]. Under certain conditions, wild-type *E. coli* can use this enzyme to produce and secrete lactate as a fermentation product [34]. Succinate is also predicted to be produced by this strain, but at a very low rate. The predicted optimal anaerobic growth rate with a glucose uptake rate of 20 mmol/gDW/h is 0.31 h^{-1}.

A growth-coupled lactate production strain with a similar design to this one was recently constructed and confirmed to produce lactate with a very high yield [Feist et al., in preparation]. This strain required supplementation with a high level of yeast extract during *in vivo* growth, while the strain designed in this study is predicted to grow in unsupplemented glucose minimal media. The previous design also included extra gene knockouts for several transport reactions and the *mutS* (b2733) DNA repair gene. Still, the fact that this strain worked well, as have other designs for lactate producing *E. coli* strains, indicates that the present strain is very likely to behave as predicted *in vivo*.

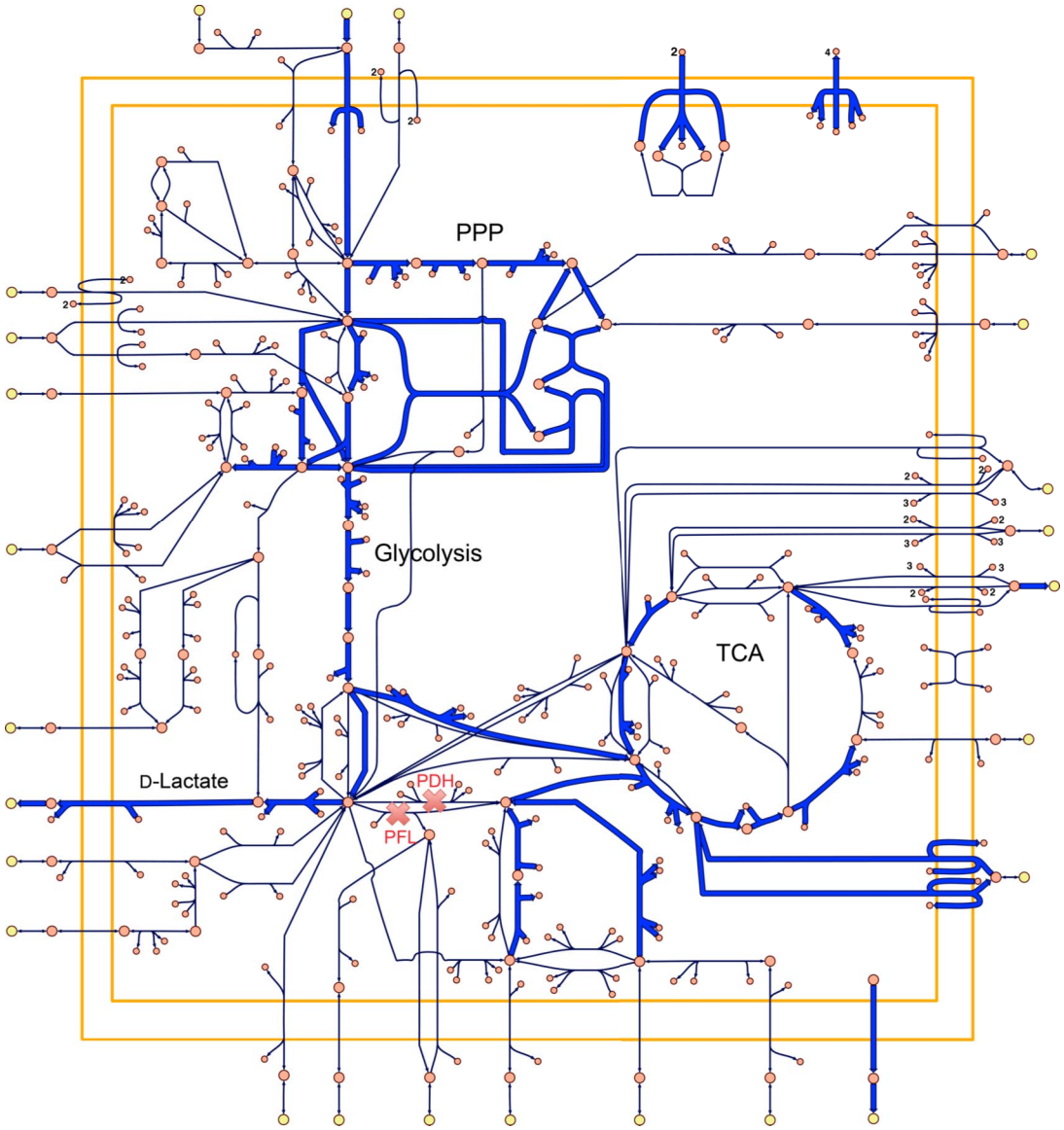

Figure 5.3 The predicted flux distribution for the *PFL PDH* knockout lactate production strain. Without these two routes to acetyl-CoA, *E. coli* is not able to produce ethanol or acetate efficiently, and is forced to produce another fermentation product. Pyruvate is produced by glycolysis, and is then converted to lactate by the D-Lactate Dehydrogenase reaction.

5.2.5 Predicted properties of 12PDO production strain

The second strain selected for *in vivo* construction and validation was for the production of R-1,2-propanediol (12PDO) from glucose anaerobically. This strain design

was also predicted to yield 12PDO from xylose. 12PDO, also commonly known as propylene glycol, is used as a solvent in pharmaceuticals and cosmetics, as an anti-freeze and de-icing compound, and as a precursor for many other chemicals [29, 35]. The strain design selected for this product has knockouts of alcohol dehydrogenase (*ALCD2x*), catalyzed by *adhE* (b1241) [36]; D-lactate dehydrogenase (*LDH_D*), catalyzed by *ldhA* (b1380) [33]; acetaldehyde dehydrogenase (*ACALD*), catalyzed by either *adhE* or *mhpF* (b0351) [37]; and malate dehydrogenase (*MDH*), catalyzed by *mdh* (b3236) [38]. These knockouts block the most efficient production routes of the more desirable fermentation products ethanol, lactate, and succinate (**Figure 5.4**). With these compounds unavailable by their normal production routes, 12PDO becomes the next optimal product. The acetaldehyde dehydrogenase reaction knockout is not essential for 12PDO production, but it improves the strength of growth-coupling by blocking synthesis of acetaldehyde, a suboptimal fermentation product that only slightly reduces the maximum growth rate when produced instead of 12PDO. FBA predicts that this strain will produce acetate and formate in addition to 12PDO, so it is not a homofermentor, but no designs that produce only 12PDO could be found. An optimal growth rate of 0.27 h^{-1} with a glucose uptake rate of 20 mmol/gDW/h is predicted.

12PDO is predicted to be synthesized by a pathway that begins with the glycolytic intermediate dihydroxyacetone phosphate. This compound is converted to the toxic intermediate methylglyoxal by methylglyoxal synthase (*MGSA*), which is then converted to D-lactaldehyde by D-lactaldehyde:NAD+ 1-oxidoreductase (*LALDO2x*), and then to 12PDO by lactaldehyde reductase (*LCARR*). An alternate route starting from methylglyoxal uses aldose reductase (methylglyoxal) (*ALR2*) to form acetol, which is

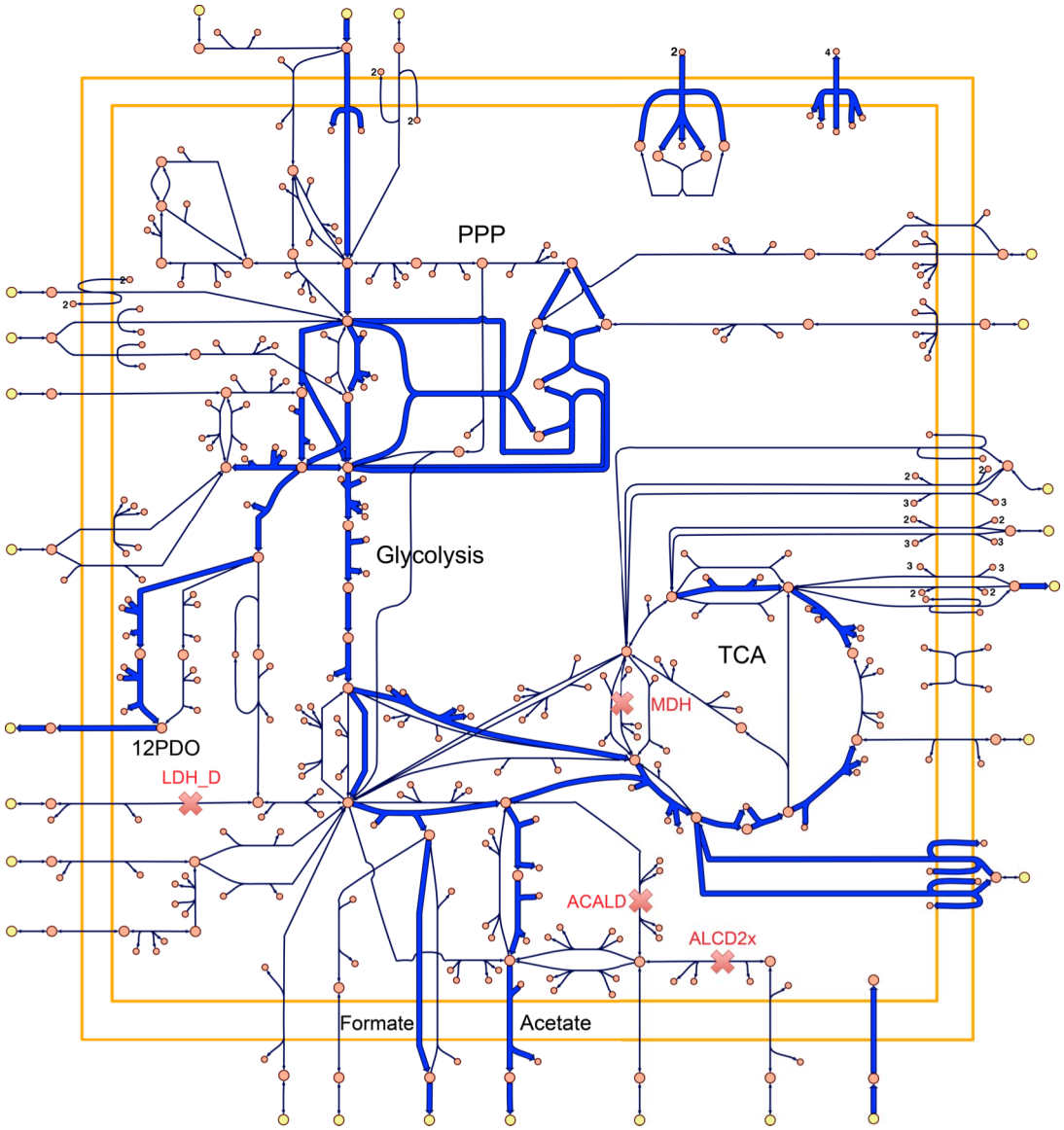

Figure 5.4 The predicted flux distribution for the *ACALD*, *ALCD2x*, *LDH_D*, *MDH* knockout 12PDO production strain. The most efficient routes for the production of ethanol and succinate have been eliminated, making 12PDO the optimal fermentation product. 12PDO is produced by a pathway beginning with dihydroxyacetone phosphate in glycolysis.

then converted to 12PDO by aldose reductase (acetol) (*ALR4x*). The necessary genes for 12PDO production in *E. coli* are known to be functional and capable of producing this product. In a previous study, they were overexpressed on a plasmid in *E. coli*, and this

strain was confirmed to produce 12PDO with a final titer of about 0.5 g/L from glucose [35].

5.3 Discussion

In this study, the application of systems biology methods for metabolic engineering is demonstrated. In order to design strains for the production of useful chemicals, many different types of genetic manipulations are utilized [15]. The simple pathway-level effects of individual gene knockouts may be relatively easy to predict based on current knowledge for many organisms, but the entire metabolic network of any organism is usually very complex and highly interconnected. Thus, a single knockout can have effects on multiple pathways and systems that are difficult to predict. The difficulty in predicting the effects of gene knockouts is increased greatly by the fact that most metabolic engineering designs require multiple knockouts or other genetic alterations. The computational methods of systems biology are useful for understanding these types of complex interactions. Computational tools can also allow for large-scale screens of thousands or even millions of possible designs to be analyzed quickly.

One of the most promising applications of systems biology methods to metabolic engineering is in the design of growth-coupled production strains [39]. For most compounds that a strain can produce and secrete, there is a tradeoff between production and cellular growth. As more mass and energy are directed to production of this product, less is available for production of biomass. As mutations occur in these strains that divert metabolic flux away from production of the target compound and towards production of

biomass components, the growth rate of the cells will increase. Production of the target compound, however, will decrease. In a continuous culture, these faster growing cells will be selected for, and will eventually overtake the slower growers. Production strains that behave in this way are not growth-coupled, and will be unstable in a continuous bioprocessing environment. On the other hand, in a growth-coupled strain, production of the target compound will be essential for the strain to achieve its highest possible growth rate. This can be accomplished by coupling the target compound to production of an essential biomass component or by coupling it to an optimal set of redox reactions required for maintaining ideal cellular NAD^+/NADH and $NADP^+$/NADPH balancing [40]. When grown in culture, as mutations in a growth-coupled strain occur to increase the growth rate, target compound production rate will also increase. Because of this, the high-production rate phenotype will be stable over time in a growth-coupled strain.

