Reconstruction, Reconciliation, and Validation of Metabolic Networks

A DISSERTATION

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Elias William Krumholz

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Igor IG Libourel

May 2018

Acknowledgements

Foremost, I would like to gratefully acknowledge my advisor, Dr. Igor IG Libourel, who provided me the opportunity and guidance to pursue my PhD in the biological sciences. Throughout my PhD career, Dr. Libourel shared his knowledge, wisdom, and passion for scientific rigor and integrity that shaped my understanding and approach to scientific endeavors. In addition, Dr. Libourel graciously provided emotional support and encouragement at critical moments in my development as a scientist, for which I am very grateful.

I would also like to acknowledge the support and guidance of my thesis committee members, Professors Fumiaki Katagiri, Chad Myers, and Nathan Springer, all of whom thoughtfully shared scientific, career, and life advice both explicitly and through example. I am very grateful for their professionalism and commitment to understanding and guiding my research. I feel very fortunate to have their mentorship.

I would like to thank all of my graduate cohort and friends in the graduate programs. Together we shared a unique experience that will bind us together. Not only do I value the kindness, humor, and camaraderie, but I am also grateful for the high standard set by their actions, which motivated me to do my best and to keep pushing.

Finally, I am forever grateful for my family, close friends, and Carly Schramm, who have encouraged me in all that I have done and have supported me through thick and thin – thank you.

Dedication

This thesis is dedicated to my parents, Bill and Phyllis Krumholz, who encouraged me in all my endeavors and provided me with countless opportunities to learn and grow throughout my life.

Table of Contents

List of Tables viii

List of Figures ix

Chapter 1: Introduction 1

Metabolic Networks Provide a Scaffold for Systems Biology 1

Automated Gene Annotation Provides a Functional Parts List of a Cell 3

Metabolic Network Reconstruction Assembles Enzymes in a Computational Framework 4

Incomplete Networks Require Reconciliation with Observed Phenotypes 6

Network Constraints Aim to Align Simulated Phenotypes with Observed Phenotypes 9

Chapter 2: Genome-wide Metabolic Network Reconstruction of the Picoalga *Ostreococcus* 14

Synopsis 14

Introduction 15

Materials and Methods 18

Gene comparisons 20

Elemental balance of reactions 20

Reversibility index for reactions 21

Phylogenetic reconstruction 22

Network visualization 23

Results and Discussion 24

Functional network analysis 24

Gap-filling 24

Gap-filling of the *Ostreococcus* networks 27

Network comparison 28

Comparison to existing reconstructions 29

Conclusions 30

Tables 32

Table 1 32

Figures 33

Figure 1 33

Figure 2 35

Figure 3 37
Figure 4 39
Figure 5 40
Figure 6 41
CHAPTER 3: Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks 44
Synopsis 44
Introduction 45
Experimental Procedures 48
Metabolic networks, biochemistry database, and gene annotations 48
Identification of functional roles 49
Gene essentiality and metabolite production 50
Gap-filling algorithm 50
Software 51
Results 52
Metabolic networks require gap-filling 52
Network completion requires reactions with no enzyme sequence support (ESS) 53
Non-producible biomass metabolites are distributed across metabolism 55
Blast-weighted gap-filling 56
Quadratic programming reveals the gap-filling solution space 57
Comparison of computational and experimental gene essentiality 59
A subset of knockout predictions are sensitive to weight changes 60
Analysis of gap-filling reactions with high ESS values 62
Discussion 63
Tables 66
Table 2 66
Table 3 67
Table 4 68
Figures 69
Figure 7 69
Figure 8 70
Figure 9 71
Figure 10 72

Figure 11 73
Figure 12 74
Figure 13 76
Chapter 4: Thermodynamic Constraints Improve Metabolic Networks 77
Synopsis 77
Introduction 78
Experimental Procedures 82
Model SEED models and the reaction database 82
Network reconciliation 82
RA and RCR weighting vectors 84
Gene essentiality simulations 85
Sensitivity analyses 85
Results 86
Thermodynamically-informed constraints outperform random constraints 86
Uninformed constraints degrade predictions 87
Model predictions reveal tradeoffs in NR strategies 89
$\Delta_r G'^\circ$ values can qualitatively guide network reconciliation 91
$\Delta_r G'^\circ$ weighted NR rationally overturns heuristics 92
Corroboration through alternate network quality assessment 94
Discussion 94
Gene essentiality informs on network-wide parameters 96
Thermodynamically informed RCs improve metabolic networks 96
Weighted NR balances competing NR objectives 98
Figures 100
Figure 14 100
Figure 15 101
Figure 16 102
Figure 17 103
Figure 18 104
Figure 19 106
Figure 20 107
Figure 21 108
Chapter 5: Conclusions 109

References.....112

List of Tables

Table 1. Scale and functionality of networks. 32
Table 2. Model SEED database and model summary. 66
Table 3. Biomass components that require gap-filling. 67
Table 4. Essentiality predictions by gap-filled networks. 68

List of Figures

Figure 1. Stepwise database generation. 33
Figure 2. Network quality metric. 35
Figure 3. Comparison of top-down and bottom-up gap-filling. 37
Figure 4. Reactions added by gap-filling to *O. tauri* and *O. lucimarinus*. 39
Figure 5. Comparison of *O. lucimarinus* and *O. tauri*. 40
Figure 6. *O. tauri* compared with *O. lucimarinus* before and after gap-filling. 41
Figure 7. Gap-filling algorithm. 69
Figure 8. Comparison of metabolites that required unsupported reactions to become producible. 70
Figure 9. Role of reactions in *E. coli* gap-filling solutions. 71
Figure 10. LP vs. QP gap-filling. 72
Figure 11. Overlaps between BLAST-weighted QP, BLAST-weighted LP, uniformly-weighted LP, and Model SEED gap-filling. 73
Figure 12. *S. pneumoniae* gap-filling solutions. 74
Figure 13. Unsupported metabolite gap-filling reactions. 76
Figure 14. DOR and Percent Correct GEP for networks with randomly shuffled reaction direction constraints. 100
Figure 15. DOR of GEP as a function of uninformed constraints. 101
Figure 16. Network reconciliation overview. 102
Figure 17. NR approaches with highest DOR of GEP as function of rcrs. 103
Figure 18. Sensitivity analysis of NR parameters. 104
Figure 19. Reactions used in *E. coli* NR. 106
Figure 20. Alternate validating of reconciled networks. 107
Figure 21. Gene essentiality prediction confusion matrix. 108

Chapter 1: Introduction

Metabolic Networks Provide a Scaffold for Systems Biology

With the advent of whole genome sequencing, systems biology has emerged as a new approach to understanding biology. A chief goal of systems biology is to predict phenotypes from genotypes by creating comprehensive models of cells (*1*). Due to the massive scale of this challenge, cells are not modeled directly as the interaction of trillions of molecules, but instead as abstracted sub-models that simulate specific cellular processes (*2*). Metabolic network reconstruction is a subset of systems biology focused on associating elements of the genome with metabolic functions and reconstructing the network of chemical reactions that make up an organism's metabolism (*3*).

Following the sequencing and annotation of the *Haemophilus influenza* genome (*4*), the first draft of a complementary genome-scale metabolic network was reconstructed, encompassing 488 reactions and 343 metabolites (*5*). A key metric of network completeness for this network and all networks to follow was the production of vital chemical compounds, referred to as biomass, from available nutrients (*6*). Organisms that are known to live on well-defined nutrient sources should be able to produce all essential biomass metabolites from available nutrients and energy using enzymes encoded in their genome. The combination of well-defined chemical laws and relatively straightforward measures of success have allowed metabolic networks to become the first truly "genome-scale" models that can provide a framework for interpreting the function of genes at the scale of an entire cell.

Due to their relative simplicity and the depth of existing research, single-cell microorganisms such as *Haemophilus influenza* and *Escherichia coli* were early targets for metabolic network reconstruction (*5*, *7*). Ongoing efforts have greatly expanded the reconstructed networks of single-cell model organisms, producing multiple versions of *E. coli* and *Sacchromyces cerevisiae* networks (*8*, *9*). The scope of metabolic network reconstruction has also broadened to include many more bacteria (*10*), plants (*11*), animals (*12*, *13*), and humans (*14*, *15*). While early network reconstructions treated cellular metabolism as a single compartment of metabolites and enzymes, now networks incorporate multiple compartments each with well-defined transporters to model eukaryotic cells (*16–19*). The scope of genome-scale modeling of metabolism has steadily expanded to include more cellular systems, including gene expression (*20*), protein synthesis and stability (*21*, *22*), and microbial community interaction (*23–25*). Ambitious work is already underway to create comprehensive "whole cell" models integrating many systems into a coherent whole (*2*).

Although work to model whole cells has aimed to comprehensively model all major cellular systems, the modeling approach for each system necessarily makes assumptions and simplifications to allow for tractable computational simulations. The aphorism attributed to statistician George Box applies well here: "All models are wrong, but some are useful." As such, whole-cell models are only as valid and useful as the component sub-models that they are composed of, and further effort is required to improve the theory, methods, and validation of each. The thesis presented here focuses on advancing the reconstruction, reconciliation, and validation of metabolic networks of

microorganisms using genome sequences and thermodynamic properties of biochemical reactions.

Automated Gene Annotation Provides a Functional Parts List of a Cell

Metabolic networks are composed of a collection of annotated enzymes discovered in an organism's genome, so assembling all available gene annotations for a target organism is the first step in building the "parts list" of a metabolic network reconstruction. New enzymes are discovered and annotated frequently, but improvements in sequencing technology have allowed the rate of genome sequencing to rapidly outpace the ability to experimentally probe the function of individual genes (*26*). In response, automated genome annotation tools and services have been developed to keep pace with sequencing. Early approaches used gene comparison techniques, such as BLAST (*27*, *28*) and HMMer (*29–31*), to compare new sequences to known genes in order to assign high confidence gene annotations (*32*, *33*). Newer approaches use ensembles of genes to guide annotation rather than relying on individual comparisons. One notable approach is the Rapid Annotation using Subsystems Technology (RAST) service, which developed the concept of functional subsystems to guide gene annotation (*34–36*). Subsystems are expert-defined modules of genes that together carry out a coherent function. All constituent parts of the subsystem are thought to be present in the organism's genome for the function to operate. For instance, if 90% of genes in the isoprenoid biosynthesis subsystem were annotated, the remaining 10% of missing subsystem roles could be prioritized for more rigorous analysis with the assumption that they exist in the genome,

but have not yet been found. In this way, the many elements of a genome are used to mutually inform on individual gene function.

Metabolic Network Reconstruction Assembles Enzymes in a Computational Framework

While gene annotations from services such as RAST provide a "parts list" of enzymes for a draft metabolic network reconstruction, the list of enzymes needs to be converted into a format that can represent the "network" aspect of metabolism. Several services and tools have been developed to facilitate the initial reconstruction process of integrating the enzymes into a coherent model of metabolism (*10*, *36–42*). The Kyoto Encyclopedia of Genes and Genomes (KEGG) provides a large database of enzymes and associated reactions and metabolites, while also providing comprehensive human readable maps of metabolism and extensive literature citations (*37*). MetaCyc is another database that catalogs metabolic pathways and provides an online environment for viewing genome elements, pathway maps, logical relations between genes and metabolic functions, as well as the related Pathway Tools software suite for computationally working with the database (*42*). EcoCyc is a subset of the MetaCyc database that catalogs pathways and related information specific to *E. coli*, enabling a systems biology approach to researching the important model organism (*43*). Metabolic databases such as KEGG and MetaCyc are especially important for network reconstruction because they represent the space of known metabolic reactions, and all organism-specific reconstructions are necessarily a subset of a metabolic database. The Biochemical, Genetic and Genomic (BiGG) knowledgebase created at the University of San Diego took a similar approach

but instead focused on creating computational models of metabolism, with a primary focus on building constraint-based models, which define a space of allowable metabolic phenotypes that are constrained by properties such as reaction irreversibility and available nutrient transport mechanisms (*41*). The COnstraints Based Reconstruction and Analysis (COBRA) toolbox, which was developed in parallel to BiGG, provided tools for manipulating metabolic network representations, adding or removing constraints, and performing computational analyses (*40*, *44*). The Raven Toolbox is similar to the COBRA toolbox but adds more integrated reconstruction tools to facilitate automated reconstruction of draft metabolic networks (*39*). Recently, the Model SEED (*10*, *11*) and the DOE KBASE (*38*) projects have aimed to provide comprehensive services from genome annotation to metabolic network reconstruction.

The Model SEED provides an online web service that automatically assembles the annotations returned from RAST into a draft metabolic network in common computable formats, such as the Systems Biology Markup Language (SBML) (*45*). In addition to assembling the draft network, the Model SEED also generates a hypothetical list of biomass components that an organism should be able to produce, given the bacterial type (e.g. gram positive or gram-negative bacteria). Reactions are automatically added to the network to allow the organism to simulate biomass production in the presence of a rich media source (*10*). This process is known as gap-filling, or network reconciliation, as the network is reconciled with an observed phenotype of biomass synthesis from available nutrients.

Extensive databases of metabolic reactions are crucial for finding network reconciliation solutions that expand a metabolic network and allow it to simulate observed phenotypes (*3*). Since databases such as KEGG, MetaCyc, and BiGG were developed independently, the databases use different chemical identifiers and nomenclature that are often incompatible. Comparing metabolism between different metabolic databases remains a major challenge, spurring efforts to create comprehensive maps between databases (*46*) and adopt rigorous chemical naming standards (*47*, *48*). The Model SEED used the KEGG database as a starting point for assembling a metabolic database, but carefully validated all chemical reactions to ensure no stoichiometric inaccuracies were introduced into draft reconstructions (*10*). This attention to detail allowed the draft reconstructions to be computationally analyzed using the COBRA toolbox, whereas uncorrected stoichiometric errors would allow network simulations to violate physical laws, such as the conservation of matter and energy (*49*, *50*). By creating a full reconstruction pipeline with a common database of enzyme, reaction, and metabolite definitions, the Model SEED generated thousands of metabolic networks which could be analyzed with computational methods and compared between organisms (*10*, *51*).

Incomplete Networks Require Reconciliation with Observed Phenotypes

Despite extensive gene annotation efforts, network reconciliation is required for all metabolic networks assembled to date. Even well-studied organisms such as *E. coli* and *S. cerevisiae* are not annotated in sufficient detail to reconstruct a functioning

metabolic network capable of producing biomass from high confidence gene annotations alone (*9*, *52*, *53*). Considerable effort has been devoted to network reconciliation, with initial work focused on developing algorithms to find solutions that made the fewest number of network modifications (*53*, *54*). Soon after, other types of data were used to select solutions that optimized various parameters, including sequence similarity (*49*, *55*–*58*), and predicted thermodynamic reversibility (*10*, *59*). Computational efficiency improvements were also pursued to reduce the time required to find gap-filling solutions (*56*, *60*).

Metabolic network reconciliation is accomplished through two primary approaches: relaxation of reaction directionality constraints and addition of transporters and reactions to a metabolic network without necessarily requiring a corresponding gene and enzyme to be identified (*53*, *61*). Relaxation of reaction directionality constraints can be simulated by adding fully reversible reaction versions in place of irreversible versions that may exist in a model; in this way, network reconciliation can be approached more simply as just the addition of new reactions to a metabolic network (*53*, *56*). A reconciliation solution is a set of reactions that when added to the metabolic network allow for the simulation of an observed phenotype, such as biomass production from available nutrients (*61*). Early reconciliation approaches, such as the SMILEY algorithm (*54*), took a bottom-up approach to gap filling, where reactions were iteratively added to metabolic networks until a gap-filling solution was found. Mixed integer linear programming (MILP) was used to search through possible solutions until a minimal solution was discovered and a variety of extensions to this approach were developed (*53*,

62–64). While bottom-up approaches showed success in finding reconciliation solutions, it was computationally intensive and was not guaranteed to find a solution with the fewest possible number of added reactions, and in some cases, bottom-up approaches were unable to find reconciliation solutions that were possible (*49*).

Top-down network reconciliation is an alternative approach to bottom-up reconciliation, where many reactions are added to a metabolic network initially, and then iteratively removed if their removal does not prevent biomass production (*54*, *55*). By adding an entire metabolic database to an organism-specific network, all biomass metabolites that could be produced with any reconciliation algorithm can be found, since all potential reconciliation reactions are present (*49*, *55*, *56*). This allows top-down gap-filling to achieve the maximal reconciliation of biomass production, whereas bottom-up reconciliation can fail to reconcile the production of individual biomass metabolites (*49*).