To identify the range of natural compounds that can be growth-coupled in the *E. coli* metabolic network through targeted gene knockouts, a large computational screen was performed. First, a new method called RandKnock was developed to randomly sample the diversity of knockout strain fermentation products. This simple algorithm was used to identify potential targets for growth-coupling. The frequency with which products occurred among the millions of random knockout strains that were generated partly indicates the ease with which growth-coupled designs can be developed. Products that occurred many times, such as lactate and succinate, can be produced in many different strains, involving many different combinations of knockouts. There are many different ways to couple these compounds to growth, so the probability of developing a growth-coupled strain *in vivo* is quite high. For rare compounds, on the other hand, there are very

few ways of coupling these products to growth. These products will be more difficult to produce. RandKnock also identified the maximum possible growth rates and production rates for each product. Some potential growth-coupled products, including D-alanine, could only be produced at a very low rate, and thus would not be good targets for computational design efforts.

After a set of 12 target compounds was chosen based partly on the results of the RandKnock screens, a computational screen involving the OptKnock and OptGene algorithms was performed. OptKnock was run first, and was used to search for growth-coupled strains with sets of three and five reaction knockouts. The utility of OptKnock is limited by the fact that it can take a very long time to run (up to one week for five knockout designs with the *i*AF1260b model). It also can only perform bi-level optimization of growth rate and production rate. OptGene was used next. This algorithm has the advantage of being able to use any user designed objective function, including nonlinear functions. It was used here to improve OptKnock solutions using three different objectives. The first objective, maximum yield of the target product, is of obvious value, as the strain with the highest yield will be predicted to produce the most total product. The second objective, maximization of substrate specific productivity, also includes growth rate as a parameter. This is useful because a faster growing strain will be able to grow to a high density and convert substrate to product at a higher rate, allowing for a faster production process. A faster growth rate is also useful for ALE, as it will allow the evolution to proceed faster [41]. As discussed below, ALE is a useful strategy for optimization of growth-coupled strains. The third objective was strength of growth-coupling. This parameter was defined as the slope of the lower part of the production

envelope, and indicates how close a strain may get to its optimal growth rate during ALE while still producing a very small amount of the growth-coupled product. A weakly growth-coupled strain can get very close to its optimal growth rate without producing the optimal product, while a strongly growth-coupled strain will be forced to produce its product even when not close to its optimal growth rate. This property, like SSP, is also important for ALE for strain production, since an *in vivo* strain may have additional constraints not accounted for by a metabolic network model, and may not reach is optimal growth rate. In a strongly growth-coupled strain, it may still achieve a high production rate anyway.

Many computational tools have been developed to design growth-coupled strains, including OptKnock and OptGene, along with more recent methods such as OptStrain [42], OptForce [43], and OptORF [44]. Constraint-based metabolic network models have also been built for multiple organisms used in bioprocessing in addition to *E. coli*, including *Saccharomyces cerevisiae* [45], *Bacillus subtilis* [46], *Clostridium acetobutylicum* [47], and *Zymomonas mobilis* [48]. Despite the extensive computational studies that have been conducted using constraint-based modeling in recent years, there have been few experimental validations of predicted production strains. One set of OptKnock designed *E. coli* strains for the production of lactate from glucose was constructed and optimized by ALE [16]. Another *E. coli* strain, based on the results of the computational study presented in this chapter, was constructed and evolved to convert glucose to lactate with a yield of nearly 100% [Feist et al., in preparation]. This strain, however, was grown in glucose minimal media culture supplemented with a high amount of yeast extract, making comparisons to model predicted phenotypes difficult. In order to

further validate the strain design predictions of model-based algorithms, two more strains from this study were selected for experimental analysis. Several different parameters were considered when selecting strains for further analysis. Strains with high predicted production rates and growth rates were investigated, as these properties are desirable for industrial production strains. An effort was also made to select strains for production of potentially useful products, and literature was investigated to ensure the feasibility of the selected strains as much as possible.

Ultimately, two strains were selected for experimental analysis. A strategy was employed of picking one strain with a very high likelihood of performing as predicted *in vivo* and one strain that appeared to be less likely to succeed in reaching its predicted optimal phenotype. The strain with a very high likelihood of working used knockouts of the pyruvate dehydrogenase and pyruvate formate lyase to couple growth to lactate production. Successful *E. coli* lactate production strains have been developed previously, and this design is closely related to the design optimized by Feist et al. The more novel strain selected for construction was for the production of R-1,2-propanediol. Few studies have been published in which *E. coli* has been engineered to produce 12PDO [3], and the strain design selected utilizes a very non-obvious set of knockouts distal to the 12PDO production pathway. Still, this strain has a high degree of growth-coupling (**Figure 5.2 b**), and the RandKnock screen identified 12PDO among the most common fermentation products of *E. coli* knockout strains. This strain was also selected because with one additional reaction knockout, for the acetate kinase reaction (*ACKr*), this same strain would then be growth-coupled to production of either L-alanine or L-valine. In the next

chapter, the *in vivo* construction and adaptive laboratory evolution of these two strains is presented.

5.4 Methods

5.4.1 Model setup and preprocessing

The COBRA Toolbox version 1.3 [49] was used to perform all constraint-based calculations and analyses in this chapter. All FBA based steps were performed using the optimizeCbModel function and the Tomlab LP solver (Tomlab Optimization Inc., Seattle, WA). For mixed-integer linear programming based methods such as OptKnock, the Tomlab MILP solver was used. For all model-based computations in this chapter, the *i*AF1260b genome-scale metabolic network reconstruction of *E. coli* was used. This updated version of the *i*AF1260 reconstruction contains six new reactions: *DHORTfum*, *MALt3pp*, *ALAt2rpp*, *GLYt2rpp*, *CITt3pp*, and *ASPt2rpp*. The default upper and lower bounds on each reaction were used, except for the reactions *CAT*, *FHL*, *SPODM*, and *SPODMpp*, which were constrained to carry zero flux.

The model was set to utilize different substrates by setting the lower bounds on the appropriate substrate exchange reaction to a value between -10 and -20 mmol/gDW/h. Anaerobic conditions were imposed by setting the lower bound on the oxygen exchange reaction (*EX_o2(e)*) to zero. For all strain design algorithms, condition specific reduced versions of the full *i*AF1260b model were used. The model bounds were set according to the desired conditions, and FVA was used to determine the maximum and minimum possible fluxes through each reaction. All reactions that could not carry any flux were

removed from the model. The upper and lower bounds of the remaining reactions were also set to the maximum and minimum possible fluxes identified by FVA. For each set of conditions, a set of allowable knockout reactions was also determined. Beginning from the full set of reactions in the reduced model, all reactions associated with experimentally identified essential genes in *E. coli* [50, 51] were removed from the list. FBA was then used to identify all essential model reactions, defined here as all reaction knockouts that reduce the maximum growth rate to less than 5% of the wild-type growth rate. These reactions were also removed. Next, all orphan and spontaneous reactions were removed. We also removed all reactions from the subsystems "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." Any reaction involving compounds with more than seven carbon atoms were removed. Finally, the sets of coupled reactions in the reduced model were identified, and all but one reaction from each remaining set was removed. The remaining sets of reaction were then used as candidate knockouts for RandKnock, OptKnock, OptGene, and AnalyzeGCdesign.

5.4.2 Implementation of strain design algorithms

RandKnock

The RandKnock algorithm was designed and implemented in the COBRA Toolbox. This algorithm selects a defined number *N* reactions from a set of target reactions and uses FBA to determine the optimal growth phenotype of a reduced model

with the *N* reactions simultaneously knocked out. The set of *N* reactions knocked out is randomly generated each time RandKnock is run. The optimal growth rate and the production rates of all excreted compounds are then determined and saved. RandKnock was run on both glucose minimal media and xylose minimal media reduced models, for *N* = 3:10 knockouts. For each combination of model and *N*, RandKnock was run 10^6 times. The number of times each compound occurred among the RandKnock products was counted. The optimal production strain for each product was determined as the strain with the highest substrate specific productivity (SSP, maximum production rate * growth rate). The production rate and growth rate for each optimal strain was recorded.

OptKnock

The OptKnock algorithm reformulated as an MILP problem [24] was encoded in the COBRA Toolbox. OptKnock was set to save intermediate solutions as well as the final, optimal solutions. A lower bound of 0.05 h^{-1} was set on growth rates for all solutions, and the algorithm was set to search for solutions with greater than or equal to a defined number of reaction knockouts. For each OptKnock run, a particular exchange reaction was set as the design objective, and a minimum number of allowed reaction knockouts was defined. A time limit between 100 and 336 h was set for most runs. OptKnock was run on reduced *E. coli* models for growth on glucose, xylose, and glycerol under aerobic conditions and glucose and xylose under anaerobic conditions. Solutions containing three, five, and ten reaction knockouts were investigated. For some OptKnock runs, a "tilted" objective function was used to avoid non-unique growth-coupled solutions. The growth objective of OptKnock, usually consisting of just the biomass

reaction, was augmented to include a slight minimization of the flux through the outer membrane transport reaction of the target compound. This forces OptKnock to consider the smallest possible production rate of the target compound as the true production rate.

OptGene

Also encoded in the COBRA Toolbox was the genetic algorithm OptGene [25]. This algorithm was modified from its published version by being allowed to predict sets of gene knockouts as well as reaction knockouts. Various parameters were adjusted to optimize performance of the algorithm for prediction of *E. coli* growth-coupled production strains. OptGene was run with a population size of 500 simultaneous strain designs, migration only in the forward direction, a migration fraction of 0.1 individuals passed to the next population, a maximum of 10 reaction or gene knockouts, a crossover fraction of 0.8, an individual mutation rate of 1/number of reactions, and a crossover mutation rate of 0.2 * the individual mutation rate. The top two optimal strains were always automatically passed to the next generation. A maximum of 10^4 generations and a time limit of 48 h were imposed on OptGene runs.

The starting strains for OptGene were the final and intermediate solutions generated with OptKnock. In addition, during the runs, the initial OptKnock solutions were added back into the populations with a probability of 1/1500, to ensure that these growth-coupled designs would not be lost early. Three different objective functions were maximized used with OptGene.

Maximum product yield:

$$Z = productionRate \quad (1)$$

Maximum substrate specific productivity (SSP):

$$Z = productionRate \times growthRate \quad (2)$$

Maximum strength of coupling (SOC):

$$Z = \frac{productionRate^2}{slope} \quad (3)$$

where *slope* is the slope of the lower part of the production envelope (**Figure 5.1 b**).

AnalyzeGCdesign

The fourth growth-coupled design algorithm to be encoded in the COBRA Toolbox was called AnalyzeGCdesign. The input to this algorithm was an OptKnock or OptGene solution for a particular reduced model and target compound. With the appropriate model loaded and the target reaction defined, AnalyzeGCdesign would replace every reaction in the initial solution one at a time from the set of selected knockout targets, remove every reaction from the solution one at a time, and add every selected reaction to the solution one a time. After each change, the value of an objective function is calculated. The single change that results in the highest objective value is selected as the optimal solution, and AnalyzeGCdesign is run again on this solution. The algorithm runs recursively until no further improvement in objective value can be identified. Eight different objective functions were used to evaluate growth-coupled designs. Several of these objective functions include deletion penalties to remove deletions that do not actually contribute to the growth-coupled phenotype.

Product yield:

$$Z = productionRate \quad (4)$$

Substrate specific productivity (SSP):

$$Z = productionRate \times growthRate \quad (5)$$

Product yield with knockout penalty:

$$Z = productionRate \times delPenalty^{numKOs} \quad (6)$$

SSP with knockout penalty:

$$Z = productionRate \times growthRate \times delPenalty^{numDels} \quad (7)$$

Growth-coupled yield:

$$Z = \frac{productionRate}{slope} \quad (8)$$

Growth-coupled SSP:

$$Z = \frac{productionRate \times growthRate}{slope} \quad (9)$$

Growth-coupled yield with knockout penalty:

$$Z = \frac{producitonRate \times delPenalty^{numKOs}}{slope} \quad (10)$$

Growth-coupled SSP with knockout penalty:

$$Z = \frac{productionRate \times growthRate \times delPenalty^{numKOs}}{slope} \quad (11)$$

5.4.3 Screening of growth-coupled strain designs

A protocol for the design of growth-coupled *E. coli* production strains was developed. First, reduced models and selected knockout candidate reaction lists were assembled. Twelve compounds were selected as potential targets, based partly on the results of RandKnock computations: R-1,2-propanediol, L-alanine, ethanol, fumarate, L-glutamate, glycerol, D-lactate, L-malate, 2-oxoglutarate, pyruvate, L-serine, and succinate.

OptKnock was run with these reduced models and targets to identify three, five, and ten knockout designs. These designs, along with additional designs resulting from test OptKnock runs, were used as seed designs for OptGene runs. OptGene was run multiple times for each starting design, with each of its three objectives sought one at a time. The most promising OptKnock and OptGene designs were then refined using AnalyzeGCdesign. Production envelopes were plotted using COBRA Toolbox functions, and the feasibility of predicted knockout strains were assessed by manual curation to identify the two strains ultimately selected for *in vivo* construction and analysis.

Acknowledgements

Chapter 5 is, in part, adapted from a paper that appeared in Metabolic Engineering, Volume 12, Number 3, Pages 173-86, May 2010. The dissertation author was a co-author of this paper, which was coauthored by Adam M. Feist, Daniel C. Zielinski, Jan Schellenberger, Markus J. Herrgard, and Bernhard Ø. Palsson.

Chapter 5 is also, in part, adapted from a paper that is being prepared for publication under the title "Adaptive laboratory evolution and characterization of computationally designed growth-coupled production strains." The dissertation author was the primary author of this paper, which was coauthored by Adam M. Feist and Bernhard Ø. Palsson.

We would like to thank Christian Barrett, Tom Conrad, Ronan Fleming, Dae-Hee Lee, Harish Nagarajan, Vasiliy Portnoy, Ines Thiele, and Karsten Zengler for their helpful comments and insights.

References

1. Lee, J.W., et al., *Microbial production of building block chemicals and polymers.* Current opinion in biotechnology, 2011.

2. Sauer, M., et al., *Microbial production of organic acids: expanding the markets.* Trends Biotechnol, 2008. **26**(2): p. 100-8.

3. Atsumi, S. and J.C. Liao, *Metabolic engineering for advanced biofuels production from Escherichia coli.* Curr Opin Biotechnol, 2008.

4. Martin, V.J., et al., *Engineering a mevalonate pathway in Escherichia coli for production of terpenoids.* Nat Biotechnol, 2003. **21**(7): p. 796-802.

5. Lee, S.Y., et al., *Metabolic engineering of microorganisms: general strategies and drug production.* Drug Discov Today, 2009. **14**(1-2): p. 78-88.

6. Bailey, J.E., *Toward a science of metabolic engineering.* Science, 1991. **252**(5013): p. 1668-75.

7. Lee, S.Y., D.Y. Lee, and T.Y. Kim, *Systems biotechnology for strain improvement.* Trends Biotechnol, 2005. **23**(7): p. 349-58.

8. Trinh, C.T., et al., *Design, construction and performance of the most efficient biomass producing E. coli bacterium.* Metab Eng, 2006. **8**(6): p. 628-38.

9. Trinh, C.T., P. Unrean, and F. Srienc, *Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses.* Appl Environ Microbiol, 2008. **74**(12): p. 3634-43.

10. Jaluria, P., et al., *Cells by design: a mini-review of targeting cell engineering using DNA microarrays.* Mol Biotechnol, 2008. **39**(2): p. 105-11.

11. Yim, H., et al., *Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol.* Nat Chem Biol, 2011. **7**(7): p. 445-52.

12. Park, J.H., et al., *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**(19): p. 7797-802.

13. Barrett, C.L., et al., *Systems biology as a foundation for genome-scale synthetic biology.* Curr Opin Biotechnol, 2006. **17**(5): p. 488-492.

14. Kim, H.U., T.Y. Kim, and S.Y. Lee, *Metabolic flux analysis and metabolic engineering of microorganisms.* Molecular BioSystems, 2008. **4**(2): p. 113-120.

15. Kim, T.Y., et al., *Strategies for systems-level metabolic engineering.* Biotechnol J, 2008. **3**(5): p. 612-23.

16. Fong, S.S., et al., *In silico design and adaptive evolution of Escherichia coli for production of lactic acid.* Biotechnol Bioeng, 2005. **91**(5): p. 643-8.

17. Dien, B.S., N.N. Nichols, and R.J. Bothast, *Recombinant Escherichia coli engineered for production of L-lactic acid from hexose and pentose sugars.* J Ind Microbiol Biotechnol, 2001. **27**(4): p. 259-64.

18. Kim, Y., L.O. Ingram, and K.T. Shanmugam, *Construction of an Escherichia coli K-12 mutant for homoethanologenic fermentation of glucose or xylose without foreign genes.* Applied and environmental microbiology, 2007. **73**(6): p. 1766-71.

19. Lee, S.J., et al., *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**(12): p. 7880-7.

20. Burgard, A.P. and S.J. Van Dien, *Methods and Organisms for the Growth-Coupled Production of Succinate*, 2006: U.S.A.

21. Alper, H., et al., *Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli.* Metab Eng, 2005. **7**(3): p. 155-64.

22. Alper, H., K. Miyaoku, and G. Stephanopoulos, *Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets.* Nat Biotechnol, 2005. **23**(5): p. 612-6.

23. Lee, K.H., et al., *Systems metabolic engineering of Escherichia coli for L-threonine production.* Mol Syst Biol, 2007. **3**: p. 149.

24. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

25. Patil, K.R., et al., *Evolutionary programming as a platform for in silico metabolic engineering.* BMC Bioinformatics, 2005. **6**: p. 308.

26. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**(121).

27. Perlack, R.D., et al., *Biomass as a feedstock for a bioenergy and bioproducts industry: The technical feasibility of a billion-ton annual supply.*, U.S.D.o. Energy, Editor 2005, Oak Ridge National Laboratory, Oak Ridge, TN.

28. Mahadevan, R. and C.H. Schilling, *The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.* Metab Eng, 2003. **5**(4): p. 264-76.

29. Fang, X., A. Wallqvist, and J. Reifman, *Development and analysis of an in vivo-compatible metabolic network of Mycobacterium tuberculosis.* BMC Syst Biol, 2010. **4**: p. 160.

30. Murphy, G.E. and G.J. Jensen, *Electron cryotomography of the E. coli pyruvate and 2-oxoglutarate dehydrogenase complexes.* Structure, 2005. **13**(12): p. 1765-73.

31. Sawers, G. and A. Bock, *Novel transcriptional control of the pyruvate formate-lyase gene: upstream regulatory sequences and multiple promoters regulate anaerobic expression.* J Bacteriol, 1989. **171**(5): p. 2485-98.

32. Reizer, J., A. Reizer, and M.H. Saier, Jr., *Novel phosphotransferase system genes revealed by bacterial genome analysis--a gene cluster encoding a unique Enzyme I and the proteins of a fructose-like permease system.* Microbiology, 1995. **141 (Pt 4)**: p. 961-71.

33. Mat-Jan, F., K.Y. Alam, and D.P. Clark, *Mutants of Escherichia coli deficient in the fermentative lactate dehydrogenase.* J Bacteriol, 1989. **171**(1): p. 342-8.

34. Clark, D.P., *The fermentation pathways of Escherichia coli.* FEMS Microbiol Rev, 1989. **5**(3): p. 223-34.

35. Altaras, N.E. and D.C. Cameron, *Metabolic engineering of a 1,2-propanediol pathway in Escherichia coli.* Appl Environ Microbiol, 1999. **65**(3): p. 1180-5.

36. Leonardo, M.R., P.R. Cunningham, and D.P. Clark, *Anaerobic regulation of the adhE gene, encoding the fermentative alcohol dehydrogenase of Escherichia coli.* J Bacteriol, 1993. **175**(3): p. 870-8.

37. Ferrandez, A., J.L. Garcia, and E. Diaz, *Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of Escherichia coli K-12.* J Bacteriol, 1997. **179**(8): p. 2573-81.

38. Sutherland, P. and L. McAlister-Henn, *Isolation and expression of the Escherichia coli gene encoding malate dehydrogenase.* J Bacteriol, 1985. **163**(3): p. 1074-9.

39. Pharkya, P., A.P. Burgard, and C.D. Maranas, *Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock.* Biotechnol Bioeng, 2003. **84**(7): p. 887-99.

40. Zhang, X., et al., *Production of L -alanine by metabolically engineered Escherichia coli.* Appl Microbiol Biotechnol, 2007. **77**(2): p. 355-66.

41. Conrad, T.M., N.E. Lewis, and B.Ø. Palsson, *Microbial laboratory evolution in the era of genome-scale science.* Mol Syst Biol, 2011. **7**: p. 509.

42. Pharkya, P., A.P. Burgard, and C.D. Maranas, *OptStrain: a computational framework for redesign of microbial production systems.* Genome Res, 2004. **14**(11): p. 2367-76.

43. Ranganathan, S., P.F. Suthers, and C.D. Maranas, *OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.* PLoS Comput Biol, 2010. **6**(4): p. e1000744.

44. Kim, J. and J.L. Reed, *OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.* BMC Syst Biol, 2010. **4**: p. 53.

45. Mo, M.L., B.Ø. Palsson, and M.J. Herrgard, *Connecting extracellular metabolomic measurements to intracellular flux states in yeast.* BMC Syst Biol, 2009. **3**: p. 37.

46. Henry, C.S., et al., *iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations.* Genome Biol, 2009. **10**(6): p. R69.

47. Lee, J., et al., *Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network.* Applied microbiology and biotechnology, 2008. **80**(5): p. 849-62.

48. Widiastuti, H., et al., *Genome-scale modeling and in silico analysis of ethanologenic bacteria Zymomonas mobilis.* Biotechnol Bioeng, 2011. **108**(3): p. 655-65.

49. Becker, S.A., et al., *Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox.* Nat. Protocols, 2007. **2**(3): p. 727-738.

50. Baba, T., et al., *Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection.* Mol Syst Biol, 2006. **2**: p. 2006.0008.

51. Joyce, A.R., et al., *Experimental and Computational Assessment of Conditionally Essential Genes in Escherichia coli.* J Bacteriol, 2006. **188**(23): p. 8259-8271.

Chapter 6: Metabolic engineering of *Escherichia coli* II: Adaptive evolution and phenotypic characterization of computationally designed strains

Growth-coupled *Escherichia coli* production strains were designed using the constraint-based metabolic model *i*AF1260. Out of hundreds of designs, two were selected for *in vivo* construction and validation. A lactate production strain was constructed and was adaptively evolved to grow on glucose minimal media, anaerobically. After evolution, this strain achieved a high anaerobic growth rate and was confirmed to convert nearly 100% of glucose consumed to lactate. This strain matched the model predicted phenotype very well. The second strain to be constructed and evolved was for the production of 1,2-propanediol. This strain was evolved for 135 days, but failed to reach the model predicted phenotype, instead producing a mixture of succinate, pyruvate, and lactate. It failed to reach the predicted optimal phenotype, instead reaching a suboptimal state. Gene expression analysis of the evolved strain confirmed this suboptimal phenotype, and by comparison to model predicted fluxes led to a prediction for the function of a currently uncharacterized gene.

6.1 Introduction

In the previous chapter, the use of constraint-based metabolic modeling to design potentially useful production strains was introduced. Modeling has the potential to improve the metabolic engineering process by potentially allowing the phenotypes of thousands of different strains to be predicted and screened *in silico* very quickly [1], and can help to identify complex designs by determining the systemic effects and interactions of many different genetic manipulations [2-4]. In particular, constraint-based modeling can be used to design growth-coupled production strains, in which the production and secretion of target compounds is required for an organism to grow at its highest possible rate. The algorithms OptKnock [5] and OptGene [6] were used to design hundreds of different *Escherichia coli* strains using the *i*AF1260 metabolic model [7]. Strains were designed for growth on three different substrates and for production of 12 different products.

Although numerous constraint-based growth-coupled design algorithms have been developed over the past decade [8-12], there have been only a few published examples in which computationally predicted strains were constructed and analyzed *in vivo* [13, 14]. One useful type of modeling for metabolic engineering includes constraint-based procedures such as flux balance analysis (FBA) [15], which can be used to predict the optimal phenotypes of different strains. To reach these predicted optimal states, adaptive laboratory evolution (ALE) can be used [16-18]. Microbial cell cultures can be passed in batch culture or diluted in continuous culture over long periods of time, and as mutations occur to increase the growth rates of particular cells, these cells will outgrow others and dominate the cultures. When starting with growth-coupled production strains,

ALE is predicted to be useful as a metabolic engineering tool. As the growth rates of these strains increase due to random mutations, the production rates of their target compounds should increase as well. Knowledge of the regulatory networks controlling metabolism is not required for this process, as the growth increasing mutations are random. This process is therefore generalizable to any microbial organism, and the growth-coupled design algorithms can be applied to any organism for which a metabolic network model is available.