Additional methods have been developed to preferentially select reconciliation reactions based on other sources of biological data. For instance, a top-down reconciliation approach developed by Christian et al. (*55*), used BLAST to quantify the support for a given reconciliation reaction and then preferentially removed unsupported reactions in a stochastic fashion to select more accurate reconciliation reactions. The consequence of tradeoffs between top-down and bottom-up had not been rigorously analyzed, and important questions remained about what approach could yield more accurate networks. Top-down and bottom-up reconciliation approaches are explored in depth in chapter 2 of this thesis, along with a hybrid algorithm that builds on the method of Christian et al. to preferentially add reactions from phylogenetically related organisms.

Early gap-filling approaches that relied on MILP led to multiple solutions due to the stochastic nature of MILP algorithms (*49*, *53*). This made it difficult to replicate specific reconciliation solutions and it was unclear if simple optimization strategies, such as minimizing the number of reconciliation reactions would yield realistic reconciliation solutions. Furthermore, the space of possible reconciliation solutions was not well understood, although it was clear that multiple solutions existed. New reconciliation algorithms are put forth in chapter 3 and 4, where linear programming (LP) and quadratic programming (QP) are used in place of MILP to find unique solutions. QP is also used to probe the space of possible reconciliation solutions.

Network Constraints Aim to Align Simulated Phenotypes with Observed Phenotypes

Constraint-based models of metabolism have become a standard representation of metabolic network reconstructions (*40*). Constraints are applied to metabolic networks to limit the space of phenotypes that a metabolic network can simulate, with the goal of aligning the model's feasible space with observed metabolic phenotypes (*44*, *65*). Examples of constraints include defining the types of transporters that allow chemicals in the environment to pass through the cellular membrane and enter the compartments of the cell (*17*, *18*), as well as constraining the direction that a chemical reaction can proceed in, effectively making certain chemical reactions irreversible in the metabolic network (*66*–*69*).

Simulating steady-state metabolic flux was a primary early motivation for developing constraint-based models, and flux balance analysis (FBA) was developed in

parallel to constraint-based models (*65*, *70*, *71*). FBA uses LP to calculate the intake of nutrients, excrement of waste products, production of biomass metabolites, and balance of energy carrying molecules while simultaneously accounting for reaction stoichiometry, reaction reversibility, upper and lower reaction rate bounds, and maximizing or minimizing a hypothesized cellular objective (*65*). FBA held great promise to massively simplify the calculation of metabolic flux, which would otherwise require computationally intensive solutions to differential equations for each metabolic reaction, and early results showed agreement between measured growth rates and predicted growth rate (*72*, *73*). However, the prediction of growth rate was directly tied to the validity of the hypothetical cellular objective optimized for using FBA (*6*), and recent results for some metabolic networks have shown only weak agreement with measured values, even with the consideration of multiple objective functions (*74–76*). Despite the potential shortcomings of FBA, it still has valuable applications, such as testing network completeness by efficiently searching for the existence of biomass production pathways, yielding Boolean growth or no-growth calls in place of quantitative growth rates (*52*, *56*, *77*).

Further constraints to metabolic networks can define the relationship between genes, enzymes, and reactions using Boolean relationships (*3*). Gene-to-enzyme-to-reaction mapping allows for gene deletions to be simulated in terms of the resulting loss of chemical reactions in a metabolic network (*77*). Simulated gene deletions can be compared to experimental gene deletions to assess the quality of the metabolic network and the constraints applied to the network (*8*, *52*, *56*). Computational gene essentiality is

a valuable metric that can be compared to experimental gene essentiality without assuming detailed quantitative network parameters related to the rates of reactions in an organism. Experimental gene deletions and observed growth or no growth on a variety of nutrient sources can be used to guide the application of constraints with the goal of aligning simulated phenotypes with experimental phenotypes. Several algorithms have been developed to improve metabolic networks using these experimental data (*62*, *78*, *79*). Experimental gene essentiality is assessed by attempting to create a mutant organism lacking a single gene, this approach has been systematically performed for every known gene in well studied microorganisms such as *E. coli* (*80*), but is also becoming possible in many new microorganisms by using transposon mutagenesis (*81*) and the CRISPER/CAS9 system (*82*).

Despite increasing ability to delete genes, determining if a gene is essential is not necessarily straightforward. The essentiality of a gene is strongly determined by the specific environmental conditions and available nutrients. To account for this, mutant libraries are typically grown on a well-defined nutrient source in standardized conditions. If a viable mutant cannot be created, or cannot sustain sufficient growth, the gene may be considered experimentally essential, but confounding variables must always be considered. Ideally, essential genes are validated by repetition using multiple knockout strategies, positive controls where the gene function is independently added back into the organism, and wildtype negative controls. However, due to the large scale of mutant libraries, such validation is often not yet feasible. Nonetheless, even imperfect calls on experimental gene essentiality can provide valuable data points for network validation

when interpreted with care (*83*). Computational gene essentiality is simulated by deleting a metabolic network gene, and in turn any reactions that are uniquely associated with that gene's enzymatic product. After the gene is deleted, the production of biomass metabolites is tested in conditions that mirror the experimental nutrient conditions, and if any essential biomass metabolite can no longer be produced, then the gene is considered computationally essential.

Genes can be classified into four categories when comparing computational gene essentiality to experimental gene essentiality. Two are considered correct predictions: computationally essential (CE) and experimentally essential (EE), as well as computationally nonessential (CNE) and experimentally nonessential (ENE). The other two cases, CE-ENE and CNE-EE, are considered incorrect predictions. Metrics such as the percent of correct predictions can be used to compare the quality of metabolic networks (*8*, *56*). It should be noted, however, that disagreements between computational and experimental essentiality do not necessarily guarantee that the network is incorrect (*3*). For instance, it is possible that a gene is experimentally essential, but computationally non-essential because the experimental deletion affects a process that is not modeled by the metabolic network, such as gene regulation. In contrast, genes that are computationally essential but experimentally non-essential indicate definite errors in the metabolic network, since some mechanism that is not modeled in the network, but exists in the actual organism allows the organism to grow without the associated gene product (*56*).

While gene essentiality can provide valuable experimental validation that probes the structure of metabolic networks, the sensitivity of gene essentiality to changes in the metabolic network has not been reported in detail. This has particular relevance to the application of network constraints, and network reconciliation, both of which affect the outcome of gene essentiality simulations. Furthermore, since gene essentiality simulations are derived from data that is independent of the application of constraints and reconciliation reactions, it has the potential to report on the quality of constraints and reconciliation algorithms. Important questions that have remained unanswered are: 1) Do constraints, such as reaction reversibility constraints, demonstrably improve the quality of metabolic networks? 2) Can biological data such as sequence similarity, or chemical data such as Gibbs free energy estimates, be used to improve network reconciliation when compared to parsimonious reconciliation algorithms that aim to make the fewest network modifications? Both of these questions require negative controls to establish baselines for network quality and ensure that rational approaches have significant benefits over randomized approaches.

The thesis research presented here focuses on the reconstruction of organism-specific metabolic networks from genome annotations and methods for improving metabolic networks by reconciling them with observed phenotypes, specifically the synthesis of essential biomass metabolites. Gene sequence similarity and estimations of thermodynamic reaction parameters are used to guide network reconciliation through the use of numerical optimization algorithms. Particular attention is devoted to the validation of metabolic networks using experimental data, such as gene

essentiality, and the development of computational controls using parameter randomization.

Chapter 2: Genome-wide Metabolic Network Reconstruction of the Picoalga *Ostreococcus*

Synopsis

The green picoalga *Ostreococcus* is emerging as a simple plant model organism, and two species, *O. lucimarinus* and *O. tauri*, have now been sequenced and annotated manually. To evaluate the completeness of the metabolic annotation of both species, metabolic networks of *O. lucimarinus* and *O. tauri* were reconstructed from the KEGG database, thermodynamically constrained, elementally balanced, and functionally evaluated. The draft networks contained extensive gaps and, in the case of *O. tauri*, no biomass components could be produced due to an incomplete Calvin cycle. To find and remove gaps from the networks, an extensive reference biochemical reaction database was assembled using a stepwise approach that minimized the inclusion of microbial reactions. Gaps were then removed from both *Ostreococcus* networks using two existing gap-filling methodologies. In the first method, a bottom-up approach, a minimal list of reactions was added to each model to enable the production of all metabolites included in our biomass equation. In the second method, a top-down approach, all reactions in the reference database were added to the target networks and subsequently trimmed away based on the sequence alignment scores of identified orthologues. Because current gap-filling methods

do not produce unique solutions, a quality metric that includes a weighting for phylogenetic distance and sequence similarity was developed to distinguish between gap-filling results automatically. The draft *O. lucimarinus* and *O. tauri* networks required the addition of 56 and 70 reactions, respectively, in order to produce the same biomass precursor metabolites that were produced by our plant reference database.

Supplementary material is available at Journal of Experimental Botany online, URL: https://academic.oup.com/jxb DOI: 10.1093/jxb/err407

Introduction

Due to the large research investments in genome projects and the rapid advancement of sequencing technologies, the number of sequenced genomes is growing exponentially (*26*, *84*). These sequences have great potential value, but their use is limited by the amount of time and effort required functionally to annotate a genome. Genes annotated with metabolic reactions are readily interpret- able at the biochemical reaction level, but their metabolic function is dependent on which other reactions are present. Flux balance analysis (FBA) (*65*) performs such a functional evaluation and has the capability to evaluate in which metabolic functions a reaction participates. FBA is therefore an excellent technology to evaluate the annotations for metabolic genes (*85*). Before such functional analysis can be performed, all the reactions associated with annotated metabolic genes must first be aggregated into a metabolic network. For prokaryotes, metabolic network reconstruction has become routine and, in many cases, sequence annotation and network reconstruction can be produced in a fully automated fashion (*10*). The quality of such machine annotations is dependent on the ability to take contextual

information into consideration during the annotation process. For instance, prokaryote annotation algorithms take the location of a gene relative to other functionally related genes into account. Eukaryotic genomes have much greater complexity, and the location of genes in eukaryotic genomes is much less informative. Consequently, annotation methods developed for prokaryotes have struggled when applied to plant genomes, requiring that these genomes still be annotated by expert teams (*86*). A metabolic network by itself can potentially provide a wealth of contextual information that is also applicable to eukaryotic systems. Metabolism can be viewed as multifaceted, highly interdependent machinery, containing functionality that is easily computer interpretable. Missing or superfluous reactions in the metabolic network can be readily identified and addressed by modifying the network in such a way that the functional metabolic unit is restored.

Here FBA has been applied on metabolic networks to evaluate the completeness of metabolic annotations for two *Ostreococcus* species. The prevalent marine microalga *Ostreococcus* (*87*) is an ideal model organism in plant biology due to its simplicity and its phylogenetic position as an early-diverging green plant lineage (*88*, *89*). *O. tauri* is the smallest known existing eukaryote (<1 μm), it can be kept in culture and can be genetically transformed (*90*). *Ostreococcus* is haploid (*90*, *91*) and has a single copy mitochondrion and chloroplast. *Ostreococcus* has been discovered relatively recently (*92*, *93*), but its importance is broadly recognized which has resulted in over 150 scientific publications, of which 70 were published in just the last two years. The significance of *Ostreococcus* is further exemplified by the complete genome sequencing of three species,

and the resequencing of 15 more. Two of these genome sequences, *O. tauri* and *O. lucimarinus*, have been manually annotated (*88*, *89*), setting the stage for metabolic network analysis.

Besides the quality of an organism's annotation, the ability to reconstruct its metabolic network algorithmically depends on a well-curated biochemical reaction database with gene-to-reaction associations. Gene-to-reaction mappings are organized in an orthology database that associates sequences of individual species with a biochemical reaction and allows for the identification of probable homology between organisms. FBA requires that the reactions included within the metabolic network be balanced at the element level. If a reaction is not elementally balanced, FBA will produce biologically meaningless solutions. For instance, a network that contains an oxygen unbalanced reaction might apply that reaction as part of a cycle consuming all oxygen produced by photosynthesis. One of the best known large ontology databases associated with biochemical reactions is produced by KEGG (*94*). This work makes use of the balanced subset of the KEGG Orthology (KO) database to associate the gene annotations with biochemical reactions. Flux balance analysis of the reconstructed draft networks reveals network functionality for some pathways, but more importantly a lack of functionality for others. Non-functional pathways can be gap-filled by adding reactions to the network until the demanded network functionality is achieved (*61*). Gap-filling requires a large reference database of reactions that may be used to fill the network gaps. For this purpose, the complete set of balanced reactions in the KEGG KO database was used. The KO database spans all kingdoms of organisms and many of the reactions exist in microbial

organisms only, making them unsuitable for the gap-filling of plant networks. To address this potential issue, a layered gap-filling approach was introduced, where the *Ostreococcus* networks were almost exclusively filled with reactions known to exist in the set of plants annotated in the KEGG database. This database of plant reactions has been called the meta-plant. The meta-plant database was curated using nested layers of the KEGG database in an attempt to retain the functionality of the complete KO database. Hence, the KO database biomass capability represents the maximum feasible biomass any model based on gene annotations from KEGG can achieve. Using a gap-filling algorithm, this functionality can be added to smaller databases through the addition of a minimal set of reactions from the KO database. The set of all eukaryotic and cyanobacteria annotations was gap-filled using the KO database to produce a reduced database, which was subsequently used to gap-fill the meta-plant.

The model systems *Ostreococcus*, *Arabidopsis*, and *Chlamydomonas* represent three clades that provide the full scope of green plant-specific genes: 'the green cut' (*95*). Curated genome-wide metabolic networks for *Arabidopsis* (*96*, *97*) and *Chlamydomonas* (*55*, *98*, *99*) already exist, and this work presents and compares the metabolic reconstructions of two *Ostreococcus* species.

Materials and Methods

Functional gene annotations were collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology (KO) database on 28 April 2011. This database contains mappings between the KEGG KO identifiers, organism-specific genes, predicted enzyme functionalities (EC numbers), and KEGG reactions. In addition, the KEGG reaction,

compound, and enzyme databases were downloaded on the same date in flat file format from the KEGG FTP website. The databases were loaded into Matlab (The MathWorks, Natick, MA) structures and organized according to the flat file field names.

To generate an SBML model from a KEGG genome, metabolic genes must be linked to metabolic reactions, but KEGG does not provide such a mapping. Instead, genes annotated with a functional role are assigned a KEGG orthology identifier (KO number). Most KO entries point directly to a set of reactions, all of which were included. If this was not the case, a KO entry often pointed to an Enzyme Classification (EC) identifier, in which case all reactions associated with the enzyme activity were added to the KO structure. The complete database mapping structure is shown in Supplementary Fig. S1 at JXB online.

The database structure was then reorganized to be rooted at the reaction level. Unique reaction identifiers were annotated with (potentially multiple) KO identifiers, EC numbers, and genes associated with these KO identifiers and EC numbers. After removal of unbalanced or incomplete reactions, SBML models (level 2, version 4) were generated using the System Biology toolbox (*100*). The SBML reaction field was populated with the organism-specific subset of the reaction database. Compounds were pulled from this reaction set and added to the SBML species field. Multi-organism models were generated by creating a union of organism-specific reaction databases. Each model was supplemented with a list of spontaneous reactions (see Supplementary Table S1 at JXB online). SBML models were subsequently converted to COBRA compliant format to access COBRA toolbox functionality. COBRA toolbox v2.0 (*40*) was downloaded from

the openCOBRA project at sourceforge.net. (http://opencobra.sourceforge.net/). The generated models were uncompartmentalized.

Gene comparisons

Genomes of organisms were downloaded in FASTA format from the KEGG database on 28 April 2011. A best gene match between genes in *O. tauri* and *O. lucimarinus* and genes in the union of KEGG plant genomes (meta-plant) was found using the Smith–Waterman algorithm (*101*) performed on a TimeLogic DeCypher (Active Motif Inc., Carlsbad, CA) gene comparison server. The union of plant genes for each reaction present in the KEGG database was used as a query sequence and the gene models of *O. tauri* and *O. lucimarinus* were used as the databases to search against. In this way, a mapping between each metabolic gene in the meta-plant genome and the best matching gene from both *O. tauri* and *O. lucimarinus* was created. Once the meta-plant model was complete, each reaction present in the meta-plant model was associated with the best scoring gene comparison. This method allowed every reaction in a large database to be annotated with a specific gene from an organism of interest and a corresponding gene from a plant database regardless of gaps in previous annotations or poor sequence similarity between available genes annotated with a particular reaction.