To validate the utility of growth-coupled strain design and ALE, two *E. coli* strain designs from the previous chapter were selected for *in vivo* construction. These strains both contain gene knockouts and no other genetic manipulations. One strain, a predicted lactate production strain named BOP338, is similar to a previously constructed and evolved growth-coupled lactate production strain. The other strain, a 1,2-propanediol (12PDO) production strain named BOP382, is entirely novel. These gene knockout strains were constructed *in vivo* and were then evolved to determine if they could reach their model-predicted optimal phenotypes. The final phenotypes of the evolved strains were characterized in detail, and gene expression analysis was performed on BOP382. The results of these evolution studies indicate that the predicted phenotypes can be reached through evolution alone for some designs, but not for others. The required mutations to achieve an optimal phenotype may be too unlikely or even impossible for some phenotypes, and additional metabolic engineering approaches could be required for these strains.

6.2 Results

6.2.1 Construction of gene knockout strains

The computational design chosen for a lactate production strain used knockouts of the reactions pyruvate dehydrogenase (*PDH*) and pyruvate formate lyase (*PFL*). The genes required for these reactions were identified from the gene-protein-reaction associations (GPRs) in the *i*AF1260 *E. coli* metabolic reconstruction. Additional literature sources were consulted to ensure that any appropriate isozymes were identified and knocked out as well if necessary. First, the *pflABfocA* (b0902-4) operon was knocked out using the standard method of homologous recombination with the lambda Red recombinase system [19]. The gene *pflB* codes for a pyruvate formate lyase enzyme [20], while *pflA* codes for an activating enzyme [21]. *focA* codes for a formate membrane transporter [22], which in the *i*AF1260 model is associated with the formate transport reactions *FORt2pp* and *FORtppi*. These reactions were not part of the original knockout design, but they were predicted not to be used at the optimal lactate production phenotype. Thus, since *focA* is adjacent to genes that must be knocked out, it was knocked out as well. The second knockout was the *pflDC* operon (b3951-2). *pflD* codes for an alternate pyruvate formate lyase enzyme [23], while *pflC* codes for its activating enzyme [23]. The third knockout was the *aceEF* (b0114-5) operon. These two genes are essential components of the pyruvate dehydrogenase enzyme complex, along with *lpd* (b0116) [24]. It was not necessary to knock out *lpd* because it cannot catalyze the pyruvate dehydrogenase reaction on its own. The final Δ*pflABfoca pflDC aceEF* strain was named BOP338.

The computationally designed 12PDO production strain required knockouts of the *i*AF1260 reactions alcohol dehydrogenase (ethanol) (*ALCD2x*), D-lactate dehydrogenase (*LDH_D*), and malate dehydrogenase (*MDH*), with knockout of the acetaldehyde dehydrogenase (acetylating) (*ACALD*) not required but increasing the strength of growth-coupling. To construct an *in vivo* strain incapable of catalyzing these reactions, the *ldhA* (b1380) gene, which codes for the primary lactate dehydrogenase in *E. coli* [25], was knocked out first. The other lactate dehydrogenases of *E. coli* are coupled to different cofactors such as FAD or ubiquinone [26], and thus are not able to produce lactate at the optimal phenotype of this strain. Next, *mdh* (b3236), coding for the malate dehydrogenase reaction, was knocked out. This is the only known malate dehydrogenase gene in this strain of *E. coli* [27]. The third knockout was the gene *adhE* (b1241). This gene codes for the primary alcohol dehydrogenase of *E. coli* [28], and also has acetaldehyde dehydrogenase activity [29]. Finally, the gene *mhpF* (b0351) was knocked out. This gene codes for another acetaldehyde dehydrogenase enzyme [30]. The Δ*ldhA mdh adhE mhpF* strain was named BOP382. After constructing this strain, an additional gene knockout was performed. The gene *ackA* (b2296), catalyzing the primary acetate kinase of *E. coli* [31], was also removed. This strain was named BOP394 and is predicted in the *i*AF1260 model to be growth-coupled for production of L-alanine or L-valine. Although this strain was constructed, it has not yet been experimentally analyzed or evolved to its optimal phenotype. BOP338, BOP382, and BOP394 all contained kanamycin resistance genes at the locus of the final gene knockout, giving them an antibiotic resistant phenotypes.

6.2.2 Adaptive evolution of lactate production strain

The experimental gene knockouts were performed while growing the strains in rich LB media. The optimal production phenotypes predicted by the model, however, were predicted for growth on glucose minimal media, anaerobically. When the two knockout strains BOP338 and BOP382 were inoculated into glucose minimal media in an anaerobic chamber, they were unable to grow. These strains were also inoculated into unsupplemented aerobic glucose minimal media and anaerobic glucose minimal media supplemented with yeast extract, but were also unable to grow in these conditions. Initially, they could only grow aerobically in media supplemented with yeast extract. These experimental observations were supported by computational predictions. When FBA is used to predict a phenotype under a particular condition, it predicts the optimal possible phenotype [15]. In the case of BOP338 and BOP382, viability is predicted by FBA on glucose minimal media, anaerobically. When a different constraint-based method, minimization of metabolic adjustment (MOMA) is used, these two strains are predicted to have a growth rate of zero under these conditions. MOMA attempts to predict the immediate effects of a gene knockout on the flux distribution of a knockout strain, and thus can more accurately predict the phenotype of a strain that has not been evolved [32]. The metabolic adjustments required for these knockout strains to grow under these challenging conditions are too great to occur immediately.

In most previous ALE studies, the strain of interest was grown in the same media conditions during the entire evolution process [14, 18, 33, 34]. However, as the BOP338 and BOP382 strains could not initially grow under their desired final conditions, a new adaptive evolution procedure had to be developed. Several adaptive evolution

experiments were conducted to determine the best way to do this. First, BOP382 was evolved to grow in glucose minimal media with no supplementation under aerobic conditions. This strain could not grow in minimal media without some kind of supplement such as yeast extract. A 100 mL culture of this strain was started glucose minimal media with 0.1g/L yeast extract. The culture was passed once per day following the standard adaptive evolution procedures, and after the growth rate appeared to stabilize after 5 days, the yeast extract was reduced to 0.01 g/L. After 6 more days, no yeast extract was added to the media, and the culture was able to grow in minimal media. This experiment demonstrated that a strain can be quickly adapted to grow in minimal media even if it is unable to do so initially. A second experiment was performed in which BOP384, a previously evolved strain that is related to BOP338 [Feist et al., in preparation], was evolved to grow in anaerobic conditions. This strain could grow in glucose minimal media thanks to a previous evolution, but it could not yet grow anaerobically. An aerobic culture was started and grown with daily passages for 11 days. When attempts were made to pass this culture to flasks in the anaerobic chamber, it did not survive. Instead, the culture was passed to a glass bottle with a screw-on lid. This lid had a small hole drilled in it with a short length of rubber tubing and a small syringe filter at the end. This small hole would severely limit oxygen diffusion into the culture, but not eliminate it completely, creating a microaerobic environment. The exact oxygen concentration in this bottle could not be easily measured, but an observed decreased growth rate was a convincing indicator of decreased oxygen. After seven days in these microaerobic bottles, the culture was passed to a flask in the anaerobic chamber, and could now stably grow under anaerobic conditions. By using the microaerobic bottles as

an intermediate step to help the strain adapt more gradually, it was possible to adapt this strain to grow anaerobically.

Several previous adaptive evolution studies have shown that the use of a chemical mutagen can increase the rate of adaptation by increasing the rate of mutation in evolving strains [35-37]. Specifically, the mutagen N-methyl-N'-nitro-N-nitrosoguanidine (NTG) was found to enhance the adaptability of some *E. coli* strains at concentrations of 2-5 mg/L in media [35]. However, when an NTG concentration of 5 mg/L was used during initial growth experiments with both BOP338 and BOP382 under aerobic conditions, growth rate decreased somewhat compared to lower amounts, so it was determined that this NTG concentration was too high. A concentration of 1 mg/L was used in subsequent adaptive evolution studies.

Once methods for evolving an *E. coli* knockout strain to grow in minimal media and under anaerobic conditions were developed, the BOP338 strain could be evolved. A single colony of BOP338 was used to start two 100 mL cultures in 4 g/L glucose minimal media with 0.1 g/L yeast extract and 1 mg/L NTG. These two cultures were named "GA" and "GB", and were independent evolutions of the same starting strain. They were grown under aerobic conditions at 37°C and passed once per day. Growth rate was estimated each day from an OD_{600} measurement (**Figure 6.1**), and this was used to determine what volume was passed to the next day's flask. At the time of each OD measurement, a sample of each culture was also taken and stored in a -80°C freezer. After 19 total days, the supplementation in both cultures was reduced to 0.05 g/L for two days, then 0.01 g/L for three days, and finally to zero on the 24^{th} day of evolution. The growth rate of strain GA was approximately 0.17 h^{-1} and the growth rate of GB was approximately 0.13 h^{-1}

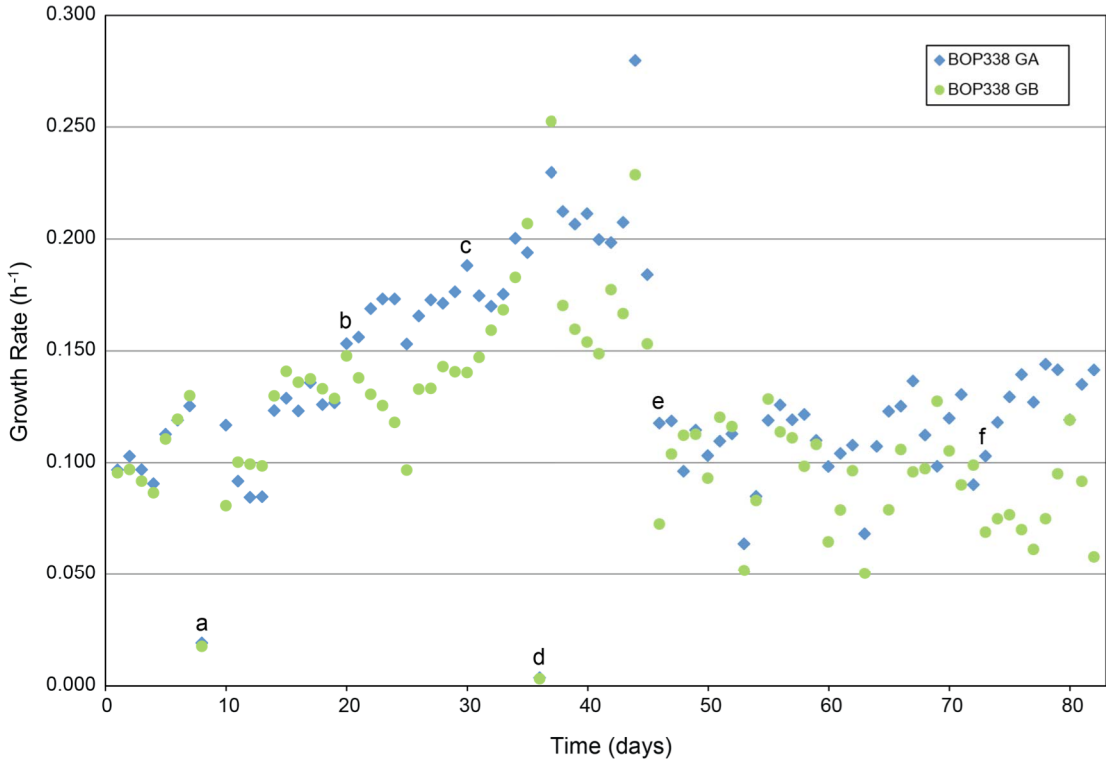

Figure 6.1 Estimated growth rates for BOP338 GA (blue) and GB (green) during the adaptive evolution to growth in glucose minimal media under anaerobic conditions, starting from aerobic conditions with 0.1 g/L yeast extract supplementation. (**a**) The first unsuccessful attempt to decrease yeast extract supplementation to 0.01 g/L. Supplementation was restored to 0.1 g/L the next day. (**b**) Yeast extract supplementation was dropped to zero. (**c**) The cultures were passed to microaerobic bottles. (**d**) The adaptive evolution experiment was halted and later resumed by starting new cultures from frozen cells. (**e**) The cultures were passed to flasks in an anaerobic chamber, and yeast extract supplementation was given again. (**f**) Yeast extract supplementation was again reduced to zero.

at this point. After 28 total days, both cultures were passed to microaerobic bottles. Both cultures were grown for a total of 17 days in the microaerobic bottles before being passed to flasks in the anaerobic chamber for fully anaerobic growth. Yeast extract was reintroduced to both cultures at 0.1 g/L. The yeast extract was reduced to 0.05 g/L after seven days and to 0.01 g/L after eight more days. Finally, the yeast extract was reduced to zero twelve days later, on the 72nd day of evolution. Both cultures were continued for

10 more days to verify that they could grow stably under anaerobic conditions in glucose minimal media. Kanamycin was added to the media approximately every 7-10 days during the evolution to prevent contamination. PCR using the primers from the knockout process (**Section 6.2.1 Construction of gene knockout strains**) was used to confirm that both cultures still retained the correct gene knockouts approximately every 3 weeks during the evolutions.

6.2.3 Phenotypic characterization of lactate production strain

After completion of the evolution, both GA and GB BOP338 strains were frozen at -80°C. These two strains were then phenotypically characterized. Each strain was grown in duplicate over a period of 48 hours, with 12 samples taken from each at different times. From each sample, the OD_{600} was measured to determine the exponential growth rate, the pH was measured, and the sugars and fermentation products in the media were analyzed by HPLC (**Table 6.1** and **Figure 6.2**). The exponential growth rate of BOP338 GA was 0.173 h^{-1} and the growth rate of GB was 0.108 h^{-1}.

Table 6.1 Properties of the evolved BOP338 strains.

	Evolved Strain	
	BOP338 GA	**BOP338 GB**
Growth rate (h^{-1})	0.174±0.001	0.108±0.003
Glucose uptake rate (mmol/gDW/h)	23.50±0.64	13.49±0.36
Lactate production rate (mmol/gDW/h)	50.21±1.17	28.49±0.13
Final lactate yield (g/g glucose)	1.06±0.01	1.01±0.01
Final pH	5.46±0.00	5.44±0.01

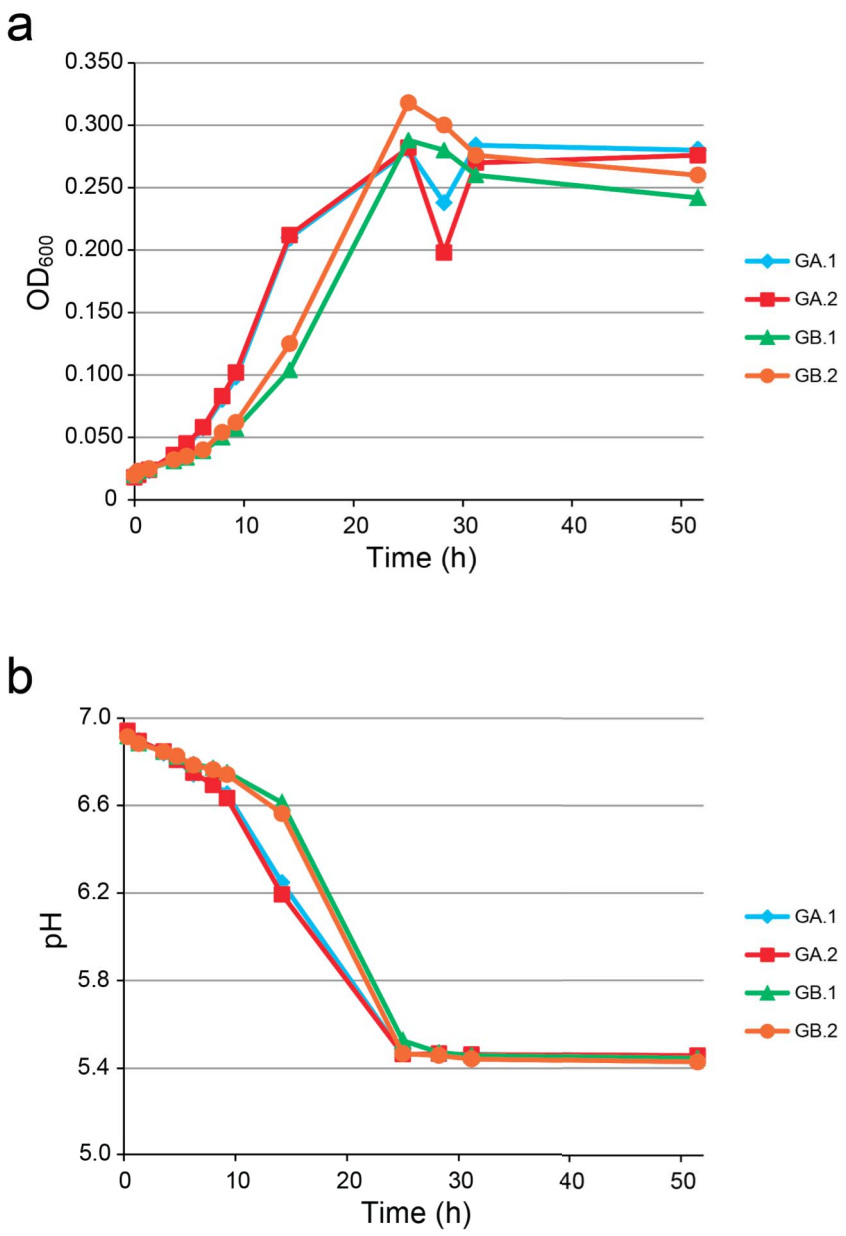

Figure 6.2 Phenotypic characterization of the two evolved BOP338 strains. The two strains (GA and GB) were grown in duplicate for their final characterization. OD_{600} (**a**) and pH (**b**) were measured over 52 hours.

Both grew to similar final densities. Strain GA produced lactate at a rate of 50.2 mmol/gDW/h and strain GB at 28.5 mmol/gDW/h, while GA consumed glucose at 23.5 mmol/gDW/h and GB at 13.5 mmol/gDW/h. The final yield of lactate from glucose in

strain GA was 1.06 g lactate/g glucose and in strain GB was 1.01 g lactate/g glucose. Both strains consumed all available glucose and reached a final pH of 5.4. These results are very similar to the computational predictions, indicating that the predicted optimal phenotypes were correct, and that the evolution process was successful. Strain GA achieved a higher final glucose uptake rate and growth rate than GB, but both achieved lactate yields of approximately 100% from glucose.

6.2.4 Adaptive evolution of 12PDO production strain

A similar procedure to the one used to evolve the BOP338 lactate production strain was used to evolve the BOP382 12PDO production strain. This strain was evolved in duplicate, with the two replicates again named "GA" and "GB." As before, evolution began from a single colony that was used to start both the GA and GB evolutions. To begin, cultures were grown aerobically in 100 mL glucose minimal media supplemented with 0.1 g/L yeast extract with 1 mg/L NTG. Growth rate was estimated by measuring the OD_{600}, and the evolving cultures were passed once per day to maintain exponential growth. The same pH and media composition measurements made during the BOP338 evolution were made during this evolution as well. The yeast extract supplementation in both cultures was steadily decreased, and was eliminated entirely in both GA and GB cultures after 14 days of evolution. After three days of steady growth in unsupplemented minimal media, both strains were passed to microaerobic bottles. The initial estimated microaerobic growth rate for GA was 0.13 h^{-1} and for GB was 0.12 h^{-1}. After 15 days of microaerobic growth (on the 31st day of evolution), both cultures were passed to flasks in the anaerobic chamber. However, both of the strains ceased growing after one day in an

anaerobic environment. The microaerobic cultures were continued in parallel during the attempted anaerobic growth, and were used to continue the evolutions. Additional attempts to pass the two evolving cultures to anaerobic flasks were made after 17, 21, and 23 days of microaerobic growth, each time with increasing amounts of microaerobic culture being passed and with increasing amounts of yeast extract supplementation added to the anaerobic media, but none of these attempts were successful. Attempts were also made to reach anaerobic growth by passing the cultures to bottles with sealed screw-on lids and by growing the cultures in the microaerobic bottles with the magnetic stirring turned off, but these strategies were not successful either. In the meantime, the growth rates of the microaerobic cultures steadily increased to approximately 0.5 h^{-1} for both GA and GB, as the strains adapted to microaerobic growth.

Finally, after 49 days of microaerobic growth, the two cultures were passed to anaerobic flasks containing 1 g/L yeast extract. These cultures continued to grow anaerobically, and the yeast extract supplementation was increased to 2 g/L after 28 days due to low growth rates (around 0.1 h^{-1}). The growth rates increased to approximately 0.25 h^{-1}. Every few days for the rest of the evolution procedure, the yeast extract supplementation was decreased, usually with a corresponding drop in growth rate. Yeast extract supplementation was decreased to as little as 0.01 g/L, but the two strains could not survive without at least a small amount yeast extract, despite numerous attempts. Because HPLC analysis of the media indicated that 12PDO was not being produced in either strain despite them begin grown in nearly minimal media conditions, anaerobically, the evolutions were halted after 135 total days (**Figure 6.3**).

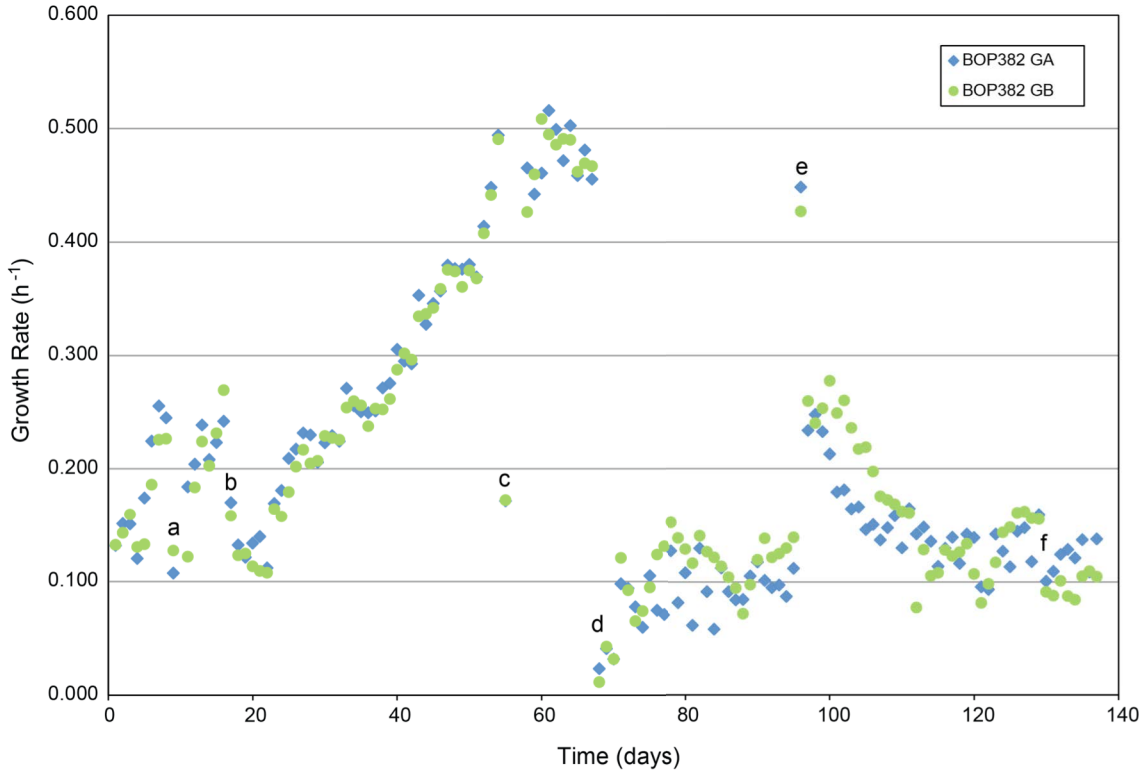

Figure 6.3 Estimated growth rates for BOP382 GA (blue) and GB (green) during the adaptive evolution to growth in glucose minimal media under anaerobic conditions, starting from aerobic conditions with 0.1 g/L yeast extract supplementation. (**a**) Yeast extract supplementation was decreased to zero. (**b**) The cultures were passed to microaerobic bottles. (**c**) Magnetic stirring was temporarily reduced to zero. (**d**) The cultures were passed to flasks in an anaerobic chamber and yeast extract supplementation was restored at 1 g/L. (**e**) Yeast extract supplementation was increased to 2 g/L. (**f**) Yeast extract supplementation was reduced to 0.01 g/L.

6.2.5 Phenotypic characterization of 12PDO production strain

Following the adaptive evolution process, the two BOP382 strains were phenotypically characterized. As was done for the evolved BOP338 strains, a growth curve experiment was performed. OD_{600}, pH, and concentrations of substrates and fermentation products were measured and rates were calculated (**Table 6.2** and **Figure 6.4**). The BOP382 strains grew at lower rates than the BOP338 strains. Strain GA grew at 0.053 h^{-1} and strain GB grew slightly faster at 0.089 h^{-1}. GB also grew to a higher final

Table 6.2 Properties of the evolved BOP382 strains.

	Evolved Strain	
	BOP382 GA	**BOP382 GB**
Growth rate (h^{-1})	0.053±0.002	0.089±0.002
Glucose uptake rate (mmol/gDW/h)	9.04±0.22	9.86±0.64
Succinate production rate (mmol/gDW/h)	9.03±0.08	5.82±0.05
Pyruvate production rate (mmol/gDW/h)	4.99±0.16	5.28±0.09
Lactate production rate (mmol/gDW/h)	2.23±0.11	6.38±0.20
Final succinate yield (g/g glucose)	0.67±0.01	0.48±0.02
Final pyruvate yield (g/g glucose)	0.27±0.01	0.30±0.01
Final lactate yield (g/g glucose)	0.12±0.01	0.26±0.01
Final pH	5.13±0.02	5.07±0.01

OD_{600} than GA, indicating more efficient conversion of glucose to biomass in this strain. Both products produced combinations of succinate, pyruvate, and lactate, but in different ratios. Strain GA produced primarily succinate, with smaller amounts of pyruvate and lactate. Strain GB also produced primarily succinate, but produced about twice as much lactate as GA did. The combined yields of the three products in both strains were approximately 100% from glucose. The *i*AF1260 model predicted optimal fermentation products 12PDO, acetate, and formate were not detected in the media for either strain, indicating that they were not growing at their predicted optimal phenotypes.