Elemental balance of reactions

The KEGG orthology database (28 April 2011) contained mappings to 4523 reactions. The elementary mass balance of each reaction was tested using a custom-developed Matlab routine. To prevent stoichiometric matrix errors, reactions that contained the same

metabolite as substrate and product were removed (see Supplementary Table S2 at JXB online). The results were verified with the elementary balancing functionality of the Cobra toolbox and no discrepancies were found. Generic compound equations containing (n) or R-groups were substituted with a large prime number for n or a large arbitrary group for R-groups to ensure that elemental balance was maintained in reactions with non-explicit formulas. In 601 reactions an imbalance in H, C, N, O, P or S, or an imbalance in n or R-groups was detected. These reactions were removed, with the exception of a small set of reactions that were manually balanced to retain the ability to reach five biomass precursor metabolites (see Supplementary Table S3 at JXB online).

Reversibility index for reactions

The reversibility of reactions was determined using the free energy calculations for reactions based on a group theory approach (*67*, *102*, *103*), which was further refined by the Milo laboratory (*104*). Elad Noor (Milo laboratory) kindly provided a custom reaction list adjusted to pH 7.5 and an ionic strength of 0.3 upon request. The reversibility index was generated according to the metric developed by Noor et al. (E Noor et al., unpublished data). Default metabolite concentrations were assumed to be 100 μM and allowed to vary between 3 μM and 3 mM which corresponds to an index cut-off value of 1000. Using these constraints, approximately half of the reactions in the KO database were considered irreversible. Reversibility information was included in the first two kinetic parameters of each reaction following the COBRA format. A reversible reaction was added as a chemical description of photosynthesis to allow the model to intake

energy. This reaction is listed as R99999 (equation: 2H2O+4 oxidized ferredoxin/4H → O2+4 reduced ferredoxin).

Phylogenetic reconstruction

The phylogenetic distance among all KEGG plant taxa was inferred from publicly available, fully sequenced and annotated genomes. Phylogenetic distance between species was estimated from six nuclear protein-coding genes: isoleucyl-tRNA synthetase, arginyl-tRNA synthetase, ribosomal protein L14, ribosomal protein S7, DNA-directed RNA polymerase alpha subunit, and DNA-directed RNA polymerase beta subunit (see Supplementary Table S6 at JXB online). These genes were previously identified by Ciccarelli et al. (*105*) as useful for reconstructing phylogenies among widely divergent taxa.

Gene sequences were downloaded from the KEGG Genome database using the KEGG ID as a search string. Many genes have variable copy number within and among taxa; therefore, single consensus sequences were generated for genes with multiple copies using Clustal X (*106*) by aligning the copies and generating a single, consensus sequence (see Supplementary Table S6 at JXB online). Gene sequences were then aligned across taxa using Clustal X with default parameters. The software package jModelTest 0.1(*107*) was used to select the best fitting-model of nucleotide evolution for each gene individually using the Akaike information criterion (AIC). The generalized time reversible (GTR) model with branch-specific evolutionary rates following a gamma distribution (GTR+G) and independent frequencies for each nucleotide (GTR+I+G) was chosen for isoleucyl-tRNA synthetase (K01870), while the GTR+G model with equal

nucleotide frequencies was chosen for all other genes. Genes were concatenated by hand. A maximum-likelihood (ML) tree was then inferred from the concatenated and partitioned genetic data set using Garli 2.0 (*108*). Models parameters estimated using jModelTest were used for each gene, and a cladogram based on current systematic knowledge (G Weiblen, University of Minnesota, personal communication) was enforced as a constraint to ensure accurate topology. All other parameters were left at the default values. The default termination criteria were used to determine when the run was complete. A Newick string with distances was converted to a distance matrix using the ape package for R 2.10.1 (R Core Development Team, 2009) (see Supplementary Table S7 at JXB online).

Network visualization

Cytoscape v2.8.1 (*109*) was used to generate metabolic network visualizations from SBML level 2 version 1 files. The advanced network merge plugin was used to create a difference network for *O. lucimarinus* against *O. tauri* and *O. tauri* against *O. lucimarinus*. The difference networks were combined with the union function to generate a complete difference network. The network was rendered with the VizMapper function using the yFiles 'organic' layout algorithm. Organism-specific reactions were identified in VizMapper by storing identifier strings in the sbml 'name' field of reactions (*109*).

Results and Discussion

Functional network analysis

FBA was used to investigate the ability of reconstructed networks to produce biomass components. The unedited *O. lucimarinus* and *O. tauri* networks were able to produce 18 and 0 biomass components, respectively (Table 1). Limited network functionality of draft networks reconstructed from genomic databases is not uncommon, and is predominantly caused by the presence of gaps in the network. Network gaps arise from missed annotations, and in the case of KEGG, disconnects between generic and specific definitions of the same metabolites. Other gaps arise from the removal of unbalanced equations during the network reconstruction process. The difference in network annotation between *O. tauri* and *O. lucimarinus* also suggest that *O. tauri* was annotated more conservatively. An example of the conservative annotation of *O. tauri* was the lack of Calvin cycle capability [ribulose-5-phosphate-3-epimerase (EC: 5.1.3.1) was not included in the draft network].

Gap-filling

Two complementary approaches to gap-filling of metabolic networks exist in the literature. Bottom-up gap-filling is based on a mixed integer optimization routine usually aiming to add a minimal number of reactions to a network (*54*). This method can distinguish between different classes of reactions either by adding reactions in a preferred order, or by associating different weights with the different reaction classes (*10*). The

bottom-up gap-filling method is the most commonly applied approach, and it was used for gap-filling all the reference databases described below.

A top-down approach to gap-filling, pioneered by (*55*), adds the complete gap-filling reaction database to a draft network followed by the iterative removal of the added reactions. This removal process is continued until no more added reactions can be removed without losing biomass production capability. Both the top-down and bottom-up methods are iterative approaches that do not have unique solutions. Candidate gap-filling reactions from species closely related to the target species are more likely to feature in the target species' network. By gap-filling reference networks of decreasing taxonomic diversity, a layered approach to gap-filling was used that takes advantage of this quality. (Fig. 1) With this in mind, gap-filling of the *Ostreococcus* networks was performed with a list of just the reactions that KEGG has associated with the 17 plant genomes in their database. This meta-plant network was not free of gaps itself, and has been filled with the combined reactions of eukaryotes and cyanobacteria. Similarly, the cyano–eukaryote model was gap-filled using the complete KEGG ontology reaction list. The complete network database was able to produce 44 out of the 94 defined biomass precursor metabolites, which increased to 49 after adding 12 previously H-unbalanced reactions (see Supplementary Table S3 at JXB online). After the consecutive gap-filling of increasingly small networks, the smaller networks approximated the biomass production capability of the ontology dataset (Table 1). This biomass component number fell well short of the target list of over 90 biomass precursors, but covered the fundamentals such as most amino acids and DNA. To expand the achievable biomass list, the number of

balanced reactions in the reference reaction database will have to be increased significantly beyond the current number of balanced reactions in the KEGG. To enable objective comparison between the different gap-filling solutions for the *Ostreococcus* networks, a quality score was developed. The quality score includes the sequence similarity score, and a weighting factor for the phylogenetic distance of the species to which sequences are compared.

To calculate this score, each reaction within the meta-plant network was associated with all sequences that were annotated with that reaction. Every gene associated with any reaction was than compared with all genes in the *O. tauri* and *O. lucimarinus* genomes using the Smith–Waterman algorithm. The Smith–Waterman *e*-score is a likelihood score applicable to a set of translated amino acids, which was weighted with the phylogenetic distance between respective species: $\omega = \sum_i -p_i/\log(e_i)$, where *e* is the Smith-Waterman *e*-score and p is the normalized phylogenetic distance score. Each sequence comparison thus yields a quality-score, and the best quality score, ω, for each reaction is assigned to that reaction in the meta-plant network: $\omega = \min(\omega)$, where j is the number of reference sequences associated with the reference reaction. The *e*-value was capped at a value of one to prevent sign inversion. The domain of ω is from 0 (perfect match) to N (very poor match), i.e. if a reaction had been annotated for the target organism, the value for that reaction would be zero (zero over minus infinity). A lack of a close sequence comparison would result in a value proportional to the phylogenetic distance and inversely proportional to the *e*-value. This quality metric allows for the rapid discrimination between gap-filling solutions (Fig. 2). Phylogenetic relationships among

plant species included in the KEGG database was inferred using a maximum likelihood approach, see the Materials and methods for a detailed explanation.

Gap-filling of the *Ostreococcus* networks

The *Ostreococcu*s networks were gap-filled using the meta-plant network as the reference reaction database. To enable the production of all 48 biomass components produced by the reference database, the *O. tauri* and *O. lucimarinus* networks gained 70 and 56 reactions respectively (Table 1). The iterative nature of the employed gap-filling methods resulted in multiple solutions. The results presented in Table 1, were selected based on their overall quality scores. For the *O. lucimarinus* network, alternative solutions did exist that included one less reaction, but these solutions had a lower quality score. Note that the number of reactions capable of carrying flux dramatically increased upon gap-filling, demonstrating the significantly improved connectivity of the networks. The included reactions with the worst sequence comparison (*O. lucimarinu*s: $e = 0.063$, *O. tauri*: $e = 0.21$) indicates that both networks were filled with at least one highly unlikely reaction. This is an unfortunate reflection of the lack of completeness of the gap-filling reaction database.

The bottom-up algorithm enabled the production of only a subset of the 48 biomass metabolites. A comparison of the two methods for the *O. tauri* network is shown in Fig. 3. For a valid comparison, the top-down method was made to fill the *O. tauri* network for the 36 biomass metabolites that the bottom-up method found. In this direct comparison between the best gap-filling solutions for the two methods, the bottom-up method used seven fewer reactions than the top-down approach. However, the top-down approach had

a better quality score, indicating a more realistic gap-filling solution. Finally, poor estimates of the imposed thermodynamic constraints could have led to incorrect reversibility constraints, causing unrealistic pathway shunts to accomplish the required biomass capability. However, if thermodynamic constraints had not been imposed, flux balance analysis could have found solutions that make use of thermodynamically infeasible pathways. That the thermodynamic constraints were active is readily demonstrated: FBA of the unconstrained *O. lucimarinus* network yielded a biomass flux of 2.079, compared to 0.275 for the constrained version, underlining the importance of accurate thermodynamic constraints. Other previously published studies also emphasize the importance of thermodynamic constraints to the accuracy of FBA models (*110*).

Network comparison

A large share of the reactions added during the gap-filling process were added to both networks (Fig. 4). Only a single reaction originally annotated for *O. tauri* was added to the *O. lucimarinus* network. Conversely, 13 reactions originally annotated for *O. lucimarinus* were added to the *O. tauri* network. Two and four reactions were added to the networks of *O. lucimarinus* and *O. tauri*, respectively, that were unique to the networks. Network changes resulting from gap-filling are shown in Fig. 5.

The differences between the draft *Ostreococcus* networks were visualized (Fig. 6) by calculating a difference network that only shows reactions exclusively present in only one network (logical XOR). The same difference network was generated after gap-filling, and the changes between the difference networks, which resulted from gap-filling, are shown in a third panel. Because only the connected differences between the networks are shown,

the connectivity of the difference network shows the alternative routes in central metabolism utilized by the *Ostreococcus* species connecting ribose and glyoxylate metabolism. The differentially added reactions show increased divergence between the two *Ostreococcus* networks, but a much larger number of reactions disappeared after gap-filling, illustrating the converging effect on the networks resulting from gap-filling.

Comparison to existing reconstructions

Due to the relatively recent discovery of *Ostreococcus*, little biochemical data are readily available. Consequently, the presented genome-scale reconstructions were exclusively based on genomic orthology. In comparison, for the established model green alga *Chlamydomonas*, at least three large-scale metabolic reconstructions exist (*55*, *111*, *112*). One of these models (*111*) is a detailed manual reconstruction that focuses on a comparison of predictions for heterotroph, autotroph, and mixotroph growth conditions. This reconstruction was not genome-wide (458 metabolites and 484 metabolic reactions), but it was compartmentalized and contained an extensive description of photosynthesis to investigate linear and cyclic electron transport. The first genome-scale model of *Chlamydomonas* was produced to introduce the bottom-up gap-filling algorithm (*55*). This model was uncompartmentalized, and compares most closely in scope and approach to the *Ostreococcus* models reconstructed in this paper. Recently, a second genome-wide *Chlamydomonas* reconstruction appeared (*112*) that includes cellular compartmentation. This work also addresses the role of light in algal metabolism and is the most sophisticated algal model to date. The network has roughly double the number of

reactions in the *Ostreococcus* models, with the *Chlamydomonas* model having 2019 reactions and 1069 metabolites.

Conclusions

The layered construction of the meta-plant reference database prevented incorporation of microbial reactions where possible. However, the bottom-up gap filing algorithm was unable to maintain the full biomass capability whilst gap-filling networks of reduced taxonomy (see Supplementary Tables S8 and S9 at JXB online). This limitation was also encountered during gap-filling of the *Ostreococcus* networks using the meta-plant network. By contrast, the trimming algorithm was able to retain all biomass functionality albeit at the cost of requiring more reactions.

The complete reference database was able to produce just over half of the target biomass metabolites. This biomass list is more extensive and varied than most models in the literature, but some common biomass components could not be produced (see Supplementary Table S9 at JXB online). This may reflect the limited list of balanced reactions contained within the KEGG database. Due to the limited size of this reference database, the *Ostreococcus* networks presented should not be regarded as definitive. The development of an exhaustive, open source and curated biochemical reaction list specific to plants should therefore be a priority in the development of plant-model reconstruction technology.

Although the quality of the reconstructed *Ostreococcus* networks is not on a par with carefully manually curated networks, the reconstruction process highlighted the ability to evaluate the completeness of the genome annotation for the *Ostreococcus* species. The

large number of reactions that needed to be added for the evaluated biomass components suggests that the *O. tauri* genome annotation in particular is lacking a substantial number of enzyme annotations. Comparison of the two metabolic network reconstructions suggested that *O. tauri* had been annotated more conservatively than *O. lucimarinus*. The difference between the two networks decreased somewhat after gap-filling, suggesting that the difference network of the two species was partly the consequence of the under annotation of *O. tauri*.

Bottom-up and top-down gap-filling approaches are both iterative methods resulting in many solutions. The ability to rapidly evaluate the quality of the gap-filling attempt is essential if many iterations are run or if the network contains many gaps. The *Ostreococcus* networks contained many such gaps, and the introduced measure for network quality provided a valuable tool to discriminate between the many gap-filling solutions automatically. The inclusion of the phylogenetic distance for reactions enriched the networks with reactions of closely related species, and thus the likelihood of these reactions existing within the actual metabolic networks. Recognition of realistic gap-filling solutions and this first network-wide functional comparison between the *Ostreococcus* species will help guide the comprehensive biochemical characterization of *Ostreococcus*.

Tables

Table 1

	Gap-filling	No of reactions	No. of metabolites	Producible biomass	Feasible (%)
KEGG orthology	None	3937	3582	49	
Eukaryote–cyanobacteria	None	2970	2826	46	
	Bottom-up	2974	2827	49	
Meta-plant	None	2060	2153	42	
	Bottom-up	2068	2154	48	
O. lucimarinus	None	908	1076	18	36.47
	Top-down	964	1100	48	48.32
O. tauri	None	801	971	0	25.79
	Top-down	871	1014	48	49.82

Table 1. Scale and functionality of networks.

The eukaryote–cyanobacteria model was refined using bottom-up gap-filling from the KEGG Orthology (KO) database. The meta-plant model was subsequently gap-filled with the gap-filled eukaryote–cyanobacteria model using the bottom-up method. Finally, the *Ostreococcus* networks were gap-filled with the gap-filled meta-plant model using the top-down approach. In all gap-filling instances, more reactions were added to the model than metabolites indicating an increase in overall network connectivity. All producible biomass functionality was transferred from the KO model to the eukaryote–cyanobacteria model. However, the meta-plant model was only able to produce 48 biomass components, one less than the KO and eukaryote–cyanobacteria models. Both *Ostreococcus* models were able to produce all 48 biomass components from the meta-plant model after gap-filling even though no biomass components were producible in *O. tauri* prior to gap-filling. Gap-filling increased the percentage of reactions with feasible fluxes in both *Ostreococcus* species, suggesting a substantial improvement in network

connectivity as a result of gap-filling. Feasible (%) = 100 * number of feasible fluxes/(number of feasible fluxes+number of non-feasible fluxes).

Figures

Figure 1

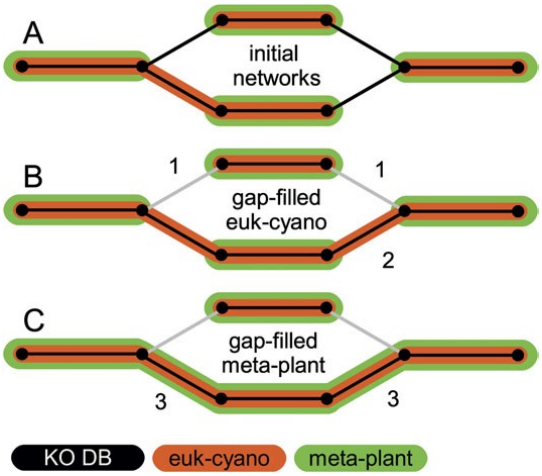


Figure 1. Stepwise database generation.

The meta-plant reaction reference database was procured using a nested gap-filling approach. The KO database (gene annotated reactions in KEGG) was evaluated for biomass production capability, representing the maximum feasible biomass that any model based on gene annotations from KEGG can achieve. Using the bottom-up gap-filling algorithm, this functionality can be added to smaller databases using a minimal set of added reactions. The set of all eukaryote and cyanobacteria annotations (A) was gap-filled using the KO database (B), and, in turn, the meta-plant network was gap- filled

using this gap-filled database (C). Reactions were therefore added with increasing priority from the database of closer phylogenetic proximity. KO reactions, which were not included in the euk- cyano model gap-filling (1) cannot be used to gap fill the meta-plant model. (2) Reaction added to gap-fill the euk-cyano model. (3) Reactions added to the meta-plant from the euk-cyano database.

Figure 2

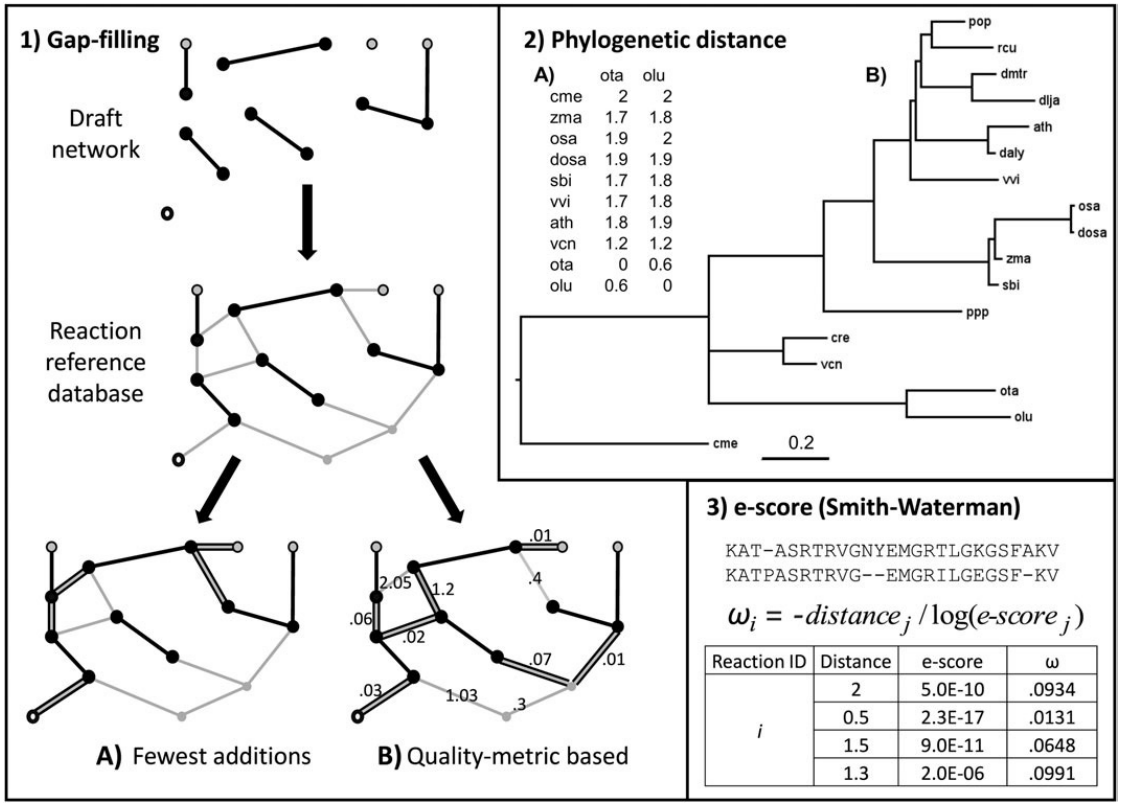
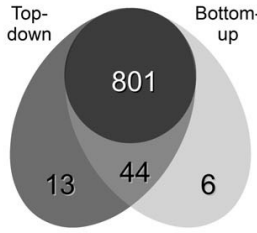


Figure 2. Network quality metric.

Draft networks can be gap-filled with the fewest number of added reactions (1A). Alternatively, reactions can be weighted for their likelihood to exist in a target organism (1B) by considering the phylogenetic proximity (2A) and sequence orthology (2B) of best matching sequences. Both factors were included in a quality metric (x) and associated with each reaction (3). This allowed the top-down gap-filling algorithm to preferentially include reactions with low (good) quality scores and was used to compare gap-filling results. All genes annotated with a particular reaction in the reaction database have a best match with a gene in both *O. tauri* and *O. lucimarinus*. A gap-filling solution with the

lowest sum of quality metrics for all reactions in a network was considered the best solution.

Figure 3



	O. tauri network	Before gap-filling	Bottom-up	Top-down: 36 biomass targets	Top-down: 48 biomass targets
	Reaction	801	851	858	871
	Metabolites	971	996	1000	1014
	Biomass	0	36	36	48
	Quality score	NA	6.204	6.044	7.252
	Max ω-value	NA	0.21	0.21	0.21
	Feasible (%)	25.79	48.03	48.93	49.82

Figure 3. Comparison of top-down and bottom-up gap-filling.

Fair comparison between top-down and bottom-up gap-filling requires the biomass targets to be identical. Because the bottom-up algorithm was only able to produce 36 biomass components, these same 36 components were used as a target for the top-down method. Top-down gap-filling added a total of 57 reactions to the draft *O. tauri* metabolic network of 801 reactions. The bottom-up algorithm added 50 reactions, 44 of which were also present in the top-down solution. Although the bottom-up algorithm included seven fewer reactions than the top-down algorithm, the combined quality score for the 50 added reactions was 6.204 whereas the top-down method scored 6.044 for 57 reactions.

Comparison of gap-filling results

Before gap-filling 25.8% of the reactions in the *O. tauri* network were capable of carrying flux (Feasible) using the available uptake metabolites while allowing all other metabolites to export. After bottom-up gap-filling 48% of the reactions were capable of carrying flux. After top-down gap filling 48.9% of reactions could carry flux. When the top-down approach was made to produce the entire meta-plant biomass target of 48 reactions, 70 reactions were added with a combined quality score of 7.252 and 49.8% of

the reactions in the network could carry flux. The network quality score is the sum of all quality scores of the reactions included in the network, and the maximum x-value indicates the worst reaction quality score of the included reactions.

Figure 4

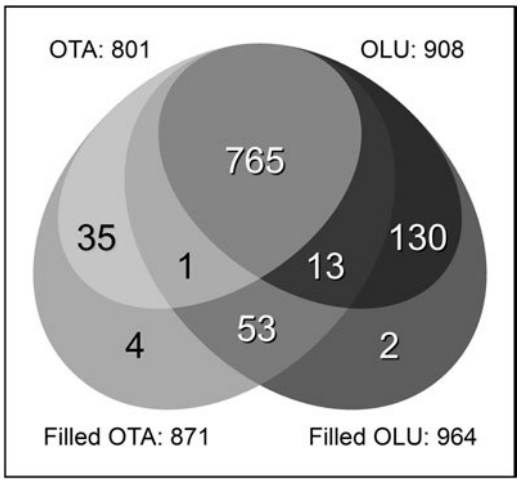


Figure 4. Reactions added by gap-filling to *O. tauri* and *O. lucimarinus*.

The draft networks of *O. tauri* and *O. lucimarinus* consisted of 801 and 908 reactions respectively, and contained an overlap of 765 reactions. The gap-filling process added a total of 70 reactions to *O. tauri* and 56 reactions to *O. lucimarinus*. *O. tauri* and *O. lucimarinus* shared 53 gap-filled reactions, which were present in neither draft network. *O. lucimarinus* donated 13 reactions to *O. tauri* during the gap-filling process, but *O. tauri* only donated one reaction to *O. lucimarinus*. Only six gap-filled reactions were unique to a single network, four were added to *O. tauri* and two were added to *O. lucimarinus*. Despite the large amount of shared reactions added during the gap-filling process, the networks retained many of their unique reactions (reactions present in only *O. tauri* or *O. lucimarinus*): the gap-filled *O. tauri* contained 39 unique reactions, and the gap-filled *O. lucimarinus* contained 132 unique reactions.

Figure 5

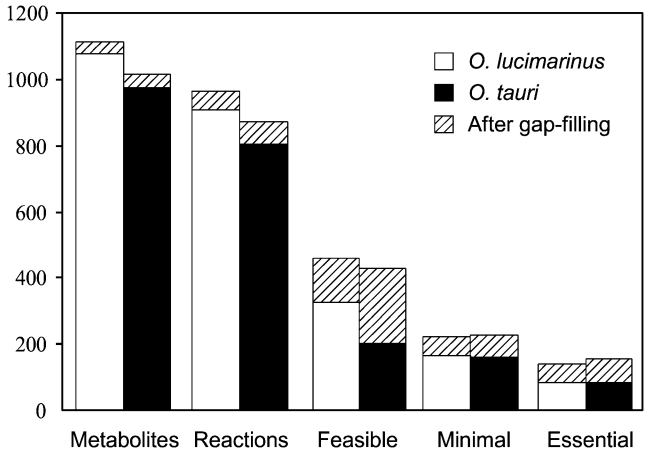


Figure 5. Comparison of *O. lucimarinus* and *O. tauri*.

To produce all 48 meta-plant biomass targets, the top-down gap-filling added 56 reactions and 34 metabolites to O. lucimarinus, and 70 reactions and 43 metabolites to *O. tauri*. Fewer metabolites than reactions were added in both cases, indicating that the network connectivity had improved for both *O. lucimarinus* and *O. tauri*. Upon gap-filling, an additional 133 reactions in *O. lucimarinus* and 224 reactions in *O. tauri* were able to carry flux (labelled 'Feasible'). 153 reactions where required for the minimal geometric FBA solution in *O. tauri* compared with 138 reactions in *O. lucimarinus* (labelled 'Minimal'). The numbers of essential reactions, as determined by reaction knockouts, in both *Ostreococcus* networks were similar: 82 in *O. lucimarinus* and 83 in *O. tauri*. 14 more essential reactions were added to *O. tauri* after gap-filling (labelled 'Essential').

Figure 6

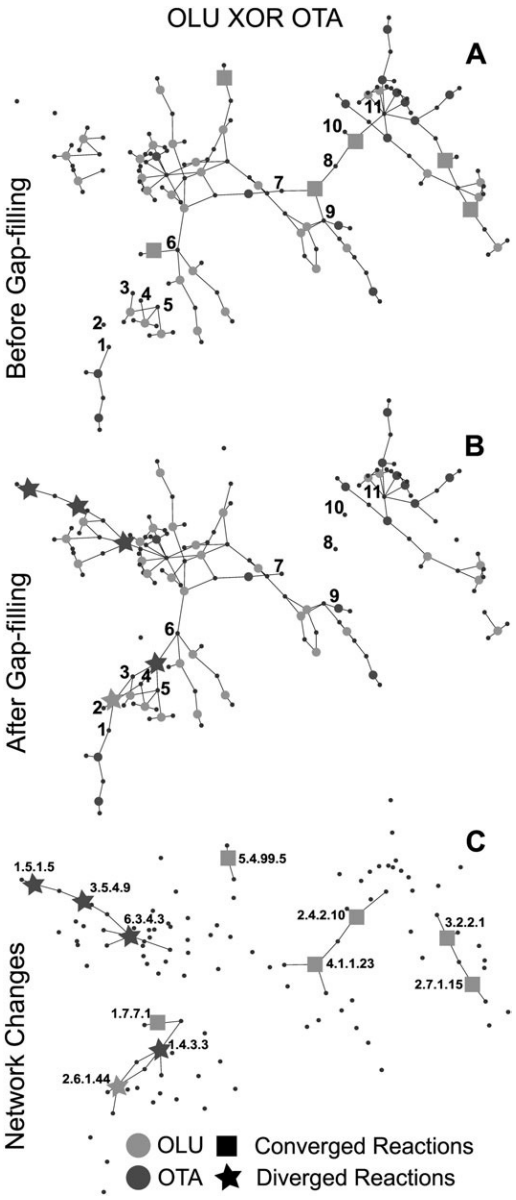


Figure 6. *O. tauri* compared with *O. lucimarinus* before and after gap-filling.

(A, B) Reactions present only in *O. tauri* or *O. lucimarinus* before and after gap-filling following binary XOR logic. Reactions present only in *O. tauri* are shown as light-grey nodes and reactions present only in *O. lucimarinus* are shown as dark-grey nodes,

metabolites are represented by small black nodes. (C) The third network shows the differences between the *O. tauri* XOR *O. lucimarinus* before and after gap-filling, including the EC numbers of the selected reactions. The networks show the largest connected component of the XOR graphs in the union of the before and after conditions, and are thus a subset of the total XOR networks between *O. tauri* and *O. lucimarinus*. Reactions represented with squares were removed from the *O. tauri* XOR *O. lucimarinus* network during the gap-filling process by adding the corresponding reaction to the other species' metabolic network. These reactions represent functionality that converged as a result of gap-filling. Stars indicate new additions to the *O. tauri* XOR *O. lucimarinus* network as a result of gap-filling. Star reactions were required by only one organism during the gap-filling process and represent diverged functionality. The only star reaction for *O. lucimarinus* shown here is EC 2.6.1.44 alanine–glyoxylate transaminase. This reaction converts glyoxylate (3) and L-alanine (2) into pyruvate (1) and glycine (4) and is not present in *O. tauri*. *O. tauri* also diverged with four unique reactions, one of which (1.4.3.3, D-amino-acid oxidase) involved the interconversion of glyoxylate (3), hydrogen peroxide (5), and ammonia (6) to glycine (4), O2 (not shown), and H2O (not shown). The converged reactions 2.4.2.10 and 4.1.1.23 demonstrate the gap-filling of a missing reaction in *O. tauri* by incorporating reactions from *O. lucimarinus*. Reaction 2.4.2.10 converts ortidine 5'-phosphate (8) and diphosphate (not shown) into orotate (10) and 5- phospho-alpha-D-ribose 1-diphosphate (11). Reaction 4.1.1.23 converts orotidine 5'-phosphate (8), into uridine monophosphate (UMP) (7) and CO2 (9). EC number key: 1.5.1.5 methylenetetrahy- drofolate dehydrogenase; 3.5.4.9

methenyltetrahydrofolate cyclohydrolase; 6.3.4.3 formate-tetrahydrofolate ligase; 2.6.1.44 alanine–glyoxylate transaminase; 1.7.7.1 ferredoxin-nitrite reductase; 1.4.3.3 D-amino-acid oxidase; 5.4.99.5 chorismate mutase; 2.4.2.10 orotate phosphoribosyltransferase; 4.1.1.23 orotidine-5'-phosphate decarboxylase; 3.2.2.1 purine nucleosidase; 2.7.1.15 ribokinase.

CHAPTER 3: Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks

Synopsis

Genome-scale metabolic models are central in connecting genotypes to metabolic phenotypes. However, even for well-studied organisms such as *Escherichia coli*, draft networks do not contain a complete biochemical network. Missing reactions are referred to as gaps. These gaps need to be filled to enable functional analysis, and gap-filling choices influence model predictions. To investigate if functional networks existed where all gap-filling reactions were supported by sequence similarity to annotated enzymes, four draft networks were supplemented with all reactions from the Model SEED database for which minimal sequence similarity was found in their genomes.

Quadratic programming revealed that the number of reactions that could partake in a gap-filling solution was vast: 3,270 in the case of *E. coli,* where 72% of the metabolites in the draft network could connect a gap-filling solution. Nonetheless, no network could be completed without the inclusion of orphaned enzymes, suggesting that parts of the biochemistry integral to biomass precursor formation are uncharacterized. But, many gap-filling reactions were well-determined, and the resulting networks showed improved prediction of gene essentiality compared to networks generated through canonical gap-filling. In addition, gene-essentiality predictions that were sensitive to poorly determined

gap-filling reactions were of poor quality, suggesting that damage to the network structure resulting from the inclusion of erroneous gap-filling reactions may be predictable.