To determine why the evolved strains did not achieve the initially predicted phenotype, the BOP382 strain design was reanalyzed using the final *i*JO1366 *E. coli* model [38], rather than the *i*AF1260 model used to design this strain. With the reactions *ACALD*, *ALCD2x*, *LDH_D*, and *MDH* constrained in the *i*JO1366 model, the predicted optimal phenotype for growth on glucose under anaerobic conditions included acetate

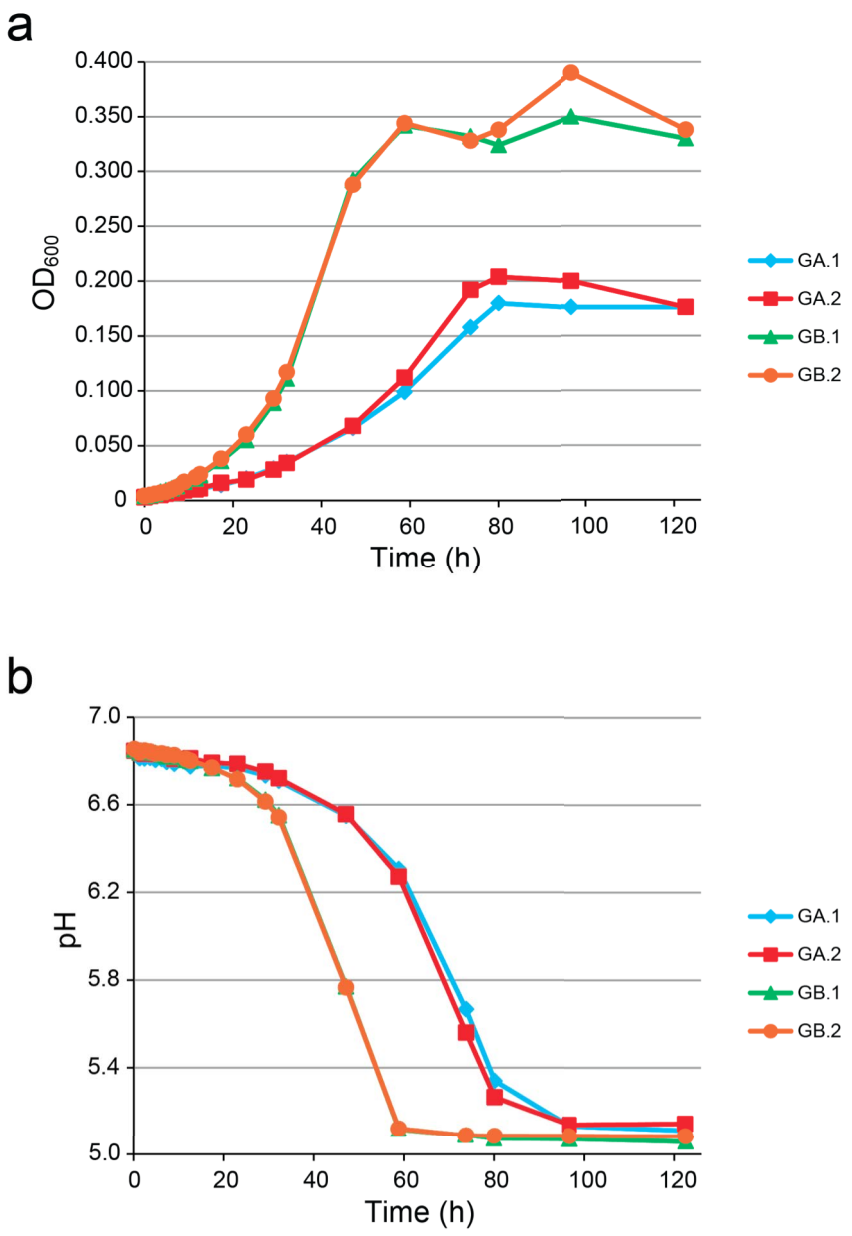

Figure 6.4 Phenotypic characterization of the two evolved BOP382 strains. The two strains (GA and GB) were grown in duplicate for their final characterization. OD_{600} (**a**) and pH (**b**) were measured over 123 hours.

and formate as growth-coupled products, but not 12PDO as initially predicted. Analysis of the *i*JO1366 model identified two reactions responsible for the qualitative change in phenotype. The reaction acetyl-CoA C-acyltransferase (butanoyl-CoA) (*ACACT2r*)

converts acetyl-CoA and butanoyl-CoA to 3-oxohexanoyl-CoA [39]. This reaction is a reversible step in fatty acid beta-oxidation, and can run in the reverse direction in *i*JO1366. In previous versions of the *E. coli* model, this reaction was irreversible. The other responsible reaction is pyruvate synthase (*POR5*), which in its reverse direction can convert acetyl-CoA to pyruvate while oxidizing a reduced flavodoxin [40]. Together, these two reactions slightly increase the efficiency of the metabolic network during anaerobic growth on glucose, causing production of 12PDO to no longer be required for optimal redox balancing. The *i*JO1366 model was constrained to the BOP382 genotype with additional constraints added, forcing secretion of succinate, pyruvate, and lactate, and it was confirmed that this is a viable but suboptimal phenotype. The two BOP382 strains thus failed to reach their optimal phenotypes (as predicted by either the *i*AF1260 or *i*JO1366 versions of the model) and remained stalled at suboptimal phenotypes despite a very long adaptive evolution process.

6.2.6 Gene expression analysis of 12PDO production strain

Expression profiling of the BOP382 GB strain was performed to help identify the suboptimal phenotype that this strain reached. RNA-seq was used to identify the transcripts during mid-log phase growth in both the unevolved and the evolved BOP382 strains. Both absolute gene expression in the evolved strain and differential gene expression between the two strains were investigated. The overall gene expression analysis pattern was assessed using a procedure called parsimonious enzyme usage FBA (pFBA) [41]. This algorithm classifies each gene in a metabolic model by the effect it has on the optimal growth phenotype. The number of genes in each category that are

determined to be expressed or not by experimental gene expression analysis can then be determined. When the BOP382 GB expressed genes were compared to the pFBA classifications of the genes generated using the *i*JO1366 model, it was found that 89.2% of the essential genes were expressed (**Table 6.3**). Of the genes not essential for growth but required for the optimal growth phenotype, only 75.2% were expressed. This number is smaller than the 82% of optimal genes determined to be expressed in a larger set of adaptively evolved *E. coli* strains [41]. Fairly high numbers of both enzymatically and metabolically less efficient were also found to be expressed in evolved BOP382. Overall, these results support the assertion that evolved BOP382 failed to reach its optimal growth phenotype. The high number of blocked model genes expressed (63%) indicates that the current metabolic model is still incomplete (see **Chapter 4: Gap-filling of the *Escherichia coli* metabolic network for model improvement and discovery**).

Next, the expression of individual genes was investigated in the context of the *i*JO1366 metabolic model. The four genes that were knocked out (*ldhA*, *mdh*, *adhE*, and *mhpF*) were confirmed not to be expressed. Some of the genes in the predicted 12PDO synthesis pathway were found to be expressed, including *mgsA* (b0963), *gldA* (b3945), *dkgA* (b3012), and *yghZ* (b3001). The genes *dkgB* (b0207) and *yeaE* (b1781) were not expressed however, so not every gene in any of the two possible pathways to 12PDO was expressed, explaining the lack of 12PDO production. Genes necessary for the production and secretion of succinate (*fumABC* (b1611-2, b4122), *frdABCD* (b4151-4), *dcuAB* (b4123, b4138), and *dctA* (b3528)) and pyruvate (*pykA* (b1854) and *pykF* (b1676)) were expressed, explaining the production of these two products. The lactate detected in the BOP382 media could be either D- or L-lactate. The known gene required to produce L-

Table 6.3 Number of genes expressed in each pFBA category for evolved BOP382.

Gene Classification	Number Expressed
Essential	198/212 (89.2%)
Optimal	164/218 (75.2%)
Enzymatically less efficient	36/56 (64.3%)
Metabolically less efficient	274/487 (56.3%)
Zero flux	89/274 (32.5%)
Blocked reactions	75/119 (63%)

lactate, *aldA* (b1415), was not expressed, making it likely that the detected lactate was D-lactate only. With the primary D-lactate dehydrogenase (*ldhA*) knocked out, there were two remaining D-lactate production pathways in *i*JO1366 (**Figure 6.5**). Both routes begin with dihydroxyacetone phosphate in glycolysis, and convert it to methylglyoxal. The gene required for this reaction, *mgsA*, was expressed. In one route, methylglyoxal is next converted to S-lactoylglutathione by the product of *gloA* (b1651), which is then converted to D-lactate by the product of *gloB* (b0212). Of these two genes, only *gloA* was expressed. The other route to D-lactate production is direct conversion of methylglyoxal to D-lactate by the reaction glyoxylase III (*GLYOX3*). This is an orphan reaction that has been biochemically tested using a purified *E. coli* enzyme, but the gene encoding this enzyme was not identified at the time [42]. This reaction is thus a candidate for D-lactate synthesis in BOP382. To identify a possible gene for this reaction, the set of highly expressed genes in evolved BOP382 compared to unevolved BOP382 were investigated. Among the 25 most highly differentially expressed genes in the evolved strain was the gene *ydbD* (b1407). Little is known about this gene, but it has been annotated as "protein involved in detoxification of methylglyoxal" [43, 44]. It is possible that *ydbD* is the gene

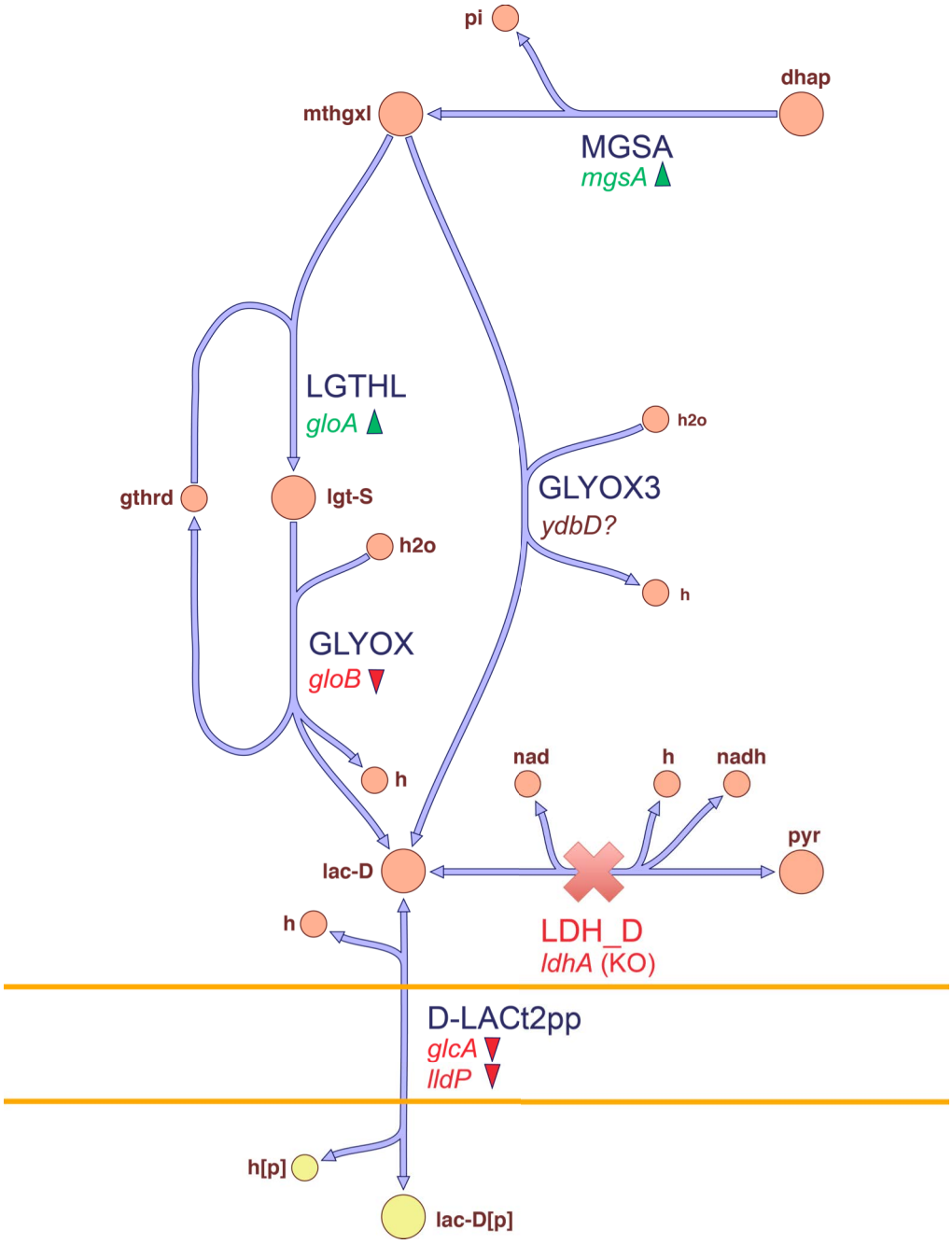

Figure 6.5 The three known pathways in *E. coli* for production of D-lactate are shown. Reaction and metabolite abbreviations are from the *i*JO1366 *E. coli* model. Genes were experimentally determined to be expressed (upward pointing arrows) or not expressed (downward pointing arrows). With the gene *ldhA* knocked out and *gloB* not expressed, D-lactate must be produced through the *GLYOX3* reaction.

coding for the unknown enzyme that was determined to provide the glyoxylase III activity in *E. coli*. This is another model-based gene function prediction that can be experimentally verified to increase our knowledge of *E. coli* metabolism.