Supplementary material is available at Journal of Biological Chemistry online, URL: http://www.jbc.org/ DOI: 10.1074/jbc.M114.634121

Introduction

Metabolic network reconstructions are instrumental in aggregating metabolic knowledge about organisms (1–3). Network reconstructions have steadily grown in size, reflecting increasingly comprehensive genome annotations (4–7). In addition, reconstructions have grown in complexity. Current reconstructions contain detailed gene-to-protein-to-reaction (GPR) mappings, thermodynamic constraints, and in some cases, signal transduction layers (8–10). The most sophisticated reconstructions have been extensively curated (6, 11, 12), but draft reconstructions are now mostly machine-generated or machine-assisted (7, 13–16). The Model SEED uses annotations from the "Rapid Annotation using Subsystems Technology" (RAST) web service (17, 18) as part of a network reconstruction pipeline for prokaryotes (15). In addition to a starting point for curated reconstructions, draft metabolic networks facilitate interpretation of the metabolic capabilities of newly sequenced organisms or communities of organisms (7, 19, 20).

Metabolic networks are reconstructed in a bottom-up fashion from identified genes following genome annotation (21). Knowledge of metabolic pathways can guide gene annotation, as implemented by the Pathway Tools software (22). Similarly, RAST

simultaneously annotates genes that are part of a metabolic subsystem (18), utilizing mutually corroborating information on genes involved in closely related metabolic processes. As a logical extension of the subsystem approach, network-wide mutually corroborating information may be used to guide reconstructions. An application of this concept is to require draft networks to be able to carry out the production of all essential cellular building blocks, collectively referred to as biomass, from a well-defined media source (23).

Metabolic networks resulting from assembling all reactions inferred from gene annotations (draft networks) are currently unable to describe the synthesis of all biomass components. Draft networks contain gaps, isolated reactions, and reactions that cannot carry flux under any circumstances (1). Although isolated or blocked reactions are easily identified (24), it is not obvious whether they result from under-annotation or over-annotation. Hence, an isolated reaction may need to be connected through additional reactions that were under-annotated in the draft network, or the isolated reaction resulted from a spurious annotation. Gaps in the network pose the opposite problem: although a network can be readily completed to enable production of all biomass components (25), the location of the actual missing reaction may be illusive. The appearance of gaps in metabolic networks is not exclusively the result of under-annotation. Incorrect reaction reversibility assignments (thermodynamic constraints) (26), or stoichiometric constraints resulting from dead-end metabolites, may also prevent production of biomass components (25, 27). Lastly, part of the biochemistry of an organism may not have been associated with genes, or the biochemistry may yet to be discovered. Adding reactions to fill these gaps is known as

"gap-filling" and has been the subject of considerable inquiry and has been reviewed in detail elsewhere (24).

Commonly, mixed integer linear programming (MILP) optimization is used to perform bottom-up gap-filling (27). In this approach, reactions are iteratively added until production of biomass becomes feasible, often while minimizing the number of reactions required (15, 25, 27). Several other optimization strategies have been reviewed here (28). In the case of the Model SEED, reactions are prioritized based on their nature. For instance, adding an internal reaction incurs a lower cost than adding a transporter. Bottom-up gap-filling works well for well-annotated genomes, but for networks that require extensive gap-filling a top-down approach is more robust (29). In the more recently developed top-down methods, all gap-filling reactions are added, followed by the successive preferential removal of unneeded gap-filling reactions with little or no sequence similarity in the genome of the organism for which the network is reconstructed (14, 29, 30). Prioritization of the removal of reactions without sequence similarity minimizes the inclusion of locally (enzymes with an associated sequence that is not present in the target genome), and globally (reactions without sequence association) orphaned reactions. Very recently, a bottom-up MILP approach also used sequence similarity as a likelihood metric for the existence of a gap-filling reaction in the target genome (31). Gap analysis itself has been used to identify knowledge gaps in human metabolism (32) and to leverage contextual information of networks to hypothesize gene function (33).

This work investigated the need for adding gap-filling reactions to draft networks, the extent for which sequence similarity to enzymes can be found for these reactions, and how

orphaned enzymes influence gene essentiality predictions by metabolic networks. To assess the extent that sequence similarity to known enzymes can support the choice of gap-filling solutions, new linear programming (LP) and quadratic programming (QP) based gap-filling problems were formulated that minimize the utilization of unsupported reactions. All gene sequences associated with the Model SEED gap-filling reaction database (11,858 reactions, received on April 20, 2012) (15) were queried against four prokaryotic genomes and unique gap-filling solutions were retrieved that minimized the utilization of unsupported reactions. Unlike recently reported BLAST-weighted MILP-based work (31), the networks resulting from our approach outperformed networks gap-filled by the Model SEED (15).

Experimental Procedures

Metabolic networks, biochemistry database, and gene annotations

Metabolic networks for *Streptococcus pneumoniae*, *Bacillus subtilis*, *Escherichia coli* MG1655, and *Acinetobacter baylyi* ADP1 were downloaded from the Model SEED (http://seed-viewer.theseed.org/) on May 3, 2013 (15) along with media conditions and biomass formulations. The Model SEED gap-filling biochemistry database, experimental gene-essentiality results, and associated media formulations were kindly provided by Chris Henry (Argonne National Laboratory, IL). Gene annotations for 891 prokaryotic species were downloaded from the RAST sapling server (34), totaling 690,445 genes encoding 7,218 functional roles. The biochemistry database maps genes to reactions through the use of functional roles and enzyme complexes made up of functional roles

(15). Table 2 includes a summary of the size of the downloaded Model SEED database and draft metabolic networks.

Identification of functional roles

For each functional role in the biochemistry database, a BLAST amino acid database was generated using all protein sequences associated with that particular role. The complete genomes of target organisms were queried against each functional role BLAST-database using BLASTX (35) with the BLOSUM62 scoring matrix (36). The E-values for the best BLAST high-scoring segment pairs (HSPs) from each functional role database query were used to weight biochemical reactions. E-values were chosen because they are comparable between different calls against distinct functional role databases and they correct for multiple comparisons by penalizing the score by both the length of the enzyme database and the length of the target genome (37). Only the lowest E-value was recorded. To adjust the weights for each enzyme complex independently, duplicate reactions were created so that each complex had an independent mapping with a reaction. Reactions were weighted with the geometric mean of the E-values for the constituent roles of an enzyme complex. This treats the E-values as probabilities in determining the support for the existence of an enzyme, which is here defined as enzyme sequence support (ESS). Reactions with an ESS of less than 1.0E-240 were set to the value of 1.0E-240. Reactions were weighted by the logarithm of the ESS values of the associated enzymes:

$$W_R = \ln(E_R) - \ln(E_{\min}) \qquad (1)$$

where W_R is the weight for a reaction, E_R is the ESS for a reaction, and E_{min} is the minimum E-value. This formulation results in small weights for well-supported reactions

relative to unsupported reactions, while constraining the weights to a smaller numerical range, which improved the numerical stability of the LP and QP solver software.

Gene essentiality and metabolite production

Flux balance analysis (FBA) (38) was used to check for the existence of a synthesis route for individual biomass components. A gene was classified as computationally essential if removing the reaction(s) uniquely associated with an enzyme complex resulted in a network that could not carry flux greater than 1.0e-6 to biomass. Similarly, an individual metabolite was classified as producible if a flux solution could be found that carried a flux >1.0e-6 of the tested metabolite through an export reaction that was temporarily added for testing purposes (25).

Gap-filling algorithm

The BLAST-weighted LP gap-filling algorithm was formulated as:

$$\min_{\mathbf{v}} f(\mathbf{v}) = \mathbf{w}^{\mathrm{T}}\mathbf{v} \qquad (2)$$

$$\text{Such that:} \quad \begin{aligned} &\mathbf{Sv} = \mathbf{0} \\ &\mathbf{0} \le \mathbf{v} \le \mathbf{v}_{\mathbf{max}} \\ &v_{bio} = 1e-3 \end{aligned} \qquad (3)$$

where **w** is a column vector of weights (Eq. 1) and **v** is a column vector of reaction fluxes including separate terms for forward and reverse reactions. The stoichiometric matrix (**S**) relates reactions to metabolites through stoichiometric coefficients. A negative value in the stoichiometric matrix specifies a metabolite that is consumed by a reaction, and a

positive value describes the production of a metabolite by a reaction. **S** has dimensions m (metabolites) by 2n (n reactions, 2n for both directions). The constraint **Sv=0** enforces that all metabolites have a net balance of production and consumption, known as a mass balance constraint. $\mathbf{v_{max}}$ is a vector of upper bounds on reaction fluxes, v_{bio} is the required flux through the biomass reaction. Similarly, a weighted QP gap-filling algorithm was formulated as:

$$\min_{\mathbf{v}} f(\mathbf{v}) = \mathbf{v}^{\mathbf{T}}\mathbf{W}\mathbf{v} \qquad (4)$$

$$\text{Such that:} \quad \begin{aligned} &\mathbf{Sv} = \mathbf{0} \\ &\mathbf{v_{min}} \le \mathbf{v} \le \mathbf{v_{max}} \\ &v_{bio} = 1e-3 \end{aligned} \qquad (5)$$

where **W** is a diagonal matrix of the weights. Other terms were identical to the LP formulation, except that reactions did not need to be divided into forward and reverse directions and were constrained directly using two vectors: $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$. The QP formulation results in fluxes that minimize the sum of weighted squared fluxes, which effectively distributes fluxes across available biomass routes inversely proportional to the weights and number of reactions of a given route.

Software

LP and QP problems were solved with CPLEX™ (IBM, Armonk, New York, http://www.ibm.com/). LP problems were solved using the dual simplex solver to minimize constraint violations. Custom Matlab™ (Mathworks, Natick, Massachusetts, http://www.mathworks.com/) and Python (www.python.org) scripts were used for the

preparation of matrices and databases. Producibility of metabolites from media components was tested by FBA using the COBRA Toolbox (39).

BLASTX comparisons for the four target genomes were run on a commodity quad core Intel i3 desktop computer, taking roughly eight total hours to complete. Further processing of the BLASTX output using custom Python scripts required approximately four hours of computation time.

Results

Metabolic networks require gap-filling

Draft metabolic networks for *S. pneumoniae*, *B. subtilis*, *E. coli* MG1655, and *A. baylyi* ADP1 were downloaded from the Model SEED on May 3, 2013. Corresponding experimental gene essentiality results of genome-scale single gene knockout libraries (40–43), along with mappings to the model genes were provided by the Model SEED upon request in October 2011. The four organisms were selected because the associated gene knockout libraries were generated through full-length gene deletion methods rather than transposon insertion methods. Transposon knockouts can display complex gene knockdown behavior which complicates the interpretation of gene essentiality predictions (44). The gap-filling reactions added by the Model SEED were stripped from the downloaded models. In addition, all genes not directly associated with metabolic reactions were not evaluated to limit gene essentiality evaluations to precursor metabolism only.

The draft networks now contained gaps resulting from under-annotation of the genome, incorrect reaction reversibility constraints resulting from inaccurate Gibbs free energy estimates, and stoichiometric constraints caused by dead-end metabolites. Consequently, there were three approaches to gap-fill metabolic networks by addressing each of the three causes. To test if removal of thermodynamic constraints alone could enable biomass production, all reactions were made reversible. The existence of a route to biomass was tested using FBA by maximizing flux through the biomass reaction. No such route existed for any of the four networks, demonstrating that removal of thermodynamic constraints alone was insufficient to gap-fill the tested networks. Removal of stoichiometric constraints caused by dead-end metabolites by allowing all metabolites to leave the network was also insufficient. Furthermore, a combination of relaxing thermodynamic constraints and allowing metabolites to leave the network (all reactions in the metabolic network were made reversible and all metabolites could leave the network) also did not result in feasible biomass production in the networks. Hence, addition of reactions to all tested metabolic networks was necessary.

Network completion requires reactions with no enzyme sequence support (ESS)

For all further gap-filling approaches, the reactions from the Model SEED gap-filling database (11,858 reactions, Table 2) were used as candidates for gap-filling (Fig. 7). The database included a subset of curated transport reactions, and had been thermodynamically constrained using the group contribution method (45). The sequences of the RAST annotated genes of all organisms in the Model SEED database were

extracted, and a sequence database was generated for each gene. Using the RAST mapping between genes, functional roles, enzyme complexes, and reactions, an organism-specific weight (Eq. 1) was calculated for each reaction of the gap-filling database (Fig. 7, see Experimental Procedures for details).

To test if networks could be completed by restricting incorporation of reactions with a predefined level of support, reactions were divided into three tiers: highly-supported reactions (ESS of 1.0e-240), significantly supported reactions (30) (ESS≤1.0e-10), and unsupported reactions (ESS>1.0e-10). For each tier, all reactions were added to the base models, and FBA was used to evaluate if biomass could be produced. This revealed that no networks could be completed with only highly-supported reactions, or even significantly supported reactions (Table 3). This suggested that the tested networks required locally orphaned enzymes (no similarity to known enzymes in the organism), or globally orphaned enzymes (no known sequence) to produce all biomass components. Hence, orphaned metabolic functionality was integral to the core of metabolic networks and included reactions essential to biomass production. However, after releasing all thermodynamic constraints including those in the gap-filling reactions and stoichiometric constraints caused by dead-end metabolites, networks containing only significantly supported reactions were able to produce all biomass components. The role of thermodynamic constraints on network completion was investigated in detail and will be reported in a specialist journal.

Non-producible biomass metabolites are distributed across metabolism

The four tested organisms were unable to produce a significant portion of their biomass components, in each case spanning multiple classes of metabolites (Figs. 8 and 9). Some of the biomass components in *S. pneumoniae, B. subtilis,* and *A. baylyi* that were not producible in the base models, were producible with the models that were augmented with the strongly supported reactions only. This suggested that the original networks were under-annotated. This was especially true for *S. pneumoniae,* which could not produce half of its biomass metabolites (39 out of 79), yet 33 metabolites could be produced solely using significantly supported reactions (Fig. 9, Table 3). Only six to eight biomass metabolites could not be produced with only supported reactions in each organism. The ability of the augmented networks to produce often many more biomass components than the base models, even if only the highly supported reactions were used, suggests that there was sufficient potential for ESS values to guide gap-filling solutions. Two biomass components, acyl carrier protein (ACP) and peptidoglycan polymers, could not be produced by any of the organisms. Spermidine and thiamine pyrophosphate (TPP) could also not be produced by any organism, but were imported from the media by *S. pneumoniae* (Fig. 8). The gram-positive bacteria *S. pneumoniae* and *B. subtilis* required the cell wall precursor glycerol teichoic acid (GTA). ACP, peptidoglycan polymers, calomide, and GTA could be produced in isolation with significantly supported reactions if the biomass reaction was replaced by independent export reactions for all biomass components. However, ACP, peptidoglycan polymers, and calomide biosynthesis was not required for total biomass production in the models because their precursors were

regenerated by the biomass equation itself. All other non-producible metabolites are discussed in detail below.

Note that not all biomass components were necessarily essential, for instance, spermidine is part of the canonical *E. coli* biomass equation, but may not be essential (5, 46). In some cases genetic evidence may support the classification of a metabolite as essential if a sole pathway synthesizes the metabolite, riboflavin is one such example. In the biosynthesis of riboflavin in *E. coli*, all genes associated with riboflavin biosynthesis were experimentally essential (40), suggesting that riboflavin is indeed an essential biomass metabolite. It was surprising that the complete synthesis of riboflavin required unsupported reactions, despite the final steps in riboflavin synthesis being present in the metabolic network (see below).

Blast-weighted gap-filling

A weighted LP problem was formulated to incorporate reactions into gap-filling solutions depending on their ESS. Each reaction in the gap-filling database was weighted inversely proportional to the associated ESS scores (Experimental Procedures). The improbability (approximated by 1-E-value) of a sequence similarity score occurring by chance was treated as the level of support for an enzyme activity existing in the network. This simplification was vulnerable to detecting false positives caused by strong similarity to a short sequence or domain only, but assessment at network level immunizes this approach to most effects of false positives. Hence, for an incorrect pathway to be included, all reactions in a pathway would have to be false positives. Treating support for a reaction as a probability, the support for a pathway was expected to scale with the product of the

support of the underlying ESS values. Support for the existence of a pathway of *n* reactions may then be described as the product of the ESS values for the individual reactions: $\prod_{i=1}^{n} ESS_i$. To avoid penalties against existing reaction annotations, reactions already included in the draft network received a weight of zero. The linear and quadratic programming objectives were minimized, while requiring a set flux through the biomass reaction (Experimental Procedures). Utilization of LP and QP made gap-filling very fast. On a typical desktop computer, solutions were retrieved in seconds, compared to minutes for MILP.