6.3 Discussion

Using constraint-based methods such as OptKnock and OptGene, growth-coupled production strain designs for various microorganisms can be generated very quickly. In the previous chapter, a study was performed in which hundreds of gene knockout designs for *E. coli* were generated. Unfortunately, the actual *in vivo* construction and analysis of these types of strains remains significantly slower than the *in silico* design process. Because of this, very few studies have been performed to date to validate computationally design growth-coupled production strains. In one previous study, several different *E. coli* strain designs were constructed and adaptively evolved on glucose minimal media, and were shown to produce lactate as predicted [14]. In another unpublished study, a separate computationally designed *E. coli* was evolved and produced lactate from glucose at a very high rate [Feist et al., in preparation]. This strain had an extremely high glucose uptake rate (over 40 mmol/gDW/h) and growth rate (0.86 h^{-1}), but it also required significant yeast extract supplementation in order to achieve this phenotype. Because of this, the *in vivo* growth conditions did not exactly match the *in silico* growth prediction conditions, so an exact comparison of predicted and observed growth is not possible. In order to truly validate the feasibility of a computationally designed growth-coupled strain, the *in vivo* experiment must be performed in minimal media as predicted. In

supplemented media, the additional uncharacterized nutrients present in yeast extract many compensate for model predicted use of pathways that are actually unrealistic under the *in vivo* conditions.

The purpose of the study presented in this chapter is to validate two different computationally designed growth-coupled *E. coli* knockout strains. In order to make direct comparisons to the model predictions, these strains were evolved to grow in unsupplemented minimal media. Also, in order to assess the utility of ALE as a strain engineering tool [3], no other metabolic engineering techniques were used in this study. Genes were knocked out of *E. coli*, and the knockout strain phenotypes were optimized through evolution alone. If such a strategy could be successful, it would have many advantages. By only performing gene knockouts, the genetic engineering phase of the design can be performed fairly quickly. No additional compounds or unusual conditions are required for induction and proper activation of heterologous genes, and the knockout strains are genetically stable. There is no way for the deleted gene functions to be restored. Specific knowledge of the regulation in the engineered organism is not required, as evolution is expected to make whatever changes are necessary to reach its optimal phenotype. Finally, the evolved strain will be genetically stable, having already been adapted to the desired growth conditions.

Two strain designs were chosen for experimental investigation. One strain, BOP338, was predicted to produce lactate, and the other, BOP382, was predicted to produce 12PDO. Both strains were predicted by FBA to grow on glucose minimal media anaerobically, but neither could do so initially when they were constructed *in vivo*. Because of this, a new ALE protocol had to be developed. The evolutions began

aerobically in yeast extract supplemented media. The amount of yeast extract was gradually reduced to zero, and the strains were then passed to microaerobic bottles as an intermediate step between aerobic and anaerobic growth. Eventually, the evolving strains could be passed to fully anaerobic conditions, although yeast extract was required again initially. This complicated procedure made these evolutions much longer than most ALE studies, which typically last between 30 and 60 days [14, 17, 18, 33, 34]. The BOP338 evolution lasted 81 days, and the BOP382 evolution lasted 135 days.

The final result of this study is that the BOP338 strain reached its predicted optimal phenotype, while BOP382 did not despite an extremely long evolution. BOP338 produced lactate at a very high rate, and converted approximately 100% of the glucose it consumed to lactate for a final lactate titer of about 4 g/L. This result, along with the results from the previously mentioned growth-coupled design studies, indicates that *E. coli* can successfully be growth-coupled for the production of lactate. The model predicted designs have all been correct. The 12PDO production strain, however, did not perform as predicted. Even when the updated *i*JO1366 metabolic model predicted a slightly different optimal phenotype than the *i*AF1260 model, this phenotype was not reached. The actual observed phenotypes in two independently evolved strains, involving production of succinate, pyruvate, and lactate, were confirmed to lie within the model solution space as suboptimal solutions with a low growth rates. In this case, the predicted flux distribution proved to be too difficult to reach by evolution alone, at least within the amount of time of this experiment. This result indicates that ALE alone may not be a feasible experimental optimization strategy for many computationally designed strains. Without knowing the full regulatory network of *E. coli*, it is not possible at this time to

predict which mutations would be required to activate the optimal pathways under these experimental conditions. It is possible that additional experimental interventions, such as upregulation of the 12PDO production pathway, combined with the BOP382 gene knockouts and evolution, could be successful. Still, this negative result, in the context of the metabolic network model, can provide useful information. By expression profiling one of the evolved BOP382 strains and comparing model predicted fluxes to upregulated genes, a potential function for the poorly characterized gene *ydbD* has been proposed.

Because of the extreme amount of time and effort required to perform ALE experiments, only two strains could be evolved in this study. Construction and evolution of more growth-coupled strain designs would be useful to more generally characterize the feasibility of computational strain designs. Automated evolution procedures would be very useful for this purpose. The exact changes made during the evolution of these two strains are currently unknown. Another potentially useful future study would be to resequence the genomes of the evolved strains. The evolutions that occurred are easy to identify, although the mechanisms by which they alter the growth phenotypes may be difficult to uncover [45, 46]. Still, further experiments of this type would help to understand the adaptive evolution process, and could help us better predict which growth-coupled strain designs could be successful.

6.4 Methods

6.4.1 Strain construction by gene knockouts

The starting strain for all strains constructed in this study was *E. coli* K-12 MG1655. Each knockout strain was constructed by homologous recombination with PCR amplified fragments using the lambda Red recombinase system [19]. For each gene, primers were designed containing sequences surrounding the ORF, and these were used to amplify a kanamycin resistance gene flanked by FRT sites. The genomic ORF was replaced by the kanamycin gene through homologous recombination, and this gene insert was later removed with an FLP recombinase, leaving a small scar region in place of the original ORF. For the BOP338 lactate production strain, the *pflABfocA* operon was removed first, followed by *pflDC* and *aceEF*. To produce the BOP382 12PDO production strain, *ldhA* was removed first, followed by *mdh*, *adhE*, and *mhpF*. The kanamycin resistance gene was not excised after the final gene replacement for each strain, making the final BOP338 and BOP382 strains kanamycin resistant. To confirm the genotypes of the knockout strains, primers flanking each knockout site were designed and used to amplify genomic DNA. Amplified fragments were run on a 2% agarose gel with EtBr, and the length of the fragments indicated whether the genes were knocked out (~600 bp), contained the kanamycin insert (~1800 bp), or remained as the wild-type gene (length varies).

6.4.2 Adaptive evolution procedures

The two strains were evolved in duplicate using a batch culture protocol adapted from previous ALE studies [18]. To begin each evolution, a single colony was picked

from an LB agar plate and used to inoculate a 100 mL culture in a standard Erlenmeyer flask. M9 minimal media (6.8 g/L sodium phosphate dibasic, 3.0 g/L potassium phosphate monobasic, 0.5 g/L sodium chloride, 0.24 g/L magnesium sulfate, 0.011 g/L calcium chloride) containing trace elements (0.1 g/L iron (III) chloride, 0.02 g/L zinc sulfate, 0.004 g/L copper chloride, 0.01 g/L manganese sulfate, 0.006 g/L cobalt chloride, 0.006 g/L disodium EDTA), Wolfe's Vitamin Solution, 4 g/L glucose, 0.1 g/L yeast extract, and 1 mg/L NTG chemical mutagen (N-methyl-N'-nitro-N-nitrosoguanidine). Flasks were initially grown aerobically at 37°C with magnetic stirring. Approximately every 24 h during evolution, the OD_{600} was measured and used to estimate the growth rate by assuming that strains were in exponential growth during the entire period since the previous passage. This growth rate was used to calculate the amount of culture necessary to pass to a new flask containing 100 mL fresh media in order to maintain exponential growth. After passing, growth was continued in the previous culture flask for approximately 24 h more and its OD_{600} is measured as well in order to determine the stationary OD of the culture and ensure that the actively evolving culture is passed at mid-exponential phase. At each passage, 0.8 mL culture was also collected, mixed with 0.8 mL 50% glycerol, and frozen at -80°C. A 2 mL culture sample was also taken at each passage, syringe filtered with a 0.2 μm filter, and stored temporarily at -20°C before being analyzed by HPLC (see Section **6.4.3 Phenotypic analysis of evolved strains**). Another 2 mL sample was taken each day for a pH measurement.

Once the growth rates of the evolving strains stabilized during aerobic growth, the amount of yeast extract was gradually decreased by adding smaller amounts to the flasks each day. When the strains were able to grow aerobically without any yeast extract

supplementation, they were passed to microaerobic bottles instead of open flasks. These bottles were standard 100 mL glass bottles with screw-on lids. A hole was drilled in each lid, and a short (~2 cm) length of tubing with a small syringe filter on the end was screwed into the hole, allowing only a very small amount of oxygen to enter. The media used in these bottles had been bubbled with anaerobic N_2 gas for 30 min and media bottles were sealed tight before autoclaving. The microaerobic culture bottles were only opened for filling with media, passing, and taking measurements inside an anaerobic chamber (Coy, Grass Lake, MI). Once the bottles were inoculated sealed, they were kept in an aerobic 37°C incubator with magnetic stirring. The evolving cultures were eventually passed to flasks that were kept inside the anaerobic chamber, allowing for fully anaerobic growth. The same anaerobic media was used, and the active cultures were never allowed outside the chamber.

6.4.3 Phenotypic analysis of evolved strains

After adaptive evolution, the final strains were frozen at -80°C in glycerol. To characterize the growth rates and fermentation product phenotypes of these strains, a small amount of frozen culture was inoculated into the same final media used for the evolutions, not including the NTG mutagen. The cultures were inoculated and grown in the anaerobic chamber. The cultures were grown and passed for several days, until the growth rate recovered to the final rate at the end of the evolutions. A growth curve experiment was then performed, starting by inoculating between 10 and 20 mL of mid-exponential culture into a flask with 100 mL fresh media in the anaerobic chamber. Approximately every hour, the OD_{600} and pH were measured, and a media sample was

filtered and frozen for later HPLC analysis. Measurements were made until the cultures had been in stationary phase for approximately 24 h.

Filtered media samples were analyzed by high-performance liquid chromatography (HPLC) (Waters, Milford, MA). A sample injection volume of 50 μL was used with an Aminex 87-H ion exchange column (Bio-Rad, Hercules, CA) and a 5 mM H_2SO_4 mobile phase (0.5 mL/min). Standard curves were generated for common fermentation products and substrates such as 12PDO, acetate, ethanol, formate, glucose, lactate, maltose, pyruvate, and succinate. Standards were run in water and in M9 media with glucose and yeast extract in concentrations similar to those used in the evolutions and growth phenotyping experiments. Retention times were used to identify compounds in the media, and areas under the curve were used with the standard curves to quantify the compounds.

6.4.4 Gene expression analysis

Gene expression data was collected for B382 using a modified Illumina RNA-seq procedure. First, frozen evolved B382 (both the GA and GB evolution endpoint strains) were inoculated into fresh media and grown anaerobically until they recovered the original growth rates. Unevolved B382 was also used to inoculate an anaerobic culture with 0.1 g/L yeast extract supplementation. All three cultures were grown to mid-exponential phase, and were then stopped by adding 5 mL 5% trizol/ethanol solution to 45 mL culture. The stopped cultures were centrifuged, and the RNA in the pellets was purified using an RNeasy kit (Qiagen, Hilden, Germany) including the DNAse digestion step. Purified RNA was quantified using a NanoDrop spectrophotometer (Thermo

Scientific, Waltham, MA) and quality was assessed using an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA). Ribosomal RNA was removed using an Invitrogen Ribominus kit (Life Technologies, Carlsbad, CA) and successful rRNA subtraction was confirmed with the bioanalyzer. The remaining RNA was fragmented by mixing with a fragmentation buffer (Life Technologies, Carlsbad, CA), incubating at 70°C for 3 min, and stopping with 1 μL 200 mM EDTA. This fragmented RNA was then ethanol precipitated, and the first DNA strand was synthesized by PCR using random primers, SUPERase-In RNAse inhibitor (Life Technologies, Carlsbad, CA), and SuperScript III reverse transcriptase (Life Technologies, Carlsbad, CA). This DNA was purified by a phenol:chloroform:IAA extraction, and was then ethanol precipitated. The second DNA strand was then synthesized using a dNTP mixture containing dUTP instead of dTTP, allowing for only the positive mRNA strands to be sequenced. The double stranded DNA was purified with a QIAquick PCR clean-up spin column (Qiagen, Hilden, Germany). The ends of the DNA strands were repaired with an Epicentre End Repair Buffer and Enzyme Mix (Illumina, San Diego, CA), and again purified with a QIAquick column. dA tails were added to the fragments using a dA-Tailing Buffer and Klenow Fragment (New England Biolabs, Ipswich, MA), and dA-tailed DNA was purified a QIAquick MinElute column. Next, Illumina adaptors were added to the ends of the fragments using a DNA Adaptor Mix and Quick T4 DNA Ligase (New England Biolabs, Ipswich, MA), and the DNA was again purified with a QIAquick MinElute column. The purified DNA fragments were then run in a 2% agarose gel and were stained with SYBR Gold (Life Technologies, Carlsbad, CA). The fragments between 180 and 500 bp were excised from the gel and purified with a QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany).

Next, the dUTP nucleotides were removed from the second strands of the DNA by a USER (New England Biolabs, Ipswich, MA) digestion. The USER treated DNA was amplified by qPCR using indexed Illumina PCR primers that allow for multiplex sequencing. Unevolved BOP382 DNA was amplified with multiplex index 1 (CGTGAT), evolved BOP382 GA was amplified with index 11 (GTAGCC), and BOP382 GB was amplified with index 12 (TACAAG). Phusion DNA polymerase was used (New England Biolabs, Ipswich, MA), and the amplification was stopped shortly after exponential amplification began. The amplified Illumina libraries were purified with Agencourt AMPure beads (Beckman Coulter, Pasadena, CA). To confirm that the correct adaptors and primers were ligated to *E. coli* mRNA sequences, library DNA was TOPO cloned into zero-blunt vectors which were transformed into TOP10 *E. coli* cells (Life Technologies, Carlsbad, CA) and grown overnight. The *E. coli* colonies were PCR amplified with M13 primers and were Sanger sequenced by Eton Bioscience, San Diego, CA. After validating the libraries, they were sent to the UCSD BIOGEM facility for high-throughput sequencing (Illumina, San Diego, CA).

The raw Illumina sequence data was mapped to the *E. coli* K-12 MG1655 genome (obtained from EcoCyc [20]) using Bowtie [47]. Reads were counted and analyzed for significance of differential expression using Cufflinks [48]. To determine which genes were expressed and which were not, the logarithms of the fragments per kilobase of transcript per million mapped reads (FPKM) values were plotted and a cutoff was defined based on inspection of the values. The FPKM values were confirmed to be log-normally distributed. To compare the measured expression values to model predicted phenotypes, pFBA was used [41]. The default pFBA function in the COBRA Toolbox 2.0 [49] was

used. The *i*JO1366 model was used, with bounds on exchange reactions set to anaerobic with a glucose uptake rate of 10 mmol/gDW/h, and with the reactions *ACALD*, *ALCD2x*, *LDH_D*, and *MDH* constrained to zero. Each gene in the *i*JO1366 model was classified as either "essential," "optimal," "enzymatically less efficient," "metabolically less efficient," "zero flux," or "blocked." The number of experimentally determined expressed genes were then determined for each category.

Acknowledgements

Chapter 6 is, in part, adapted from a paper that is being prepared for publication under the title "Adaptive laboratory evolution and characterization of computationally designed growth-coupled production strains." The dissertation author was the primary author of this paper, which was coauthored by Adam M. Feist and Bernhard Ø. Palsson.

We would like to thank Mallory Embree, Dae-Hee Lee, Harish Nagarajan, Vasiliy Portnoy, Douglas Taylor, and Karsten Zengler for their helpful comments, insights, and help with experimental procedures.

References

1. Feist, A.M., et al., *Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli.* Metab Eng, 2010. **12**(3): p. 173-86.

2. Lee, S.Y., D.Y. Lee, and T.Y. Kim, *Systems biotechnology for strain improvement.* Trends Biotechnol, 2005. **23**(7): p. 349-58.

3. Barrett, C.L., et al., *Systems biology as a foundation for genome-scale synthetic biology.* Curr Opin Biotechnol, 2006. **17**(5): p. 488-492.

4. Kim, H.U., T.Y. Kim, and S.Y. Lee, *Metabolic flux analysis and metabolic engineering of microorganisms.* Molecular BioSystems, 2008. **4**(2): p. 113-120.

5. Burgard, A.P., P. Pharkya, and C.D. Maranas, *Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization.* Biotechnol Bioeng, 2003. **84**(6): p. 647-57.

6. Patil, K.R., et al., *Evolutionary programming as a platform for in silico metabolic engineering.* BMC Bioinformatics, 2005. **6**: p. 308.

7. Feist, A.M., et al., *A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.* Mol Syst Biol, 2007. **3**(121).

8. Pharkya, P., A.P. Burgard, and C.D. Maranas, *OptStrain: a computational framework for redesign of microbial production systems.* Genome Res, 2004. **14**(11): p. 2367-76.

9. Ranganathan, S., P.F. Suthers, and C.D. Maranas, *OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions.* PLoS Comput Biol, 2010. **6**(4): p. e1000744.

10. Kim, J. and J.L. Reed, *OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains.* BMC Syst Biol, 2010. **4**: p. 53.

11. Lun, D.S., et al., *Large-scale identification of genetic design strategies using local search.* Mol Syst Biol, 2009. **5**: p. 296.

12. Yang, L., W.R. Cluett, and R. Mahadevan, *EMILiO: a fast algorithm for genome-scale strain design.* Metab Eng, 2011. **13**(3): p. 272-81.

13. Yim, H., et al., *Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol.* Nat Chem Biol, 2011. **7**(7): p. 445-52.

14. Fong, S.S., et al., *In silico design and adaptive evolution of Escherichia coli for production of lactic acid.* Biotechnol Bioeng, 2005. **91**(5): p. 643-8.

15. Orth, J.D., I. Thiele, and B.Ø. Palsson, *What is flux balance analysis?* Nat Biotechnol, 2010. **28**(3): p. 245-8.

16. Portnoy, V.A., D. Bezdan, and K. Zengler, *Adaptive laboratory evolution--harnessing the power of biology for metabolic engineering.* Current opinion in biotechnology, 2011. **22**(4): p. 590-4.

17. Conrad, T.M., N.E. Lewis, and B.Ø. Palsson, *Microbial laboratory evolution in the era of genome-scale science.* Mol Syst Biol, 2011. **7**: p. 509.

18. Ibarra, R.U., J.S. Edwards, and B.Ø. Palsson, *Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth.* Nature, 2002. **420**(6912): p. 186-9.

19. Datsenko, K.A. and B.L. Wanner, *One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products.* Proc Natl Acad Sci U S A., 2000. **97**(12): p. 6640-5.

20. Sawers, G. and A. Bock, *Novel transcriptional control of the pyruvate formate-lyase gene: upstream regulatory sequences and multiple promoters regulate anaerobic expression.* J Bacteriol, 1989. **171**(5): p. 2485-98.

21. Yang, J., et al., *The iron-sulfur cluster of pyruvate formate-lyase activating enzyme in whole cells: cluster interconversion and a valence-localized [4Fe-4S]2+ state.* Biochemistry, 2009. **48**(39): p. 9234-41.

22. Suppmann, B. and G. Sawers, *Isolation and characterization of hypophosphite--resistant mutants of Escherichia coli: identification of the FocA protein, encoded by the pfl operon, as a putative formate transporter.* Mol Microbiol, 1994. **11**(5): p. 965-82.

23. Reizer, J., A. Reizer, and M.H. Saier, Jr., *Novel phosphotransferase system genes revealed by bacterial genome analysis--a gene cluster encoding a unique Enzyme I and the proteins of a fructose-like permease system.* Microbiology, 1995. **141 (Pt 4)**: p. 961-71.

24. Steiert, P.S., L.T. Stauffer, and G.V. Stauffer, *The lpd gene product functions as the L protein in the Escherichia coli glycine cleavage enzyme system.* J Bacteriol, 1990. **172**(10): p. 6142-4.

25. Mat-Jan, F., K.Y. Alam, and D.P. Clark, *Mutants of Escherichia coli deficient in the fermentative lactate dehydrogenase.* J Bacteriol, 1989. **171**(1): p. 342-8.

26. Dym, O., et al., *The crystal structure of D-lactate dehydrogenase, a peripheral membrane respiratory enzyme.* Proc Natl Acad Sci U S A, 2000. **97**(17): p. 9413-8.

27. Sutherland, P. and L. McAlister-Henn, *Isolation and expression of the Escherichia coli gene encoding malate dehydrogenase.* J Bacteriol, 1985. **163**(3): p. 1074-9.

28. Leonardo, M.R., P.R. Cunningham, and D.P. Clark, *Anaerobic regulation of the adhE gene, encoding the fermentative alcohol dehydrogenase of Escherichia coli.* J Bacteriol, 1993. **175**(3): p. 870-8.

29. Shone, C.C. and H.J. Fromm, *Steady-state and pre-steady-state kinetics of coenzyme A linked aldehyde dehydrogenase from Escherichia coli.* Biochemistry, 1981. **20**(26): p. 7494-501.

30. Ferrandez, A., J.L. Garcia, and E. Diaz, *Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of Escherichia coli K-12.* J Bacteriol, 1997. **179**(8): p. 2573-81.

31. Kakuda, H., et al., *Identification and characterization of the ackA (acetate kinase A)-pta (phosphotransacetylase) operon and complementation analysis of acetate utilization by an ackA-pta deletion mutant of Escherichia coli.* J Biochem (Tokyo), 1994. **116**(4): p. 916-22.

32. Segre, D., D. Vitkup, and G.M. Church, *Analysis of optimality in natural and perturbed metabolic networks.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15112-7.

33. Fong, S.S., A.R. Joyce, and B.Ø. Palsson, *Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states.* Genome Res, 2005. **15**(10): p. 1365-72.

34. Fong, S.S. and B.Ø. Palsson, *Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes.* Nat Genet, 2004. **36**(10): p. 1056-58.

35. Lee, D.H., et al., *Cumulative number of cell divisions as a meaningful timescale for adaptive laboratory evolution of Escherichia coli.* PloS one, 2011. **6**(10): p. e26172.

36. Conrad, T.M., et al., *Whole-genome resequencing of Escherichia coli K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations.* Genome Biol, 2009. **10**(10): p. R118.

37. Lee, D.H. and B.Ø. Palsson, *Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol.* Applied and environmental microbiology, 2010. **76**(13): p. 4158-68.

38. Orth, J.D., et al., *A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011.* Mol Syst Biol, 2011. **7**: p. 535.

39. Fujii, T., et al., *Molecular and functional characterization of an acetyl-CoA acetyltransferase from the adzuki bean borer moth Ostrinia scapulalis (Lepidoptera: Crambidae).* Insect biochemistry and molecular biology, 2010. **40**(1): p. 74-8.

40. Akhtar, M.K. and P.R. Jones, *Construction of a synthetic YdbK-dependent pyruvate:H2 pathway in Escherichia coli BL21(DE3).* Metab Eng, 2009. **11**(3): p. 139-47.

41. Lewis, N.E., et al., *Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models.* Mol Syst Biol, 2010. **6**: p. 390.

42. Misra, K., et al., *Glyoxalase III from Escherichia coli: a single novel enzyme for the conversion of methylglyoxal into D-lactate without reduced glutathione.* Biochem J, 1995. **305 (Pt 3)**: p. 999-1003.

43. Kim, I., et al., *Screening of genes related to methylglyoxal susceptibility.* Journal of microbiology, 2007. **45**(4): p. 339-43.

44. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.

45. Anderson, M.J., et al., *Crystal structure of a hyperactive Escherichia coli glycerol kinase mutant Gly230 --> Asp obtained using microfluidic crystallization devices.* Biochemistry, 2007. **46**(19): p. 5722-31.

46. Applebee, M.K., et al., *Functional and metabolic effects of adaptive glycerol kinase (GLPK) mutants in Escherichia coli.* J Biol Chem, 2011. **286**(26): p. 23150-9.

47. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

48. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

49. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox 2.0.* Nat Protoc, 2011.