Quadratic programming reveals the gap-filling solution space

The QP formulation of the weighted gap-filling algorithm minimized the squared sum of weighted reaction fluxes. Squaring the weighted reaction fluxes limited large fluxes because penalties increased quadratically with flux. Conversely, small fluxes were penalized lightly, even if the associated weights are high. This resulted in the distribution of flux through alternative gap-filling solutions inversely proportional to the combined weights of reactions in a given pathway (Fig. 10). The number of reactions used in a QP gap-filling solution can thus provide a lower bound estimate on the number of reactions that can participate in gap-filling solutions. QP revealed that several thousand reactions could participate in gap-filling reactions for each organism (Fig. 11). Importantly, QP is not guaranteed to identify all potential gap-filling routes. Combinations of irreversible reactions and reaction weights can lead to hidden gap-filling reactions (Fig. 10). Removing reversibility constraints and adding random reaction weights allowed for an extreme estimate of gap-filling solutions in the gap-filling database. It was revealed that

even more reactions could potentially participate in gap-filling. For *E. coli*, this extreme QP solution included 7,337 reactions, almost double the number of reactions in the constrained QP solution.

The LP solutions were necessarily always contained in a subset of the QP solution for a given set of reversibility constraints and reaction weights. LP solutions that were based on uniform weights were mostly, but not always, contained in the QP solution (Figs. 10 and 11). The solutions of the Model SEED were more frequently outside the QP solution space (Fig. 11). The uniformly weighted LP solutions contained the lowest number of gap-filling reactions for the four tested networks, and were likely the minimal reaction solutions in most cases. Strictly, the uniformly weighted LP solution was a minimal flux solution, making it imaginable that an alternative LP solution with fewer reactions, but that carries more flux, may exist. In contrast, the weighted LP solutions often contained high flux reactions, but only if such reactions were associated with very low weights (Fig. 9). The weighted LP solution always contained substantially more reactions than either the uniformly weighted LP or the Model SEED gap-filling solutions, suggesting that a strong enough ESS signal existed to significantly influence gap-filling. LP and Model SEED gap-filling solutions often shared several reactions, indicating that *some* biomass components may only be made producible in a limited number of ways, but no solutions were close to identical (Figs. 11 and 12).

The sheer size of the QP solution made it clear that many different gap-filling solutions existed (Fig. 12). 78% of the metabolites of the original network were used in the quadratic gap-filling solution of *S. pneumoniae* (Fig. 12), suggesting that there were no

obvious metabolites that should serve as connecting metabolites to gap-filling reactions. With the size and level of connections in mind, it was surprising that gap-filling solutions shared reactions at all. Further investigation revealed that the shared reactions were often associated with high ESS values, and sometimes represented a sole gap-filling solution to a subset of biomass components (Table 3, S1, S2, available at JBC online). The high ESS values of the shared reactions indicated that, in reality, alternative pathways to these biomass components may exist, as well as highlighted the possible existence of missing biochemistry in the gap-filling database.

Comparison of computational and experimental gene essentiality

To investigate the quality of the gap-filling solutions, gene essentiality predictions from the gap-filled networks were compared to experimental data of full-length single gene knockout libraries. Gene deletions were simulated by removing all reactions that required a given gene. Gene essentiality was then predicted from the feasibility of biomass production from the specified media using FBA. Gene deletions that resulted in networks that could no longer produce biomass were considered computationally essential. Networks that were gap-filled with weighted gap-filling (referred to as BLAST LP) predicted gene knockout outcomes better than networks filled by the Model SEED or uniformly-weighted gap-filling (Table 4). Weighted gap-filling outperformed the alternatives methods both at the essential as well as nonessential gene predictions. The improved performance for both essential gene and nonessential gene predictions was striking because the weighted gap-filling added significantly more reactions to networks, yet this did not result in fewer true essential gene predictions (except for in *B. subtilis*).

More importantly, it suggested that the ESS signal was strong enough to enhance gap-filling of draft metabolic networks, even though all the solutions included reactions associated with maximum ESS values. All genes associated with supported reactions in the BLAST LP gap-filling solutions for the *E. coli* network were consistent with RAST annotations. However, the additional functionalities associated with *AceE* and *SucB* (S1, available at Biophysical Journal online) were not supported by the literature (47) and were likely incorrect.

A subset of knockout predictions are sensitive to weight changes

Two sensitivity analyses were performed to investigate the robustness of the network support (NS) for the weighted gap-filling solutions, and the influence of the gap-filling solutions on gene essentiality predictions. NS is here defined as how well a gap-filling reaction selection is determined by the entire network. NS for a reaction is calculated from the gap-filling penalty function increase after exclusion of that reaction from the gap-filling database. To test which gene essentiality predictions were sensitive to any gap-filling solution, 100 Monte Carlo gap-filling simulations of the *E. coli* network with randomly shuffled weights were calculated. This resulted in 97 genes with alternating essentiality prediction (S2, available at JBC online), suggesting that a significant portion (9.1%) of gene predictions were sensitive to gap-filling.

In a second sensitivity analysis, weights were shifted by a small, random amount from the sequence-derived weights to test sensitivity to variations in the ESS calculation. Only 18 genes changed predicted essentiality status over 100 runs (S2, available at JBC online). This suggested that the sequence-based gap-filling approach was fairly robust to

variations in the BLAST sequence comparisons. The number of genes changing essentiality prediction was fairly insensitive to the magnitude of the added noise, ranging from a normal distribution centered at zero, with a standard deviation from 0.1 to 10 units (reaction weights scaled between 0 and 553), indicating that most essentiality predictions were well determined by the gap-filling approach.

The BLAST LP optimal gap-filling solution utilized 15 reactions, eight of which were present within all shifted weight Monte Carlo gap-filling solutions. The additional reactions varied substantially over the Monte Carlo runs, but the number of reactions that were featured at least once in the gap-filling solutions was insensitive to the magnitude of the noise (S2, available at JBC online). Of the eight reactions that were always retrieved, four had minimal ESS values and four had maximum ESS values, including one mandatory reaction for which no alternative existed. Removal of any of the eight reactions that were always included resulted in solutions with substantially higher objectives, indicating strong NS and explaining their consistent inclusion. 17 of the 18 genes with variable essentiality predictions were essential in the noiseless solution. Remarkably, 15 out of these 17 computationally essential predictions were wrong, which was on par with random predictions, considering that only 10% of genes were experimentally essential. Note that overall, 70% of the experimentally essential genes were correctly predicted as essential. However, of the computationally essential genes, only 44.6% were experimentally essential. Disregarding the 18 genes with alternating essentiality calls improved the latter statistic to 48.3%.

Combined, these results indicated that the ESS signal was strong enough to determine >80% of otherwise variable essentiality predictions. The seven gap-filling reactions for which the inclusion was sensitive to reaction weights, determined the 18 gene essentiality calls that were of very poor quality (S2, available at JBC online). Therefore, the implied presence of orphaned enzymes in all networks did not nullify the ability to find meaningful gap-filling solutions, but the poorly determined reactions significantly deteriorated a subset of essentiality calls.

Analysis of gap-filling reactions with high ESS values

A subset of the metabolites that could not be produced by significantly supported reactions still required unsupported reactions for production after breaking the biomass equation into independent export reactions. These unsupported metabolites were investigated in more detail by using the BLAST LP algorithm on the gap-filled networks to calculate flux utilization of the gap-filling reactions for unsupported metabolites (Fig. 13).

The two reactions required by *E. coli* for riboflavin, FAD, and TPP synthesis had strong NS values and were therefore always included in shifted weight sensitivity analysis. The remaining reaction associated with spermidine synthesis was included in 71 out of 100 solutions. This suggested that these reactions were strongly determined by the gap-filling approach. In contrast, only the reaction associated with riboflavin and FAD was always included in the shuffled weight sensitivity analysis because no alternative reaction was present in the gap-filling database. The reactions for TPP and spermidine synthesis were only included in four and 24 cases out of 100 in the shuffled weight solutions

respectively. This suggests that despite the lack of ESS support for these reactions, NS strongly determined gap-filling reactions among many other poor alternatives.

Two reactions implicated by NS were supported by circumstantial evidence. The gap-filling export reaction for TPP may be through spontaneous diffusion due to the chemical properties of 4-hydroxy-benzyl alcohol, a byproduct of TPP synthesis. Riboflavin and FAD required a reaction for which no alternative existed in the Model SEED database. This reaction has been hypothesized in the literature and only recently a gene in *E. coli* has been associated with the activity (48).

Discussion

Draft metabolic networks of four species were investigated for the ability to produce a complete set of biomass metabolites. The observed inability of networks to produce biomass, even after removal of thermodynamic and stoichiometric constraints caused by dead-end metabolites, necessitated the addition of gap-filling reactions. Although each network could be readily filled using the Model SEED biochemistry database, no networks could be filled solely with reactions that were supported by sequence similarity to known enzymes. The need for orphaned enzymes implied that all metabolic networks were missing essential biochemistry annotations. Possibly, these reactions are of unknown biochemistry, suggesting fundamental gaps in our biochemistry knowledge for even the best-studied organisms. This realization suggests that our biochemistry knowledge or inclusion of this knowledge in the database, rather than the quality of machine annotations, is limiting our ability to further improve automated network reconstructions. Note that given the very small flux requirement through unsupported

reactions (Fig. 9), it is conceivable that some of the orphaned activities may be attributed to secondary catalytic activity of promiscuous enzymes.

The presence of orphaned enzymes in gap-filling solutions, and the very large size of the solution spaces, made evident by the quadratic programming, prompted the question of how robust the gap-filling solutions were in response to noise, and to what extent gene essentiality predictions were influenced by gap-filling solutions. One hundred repeated gap-filling runs using randomly shuffled weights for the *E. coli* network showed that a substantial number of reactions could be part of the gap-filling solution, which resulted in many alternate gene essentiality assignments (S2, available at JBC online). However, in response to noise added to the correct weights, a much smaller subset of genes showed alternating gene essentiality. This suggested that many gene essentiality predictions sensitive to the gap-filling solutions were strongly determined by the sequence-derived weights. Additionally, eight gap-filling reactions were always present in gap-filling solutions, suggesting that they were strongly determined by NS. Interestingly, the essentiality of the group of genes sensitive to gap-filling was predicted very poorly, which suggested that the fallout of the partially arbitrary gap-filling process due to a simplified relationship between E-values and ESS, as well as the addition of orphaned enzymes, may be limited to a small subset of gene essentiality predictions.

LP- and QP-based gap-filling algorithms generated fast and meaningful gap-filling solutions. LP optimization resulted in gap-filled networks that performed superior in gene essentiality predictions in comparison to networks that were filled with existing gap-filling technology. The large majority of gap-filling reactions was supported by sequence

similarity and had often been identified by RAST, yet these reactions had not been included in the Model SEED draft models. The fairly insignificant computational time to establish ESS values (2 hours per organism on a quad core Intel i3 desktop computer) should be well worth the effort even though the network quality improvement may be modest. This is particularly true for the inclusion of BLAST LP in network reconstruction pipelines.

This work demonstrated that orphaned enzymes were integral to essential metabolic functions, and that a fully supported and functionally complete metabolic network could not be assembled even with the extensive compilation of enzymes and biochemistry from RAST and the Model SEED. Nonetheless, sequence similarity driven gap-filling improved the quality of the networks and identified deficiencies in our biochemistry knowledge. The large set of significantly supported gap-filling reactions in all gap-filling solutions showed the potential for network-based identification of candidate gene annotations. Truly realistic models will likely require further expansion of the Model SEED biochemistry database, or the discovery of not yet observed metabolic reactions and their gene associations.

Tables

Table 2

	Model SEED Gap-Filling	*Streptococcus pneumoniae*	*Bacillus subtilis*	*Escherichia coli* K-12 MG1655	*Acinetobacter baylyi* ADP1
Number of functional roles	7,218	1,496	2,606	3,658	2,200
Number of unique reactions	10,516	880	1,537	1,638	1,287
Number of metabolites	7,732	848	1,280	1,278	1,095
Number of enzyme complexes (equal to number of reactions	11,858	NA	NA	NA	NA
Number of genes	690,445	480	952	1067	701

Table 2. Model SEED database and model summary.

Metabolic networks produced by the Model SEED are subsets of the complete Model SEED gap-filling database. Relationships from gene to functional role to enzyme complex to reaction are encoded as a gene to reaction relationship in draft metabolic networks, thus removing the enzyme complex abstraction from the model. This compact encoding of relationships allows gene knockouts to be quickly translated into reaction knockouts in draft networks.

Table 3

Biomass metabolites producible with:	*Streptococcus pneumoniae*	*Bacillus subtilis*	*Escherichia coli* K-12 MG1655	*Acinetobacter baylyi* ADP1
Base model	39	57	60	47
Highly-supported reactions	41	74	60	51
Significantly supported reactions (FSS ≥ 1e-10)	71	76	66	61
All reactions in gap-filling database	79	83	73	67

Table 3. Biomass components that require gap-filling.

To investigate which biomass metabolites required gap-filling, FBA was used to maximize the export of each individual biomass component, given the stoichiometric and thermodynamic constraints imposed on the network. An exchange reaction was added for each biomass component that was tested, and FBA was used to maximize flux through each component exchange reaction in turn. Metabolites that could not be exported at a flux greater than a numerical cutoff of 1.0e-6 were considered non-producible. Production of the individual biomass components was attempted using gap-filling reaction sets with three different levels of support, as well as the base models with no gap-filling reactions.

Table 4

Essentiality	Gap-filling method	*E. coli*		*B. subtilis*		*A. baylyi*		*S. pneumoniae*	
		EE	**ENE**	**EE**	**ENE**	**EE**	**ENE**	**EE**	**ENE**
Computationally essential	BLAST LP	75	93	60	84	111	37	38	49
	Model SEED	75	94	58	77	110	39	37	50
	Uniform LP	75	93	58	83	111	36	37	50
Computationally nonessential	BLAST LP	32	864	39	765	110	333	38	157
	Model SEED	32	863	41	772	111	331	39	156
	Uniform LP	32	863	39	766	110	334	39	156

Table 4. Essentiality predictions by gap-filled networks.

Compared to the Model SEED and uniformly-weighted gap-fillings, BLAST LP resulted in metabolic networks that had equal or improved predictions for both essential and nonessential genes in three out of four organisms. Surprisingly, the uniformly-weighted solutions, which always contained the fewest reactions, did not result in networks with more computationally essential genes. Essentiality predictions are compared to experimentally essential (EE) and nonessential (ENE) observations.

Figures

Figure 7

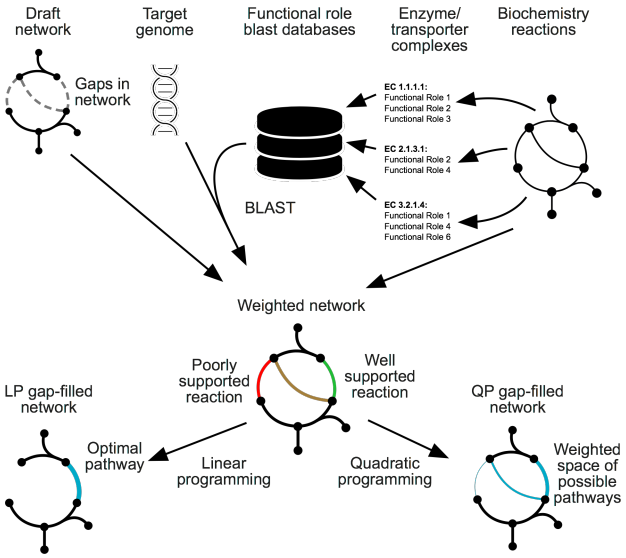


Figure 7. Gap-filling algorithm.

Weighted biochemistry databases were generated for target organisms by comparing the target genomes to functional role-specific BLAST databases for each known enzyme functional role in the RAST database. The best HSP returned from each database search was translated into a weight value for the reactions associated with the enzyme function. LP was used to select an optimally supported gap-filling solution from the weighted database and QP was used to identify a space of possible gap-filling solutions.

Figure 8

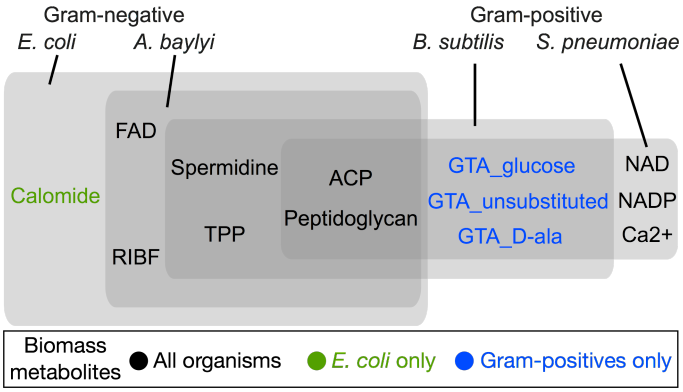


Figure 8. Comparison of metabolites that required unsupported reactions to become producible.

All four organisms shared a small subset of metabolites that required unsupported reactions. Further shared metabolite groups were Gram-specific, with Gram-negative species requiring fewer unsupported metabolites. Not all organisms had identical biomass equations, metabolites colored black were shared in all biomass equations, but metabolites colored green were specific to *E. coli* and metabolites colored blue were specific to *B. subtilis* and *S. pneumoniae*. Metabolite abbreviations: (FAD: flavin adenine dinucleotide, RIBF: riboflavin, ACP: acyl carrier protein, TPP: thiamine pyrophosphate, GTA: glycerol teichoic acid, NADP: nicotinamide adenine dinucleotide phosphate, NAD: nicotinamide adenine dinucleotide.)

Figure 9

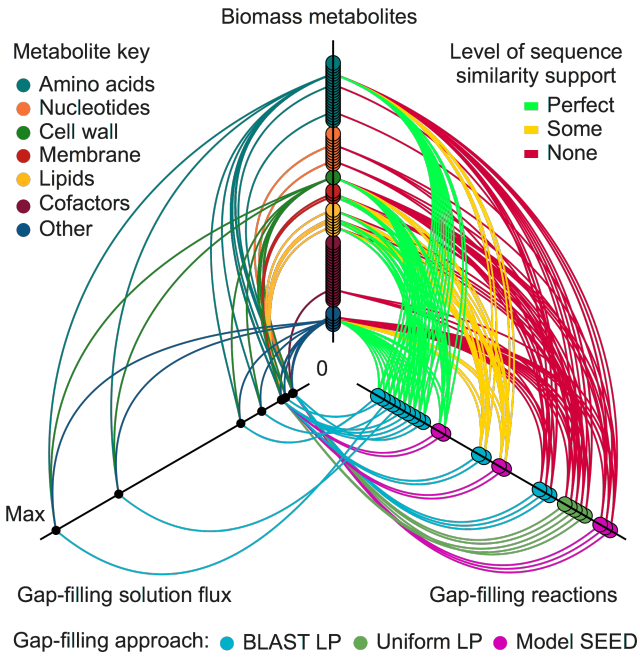


Figure 9. Role of reactions in *E. coli* gap-filling solutions.

Removal of a single reaction from the gap-filling solutions revealed metabolites for which that reaction was essential for metabolite production. This relationship is shown as lines connecting the gap-filling reaction axis to the biomass metabolites axis. Reactions are grouped by ESS and metabolites are grouped by class. A third axis illustrates the amount of flux through gap-filling reactions required for the production of a set biomass flux. In all cases, the gap-filling solutions included reactions with maximum ESS values, and for which no alternatives existed, in spite of the large space of potential gap-filling solutions. The BLAST LP gap-filling solutions minimized flux through unsupported reactions, yet a small flux through unsupported reactions was always required.

Figure 10

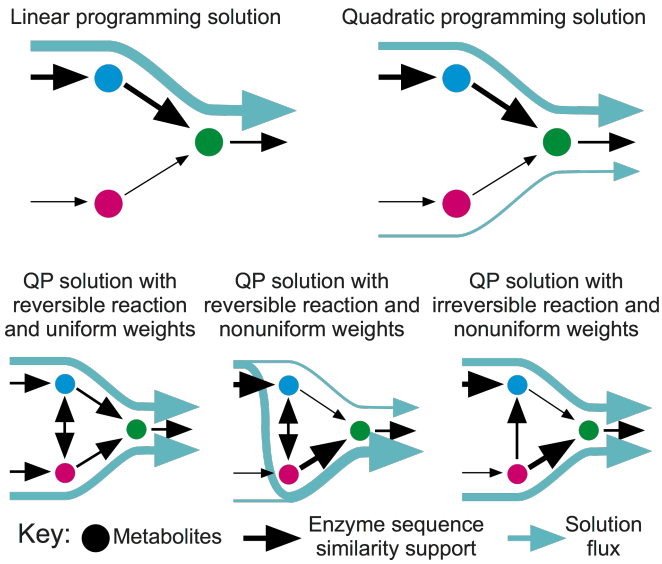


Figure 10. LP vs. QP gap-filling.

LP minimizes the weighted reaction fluxes to select the most supported pathway. QP minimizes the weighted squared flux, which distributes flux in inverse proportion to the pathway weights. However, QP does not result in flux through all possible solutions. Irreversibility of reactions may result in exclusion of reactions. The bottom figure shows how two different reaction weightings on the same network lead to two different flux solutions. If the reaction converting magenta metabolites to blue metabolites is irreversible, it will only be used in a QP flux solution if the blue to green pathway is favorable relative to the magenta to green pathway.

Figure 11

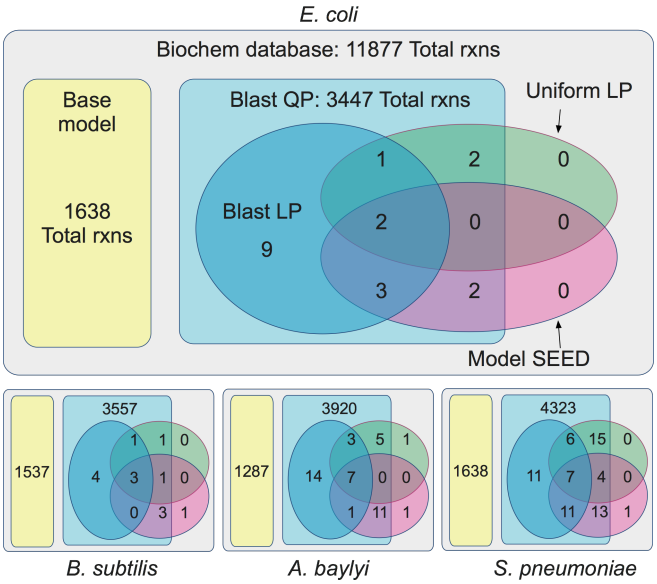


Figure 11. Overlaps between BLAST-weighted QP, BLAST-weighted LP, uniformly-weighted LP, and Model SEED gap-filling.

The QP gap-filling approach includes vastly more reactions than the other three gap-filling approaches, and almost all reactions from other methods were contained in the quadratic solution. Only the BLAST LP solution is guaranteed to be a subset of the QP gap-filling solution, because they use identical weights. The BLAST LP, uniformly-weighted LP, and Model SEED gap-filling approaches overlap, but are all unique and lead to distinct gene knockout predictions.

Figure 12

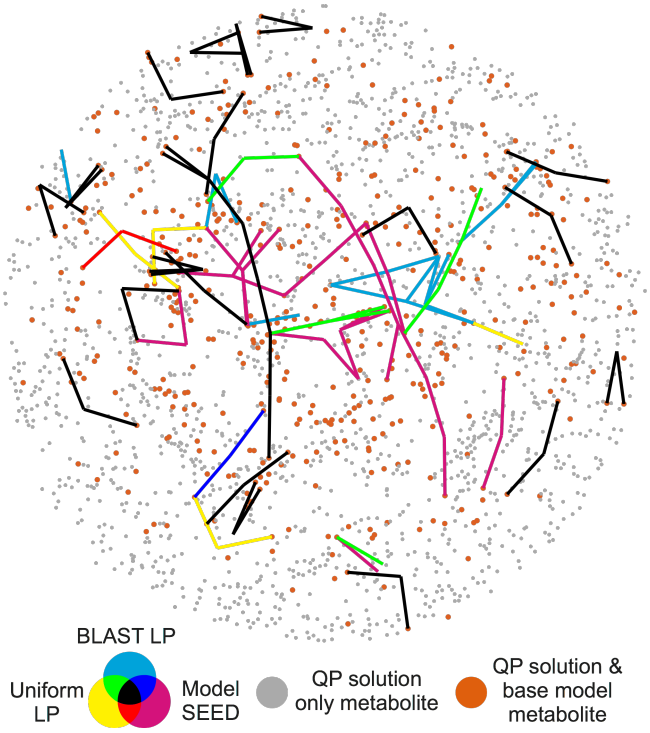


Figure 12. *S. pneumoniae* gap-filling solutions.

Metabolites contained in the QP solution were organized using force-directed network visualization. The BLAST LP, uniformly weighted LP and Model SEED gap-filling solutions are shown in separate colors. Metabolites that exist in both the draft metabolic model and the quadratic gap-filling solutions are orange, while metabolites that only exist in the quadratic gap-filling solution are gray. The QP gap-filling solution reveals the large space of potential gap-filling routes, as well as the high degree of connectivity with the draft metabolic network. Gap-filling solutions can begin and end in many parts of the

metabolic network, yet the network can be filled with a small subset of potential gap-filling reactions, as few as 32 reactions in this case.

Figure 13

Unsupported metabolite	Unsupported gap-filling reaction	Affected organisms
	Reaction formula	
	Comment	
Riboflavin (RIBF)	**rxn05039**	***E. coli, A. baylyi***
	5-Amino-6-5-phosphoribitylaminouracil + H2O	
	-> Phosphate + 4-1-D-Ribitylamino-5-aminouracil + H+	
The enzyme 5-amino-6-(5-phospho-D-ribitylamino)uracil phosphatase was recently discovered in *E. coli* (48) and was therefore not included in the Model SEED biochemistry database.		
FAD	**rxn05039**	***E. coli, A. baylyi***
	5-Amino-6-5-phosphoribitylaminouracil + H2O	
	-> Phosphate + 4-1-D-Ribitylamino-5-aminouracil + H+	
The same gap-filling reaction identified for riboflavin was also selected for FAD. Riboflavin is a precursor to FAD synthesis, and this result suggests that only the precursor metabolism was missing.		
Spermidine	**rxn00125**	***E. coli, A. baylyi, B. subtilis***
	H2O + S-Adenosyl-L-methionine	
	<=> L-Homoserine + H+ + MTA	
The activity of adenosylmethionine hydrolase has been observed in soil bacteria (49), but no associated gene was cataloged in the Model SEED, KEGG (50) (accessed May 7, 2015), or Biocyc (51) (accessed May 7, 2015) databases.		
Spermidine	**rxn13085**	***A. baylyi***
	Putrescine + S-Adenosyl-L-methionine	
	<=> CO2 + MTA + Spermidine	
The *A. baylyi* model was missing the spermidine synthesis step. The *A. baylyi* genome had no significant similarity to genes in the Model SEED database that could fill this role. Spermidine synthesis may occur through a mechanism that does not have homology with known synthesis pathways. Alternatively spermidine may be incorrectly included in the biomass equation of *A. baylyi*.		
TPP	**rxn12376**	***E. coli, A. baylyi, B. subtilis***
	4-Hydroxy-benzylalcohol	
	<=> 4-Hydroxy-benzylalcohol	
4-Hydroxy-benzyl alcohol transport does not contain a gene association in the Model SEED, KEGG, or Biocyc database. However, the ChemSpider entry for 4-Hydroxy-benzyl alcohol (http://www.chemspider.com/Chemical-Structure.122.html, accessed 14:52, May 7, 2015) lists a very high water solubility and large octane/water partition coefficient, which suggests that it may readily passively traverse membranes.		

Figure 13. Unsupported metabolite gap-filling reactions.

Chapter 4: Thermodynamic Constraints Improve Metabolic Networks

Synopsis

In pursuit of establishing a realistic metabolic phenotypic space, the reversibility of reactions is thermodynamically constrained in modern metabolic networks. The reversibility constraints follow from heuristic thermodynamic poise approximations that take anticipated cellular metabolite concentration ranges into account. Because constraints reduce the feasible space, draft metabolic network reconstructions may need more extensive reconciliation, and a larger number of genes may become essential. Notwithstanding ubiquitous application, the effect of reversibility constraints on the predictive capabilities of metabolic networks has not been investigated in detail. Instead, work has focused on the implementation and validation of the thermodynamic poise calculation itself. With the advance of fast linear programming-based network reconciliation, the effect of reversibility constraints on network reconciliation and gene essentiality predictions have become feasible and are the subject of this study. Networks with thermodynamically informed reversibility constraints outperformed gene essentiality predictions compared to networks that were constrained with randomly shuffled constraints. Unconstrained networks predicted gene essentiality as accurately as thermodynamically constrained networks but predicted substantially fewer essential genes. Networks that were reconciled with sequence similarity data and strongly enforced reversibility constraints outperformed all other networks. We conclude that metabolic

network analysis confirmed the validity of the thermodynamic constraints, and that thermodynamic poise information is actionable during network reconciliation.

Supplementary material is available at Biophysical Journal online, URL: https://www.cell.com/biophysj/home DOI: 10.1016/j.bpj.2017.06.018

Introduction

Metabolic networks provide a backbone for the integration of biological data and have emerged as a powerful complement to genome annotation by contextualizing the role of individual genes (*5*, *113*, *114*). They provide a genome-scale structure to organize organism specific knowledge (*37*, *115*, *116*) and facilitate the generation and evaluation of testable genotype-to-phenotype predictions (*117–119*). In metabolic networks, a biochemical database maps enzymes to reactions and their component metabolites. For well-studied organisms, experimental literature makes up the bulk of enzyme and associated gene annotations (*3*), leading to highly curated metabolic networks (*5*, *8*, *15*). For less studied organisms, enzyme activities are inferred from sequence similarity to known enzymes (*36*, *49*, *59*).

The explosion of sequencing data has driven efforts to automate both gene annotation and reconstruction of draft metabolic networks, and several algorithms and services have been developed to facilitate draft network reconstruction (*10*, *11*, *39*, *79*, *120*). One such service is the Model SEED (*10*), which automatically annotates genomes through Rapid Annotation using Subsystem Technology (RAST)(*34*, *36*) and reconstructs draft metabolic networks. Draft network reconstruction removes much of the initial work in creating an organism specific metabolic network. A curated common biochemistry

database standardizes reaction and metabolite names, facilitating the communication and analysis of metabolic networks (*46*). Once reconstructed, metabolic networks are used to investigate expected genetic interactions (*75*, *79*, *121*, *122*), gene essentiality (*52*, *83*, *123–125*), and characteristics of the feasible flux space (*126*, *127*). These evaluations often use Flux Balance Analysis to model metabolic flux in the network (*65*). Continuing effort is made to model flux as realistically as possible by imposing constraints on the allowable flux space (Constraint-Based Reconstruction and Analysis (*20*, *73*, *74*, *128*)). Constraints may be optimality-based, such as a maximum yield requirement or parsimony (*65*, *126*, *129*), but are also to prevent thermodynamically infeasible behavior such as circular flux (*50*, *130*) or flux against a strong thermodynamic poise (*64*, *67*, *131*). The first is easily imposed by minimizing flux through a network in combination with a fixed biomass production (*50*), and the latter can be imposed by assigning reversibility constraints to reactions (*132*).

Several heuristic approaches are used to estimate the thermodynamic poise for reactions, including "group contribution" (*67*, *133*, *134*), and more recently "component contribution" (*135*). In addition to feasible Gibbs free energy ranges calculated using group contributions, the Model SEED incorporates heuristic constraints based on literature values and canonical knowledge on reaction types (*59*). Due to the broad adoption of constraint based modeling for metabolic network reconstruction, the inclusion of reaction reversibility constraints has become ubiquitous in metabolic networks (*3*, *10*, *66*). Despite the ubiquity, the consequence of applying reaction reversibility constraints (RCs) to metabolic networks has not been investigated

extensively in relation to genome-scale gene essentiality predictions (GEP) and observations.

Note that the application of RCs on reactions reduces the accessible thermodynamically infeasible space to the model but does not eliminate this space. For instance, for one reaction to be feasible in the reverse direction, product concentrations in the upper physiological range may be required. A second reaction may produce these metabolites, but only if their concentrations are in the lower physiological range. Yet, both reactions are allowed simultaneously if reaction RCs are implemented without modeling metabolite concentrations, resulting in thermodynamically under-constrained RCs. The thermodynamically feasible flux space can be approached or achieved by using "Thermodynamics-based Metabolic Flux analysis" (TMFA) (*64*, *136*), Energy Balance Analysis" (EBA) (*137*) or TR-fluxmin (*138*), where metabolite concentrations are modeled in conjunction with fluxes.

Upon reconstruction, draft metabolic networks are incomplete and/or overly constrained (*3*, *54*, *56*) and lack the ability to synthesize a complete set of vital metabolites, such as DNA and amino acids, which are collectively referred to as biomass metabolites, or simply biomass (*3*, *6*). Before metabolic networks can be evaluated functionally, the network must be curated to ensure the inclusion of the complete synthesis of biomass from defined nutrient sources. This process of network reconciliation with observed biomass production is referred to as gap-filling in the literature (*3*), and can be accomplished through three primary modifications: (i) addition of internal reactions; (ii) addition of transport reactions, and; (iii) relaxation of reaction reversibility constraints (*3*,

53). Here, modifications (i) and (ii) are referred to as reaction addition (RA) and modification (iii) is referred to as reversibility constraint relaxation (RCR). Strictly, the terms gap and gap-filling could be considered misnomers. The term "gap" implies that the network is incomplete at a specific location that is discovered once a "gap-filling reaction" is found. This definition is problematic because there are multiple unique and independent ways of fixing incomplete metabolite synthesis, and it is often unclear which solution is more representative of the actual biochemistry of the organism (*56*). Rather, the network is inoperable, and is reconciled with an observed or inferred phenotype, specifically biomass production, although networks can be reconciled with other observations, such as growth/no-growth for a variety of mutants and nutrient conditions (*62*). Some network reconciliation (NR) strategies explicitly quantify the tradeoffs associated with each modification by assigning a penalty to each type of modification and computationally searching for a solution that has a minimal penalty (*10*, *39*, *55–57*, *138–140*). Although specific weighting schemes have been presented, no weighting scheme has been explicitly evaluated by quantifying the effect of the weighting parameters on the quality of network gene essentiality prediction (GEP).

Here, thermodynamically informed reaction reversibility constraints were evaluated for their effect on metabolic network predictions. Unconstrained models were compared to constrained ones, and uninformed constraints were compared to thermodynamically informed constraints. In addition, the role of how reaction RCs may be handled during the reconciliation stage of network reconstruction was investigated in detail. For this purpose, $\Delta_r G'^{\circ}$ values were used to select modifications among potential alternatives. To

test the quality of metabolic networks, GEP was compared to experimentally observed gene essentialities. GEP makes no assumptions on yield or rate optimalities, and only evaluate feasibility of biomass production for a given set of genes.

Experimental Procedures

Model SEED models and the reaction database

Metabolic networks for *Streptococcus pneumoniae*, *Bacillus subtilis*, *Escherichia coli* MG1655, and *Acinetobacter baylyi* ADP1 were downloaded from the Model SEED (http://seed- viewer.theseed.org/) along with media conditions and biomass formulations. The Model SEED biochemistry database was utilized for reconciliation, and models were stripped of RA modifications as described in detail previously (*56*). The Model SEED constrained the reversibility of reactions in the biochemistry database and draft metabolic models using a hybrid approach that incorporated group contribution estimation of $\Delta_r G'^{\circ}$, literature sources, and heuristic annotations based on chemical reaction classes (*10*, *59*).

Network reconciliation

Weighted linear programming, calculated using the CPLEX™ software (IBM, Armonk, New York, http://www.ibm.com) running in the Matlab™ programming environment (Mathworks, Natick, Massachusetts, http://www.mathworks.com/), was used to calculate reconciliation solutions that minimized the total flux through unsupported reactions in metabolic networks while simultaneously forcing a set flux through a defined biomass equation as previously described (*56*). RA selected by the reconciliation algorithm, as well as any RCR modifications, were retained in the organism specific reconciled

network. As before, each reversible reaction was separated into forward and reverse partial reactions for all fluxes to have a positive value, and weighted linear programming was used to minimize the sum of all reaction fluxes multiplied by the associated reaction weights. For all irreversible reactions, including those in the draft metabolic models, the disallowed partial reaction was made available as a reconciliation reaction, and a weight was added to the sequence-similarity weight. For all NR reactions, both partial reactions were assigned the same sequence-similarity weight. The reconciliation algorithm is given by:

$$\text{Minimize} \sum_{i=1}^{2r} (s_i + \text{rcr}_s rcr_i) v_i$$

$$\mathbf{S}v = 0$$

$$0 \leq v_i < ub_i$$

$$v_{bio} = 1e-3$$

Where *v*, *s*, and *rcr* are vectors of length 2*r* and v_i is the flux through a directional reaction, *i*. s_i and rcr_i are corresponding sequence and thermodynamic weights. rcr_s is a scaling factor that adjusts the relative effects of the two weighting vectors. The network is constrained to steady-state, where **S** is a *2r* by *m* irreversible stoichiometric matrix, with *r* reactions and *m* metabolites. Flux v_i, was constrained to be positive and less than the upper bound, ub_i. The biomass reaction, v_{bio}, was constrained to a flux of 1e-3, to ensure a set biomass production. Reconciled models included RA modifications for reconciliation reactions that carried a flux >1e-6 during reconciliation, and RCR modifications were retained for reactions for which the same minimal flux in the

penalized direction was observed in the linear programming (LP) solution. Reaction reversibility constraints were taken from the Model SEED (*10*), and calculated with the Von Bertalanffy toolbox (*135*) (Supplementary Material Section 1, available at Biophysical Journal online).

RA and RCR weighting vectors

For the *s* and *rcr* vectors, two specific versions were generated to test informed weights and uniform constraints of 0 and 1. Transformed BLAST sequence similarity e-values were used as the informed weighting vector for *s*, and were generated for each reaction in the Model SEED reaction data as described previously (*56*). The calculated $\Delta_r G'^\circ$ values in the Model SEED (*10*) reaction database were used as the informed *rcr* weighting vector. Negative $\Delta_r G'^\circ$ values lead to a positive weight equal to the absolute value of $\Delta_r G'^\circ$ on the reverse reaction. Positive $\Delta_r G'^\circ$ values lead to a positive weight on the forward reaction. A uniform *s* vector was generated where all NR reactions received a weight of 1 and all draft metabolic network reactions received a weight of 0. Similarly, a uniform *rcr* vector was generated based on the heuristic reaction reversibility constraints in the Model SEED reaction database. Both directions for reactions annotated as reversible received a weight of 0, and the disallowed directional reactions in irreversible reactions were weighted as 1. Thus, four NR weighted schemes were used, informed weights for both *s* and *rcr* (RCR_wRA_w), uniform weights for both *s* and *rcr* (RCR_uRA_u), and two mixed combinations, RCR_wRA_u and RCR_uRA_w.

Gene essentiality simulations

Reconciled networks were evaluated for their ability predict experimentally observed gene essentiality as before (*56*). Briefly: for each gene knockout, all reactions uniquely associated with that gene were removed from the metabolic network. The reduced network was evaluated for biomass production >1e-4 using LP. If no flux solution could be found, the gene was considered computationally essential, whereas the existence of a solution resulted in the gene being considered computationally nonessential. Computational essentiality was compared to experimental essentiality resulting from experimental whole-gene deletion studies for all four organisms (*80*, *123*, *141–143*). The quality of metabolic networks was assessed by comparing percent correct gene essentiality predictions (GEP) (correctly predicted essential genes + correctly predicted nonessential genes) / total number of evaluated genes, and the diagnostic odds ratio (DOR) of GEP (correctly predicted essential genes * correctly predicted nonessential genes) / (incorrectly predicted essential genes * incorrectly predicted nonessential genes).

Sensitivity analyses

Three types of sensitivity analyses were performed, (1) shuffling of RCs and RCR weights, (2) shuffling of RCR weights alone, and (3) shuffling of RCs and RCR weights with controlled portions of reversible and irreversible reactions. For (1), the reversibility constraints and corresponding *rcr* weights were randomly shuffled between reactions, allowing previously reversible reactions to become irreversible in either the forward or reverse direction and vice versa. Shuffling the constraints and weights ensured that the portion of reversible or irreversible reactions remained constant, along with the overall

distribution of values in the *rcr* vector. For (2), only the *rcr* vector was shuffled, leaving the original reversibility constraints in place. For (3), the total number reversible reactions were controlled by first making all reactions reversible, and then randomly selecting a fixed portion of reactions to be irreversible in either the forward or reverse direction. Uniform RCR and RA weights were used for subsequent NR, and the uniform RCR weights were generated from the randomly shuffled RCs, ensuring that the RCR weights matched the randomized RCs. The portion of forward to reverse irreversible reactions was controlled to be equal to that of the reaction database.

Results

Thermodynamically-informed constraints outperform random constraints

The space of allowable flux solutions is expected to decrease with the application of constraints to metabolic networks. To establish a realistic view of feasible phenotypic states of an organism, metabolic networks are constrained to disallow biologically infeasible states. One category of such constraints is reaction reversibility constraints, which are derived from Gibbs free energy calculations, literature sources, and heuristics based on reaction type (*59*). To test the effect of thermodynamically informed RCs on metabolic networks, networks constrained by the Model SEED RCs were compared to networks with randomly shuffled RCs and RCR weights. This shuffling approach retained the same portion of irreversible to reversible reactions, but removed the information provided by rationally defined reversibility constraints. The quality of networks was evaluated by comparing correct GEP, which was summarized as percent

correct GEP and the DOR of GEP. Genes were evaluated if they were annotated in the draft model, and if experimental gene essentiality data was available. Each random shuffling was repeated twenty times to generate representative outcomes. Following shuffling, networks were reconciled using LP to find a biomass synthesis solution while minimizing flux through poorly supported reactions or against RCs, penalizing flux through RA and RCR reactions equally.

Networks with thermodynamically-informed constraints improved both percent correct predictions and DORs over randomly shuffled thermodynamic constraints (Figure 14). For three out of four organisms, original networks achieved higher DOR than any network with randomized constraints, but for all organisms except *A. baylyi*, a network with the highest percent correct GEP resulted from one of the randomly shuffled networks. An alternative set of reversibility constraints calculated using the recently released version 2.0 of Von Bertalanffy toolbox (*135*) did not improve network quality compared to the constraints defined by the Model SEED (*10*) (Supplementary Material Section 1, available at Biophysical Journal online), and the Model SEED defined thermodynamics were used for further analyses.

Uninformed constraints degrade predictions

The effect of uninformed RCs on network quality was tested in more detail using the *E. coli* metabolic network and shuffling RCs while controlling the total portion of reversible and irreversible reactions. All reactions were initially made reversible, and successively more reversible reactions were randomly selected and constrained to be irreversible until all reactions were set to irreversible. At each step, twenty unique networks were

generated by randomly shuffling the constraints among reactions. Networks were reconciled using RCR_uRA_u NR with rcr_s=1, and evaluated for GEP to measure the quality of the reconciled networks. On average, network predictions degraded with the number of random constraints added, but network instances were found that outperformed unconstrained networks (Figure 15). Note that the unconstrained *E. coli* network (Figure 15, Supplementary Material Section 2 Table 2, available at Biophysical Journal online) outperformed the network containing default Model SEED constraints in terms of percent correct GEP, but not DOR (Figure 14, Supplementary Material Section 2 Table 2, available at Biophysical Journal online). The two networks had distinct advantages and disadvantages: The Model SEED-constrained network predicted many more essential genes and predicted them correctly much more frequently. In contrast, the unconstrained model identified more nonessential genes, and did so correctly more often (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). In the default *E. coli* metabolic network 41.34% of reactions are constrained, which suggests that the quality of the constraint-assignments fully compensates the negative effects associated with the addition of random constraints (Figure 15). Networks that were randomly constrained at 41.34%, included on average 4 more unsupported reactions and greater utilization of unsupported reactions than the network with informed constraints (Supplementary Material Section 3 Table 3, available at Biophysical Journal online). This separate measure of network quality corroborates the notion that the improvement in predictive power of a network that is constrained with informed constraints may be modest, but the informed constraints do not damage the network.

Using linear regression, we further observed that RA was slightly, but significantly, correlated with a degradation of nonessential gene predictions regardless of the support level of the NR-added reactions. However, correct essential predictions were harmed by the addition of unsupported reactions, but improved with the addition of supported reactions (Supplementary Material Section 4, available at Biophysical Journal online), suggesting that the addition of supported reactions has a small positive effect on the ability to correctly predict essential genes.

Networks with randomized constraints that were reconciled using NR that prioritized RA required fewer modifications and performed on par with networks that were reconciled while prioritizing RCR. Considering that the requirement for extensive NR was caused by random reversibility constraints, this was a surprising result. Separately, the RA-prioritized networks demonstrated that a substantial set of genes (~1/3 experimentally essential and ~1/6 of the experimentally nonessential genes) showed variable gene essentiality predictions in response to network constraints (Supplement Material Section 5, available at Biophysical Journal online).

Model predictions reveal tradeoffs in NR strategies

Erroneous reversibility constraints can result in networks that cannot produce biomass under any conditions. Consequently, attempting to overturn erroneous constraints is usually part of a NR process (*10*, *53*). Here, several NR strategies were explored to investigate how RCR and RA affects the quality of network predictions. Several existing NR algorithms make explicit tradeoffs between RCR and RA (*10*, *54*, *57*). For example, the Model SEED NR algorithm (*10*) preferentially uses RA over RCR, and individual

RCR modifications are weighted using $\Delta_r G'^\circ$ for the associated reactions. However, the consequences of the various tradeoff schemes in terms of GEP has not yet been reported. To investigate the tradeoffs between RCR and RA, a weighted linear programming-based NR algorithm was used that explicitly controlled the tradeoff through a scaling factor (rcr_s) and allowed for individual reactions to be weighed based on the calculated $\Delta_r G'^\circ$ (RCR_w) and sequence similarity values (RA_w). Large rcr_s values favored RA over RCR, while small rcr_s values lead to the opposite effect (Figure 16). Uniform weighting schemes that applied Boolean weights to RA (RA_u) and RCR (RCR_u) modifications for all reactions were compared to weighted RCR and weighted RA weighting schemes to investigate the value of the thermodynamic and sequence similarity weighting information. (Figure 17).

Networks reconciled by weighted RA showed consistently strong DOR of GEP across all organisms and for different rcr_s values (Figure 17). All organisms showed improved predictions for large rcr_s values relative to small rcr_s values. *S. pneumoniae* and *E. coli* had an optimal result for an rcr_s value of 1e-2, but *E.* coli had equal results for a range of rcr_s values from 1e-2 to 1e8. Overall, weighting of RA was more consequential than weighting of RCR. The *E. coli* metabolic network showed the clearest response to the doubly weighted NR. Doubly weighted NR with a large rcr_s value led to the best predicting network. Because the *E. coli* network was the most complete and best performing metabolic network examined, the network may be most sensitive to improvement. For the *E. coli* network, the benefit of a large scaling factor combined with

doubly weighted NR should strongly favor RA over RCR, except where RCR weights were equal to zero.

$\Delta_r G'^\circ$ values can qualitatively guide network reconciliation

Reactions with $\Delta_r G'^\circ$ values that had an opposite poise compared to the heuristically defined RC received a RCR weight of zero and were found to be the sole RCR that were observed during NR when rcr_s was large (Supplementary Material Section 6 Table 6, available at Biophysical Journal online). To investigate if the observed increase in DOR of GEP indeed resulted from heuristic reversibility constraints and RCR weights, a set of randomization controls were performed (Figure 18). For a range of rcr_s values, the *E. coli* metabolic network was reconciled using RCR_wRA_w and RCR_uRA_u. For each combination, two sensitivity analyses were performed: (i) shuffling of RC calls and RCR weights, or (ii) shuffling of only the RCR weights. Shuffling of both RC calls and RCR weights removed all information about the specific reaction directionality throughout the draft network and the reaction database, leaving the NR algorithm to find the cheapest biomass synthesis route through randomized RC and RCR weights. Shuffling just RCR weights tested the sensitivity of the NR algorithm to erroneous RCR weights, and also tested if rational weights consistently performed better than randomized weights.

All rcr_s values and NR weightings led to improved networks relative to networks with randomized RC and RCR weights, and just randomized RCR weights. This suggests that both RC and RCR weights contain information that improved network GEPs.

Randomized weighting vectors demonstrated that weighted NR in combination with large rcr_s values consistently outperformed the unweighted case relative to randomized

controls, suggesting that weighted vectors with large rcr_s values usefully guided NR. For both weighted and unweighted cases, superior networks were again observed after reconciliation with a randomized weight vector, indicating that reaction constraint sets exist that result in better DOR of GEP than rationally obtained constraints.

$\Delta_rG'^\circ$ weighted NR rationally overturns heuristics

Large rcr_s values avoided RCR in favor of RA (Figure 18), yet the RCR weighted case significantly improved networks over the uniformly weighted case. A difference between the uniform and RCR weighted cases occurred where the $\Delta_rG'^\circ$ value disagreed with the heuristic directionality assignment. In such cases, the RCR weighted NR approach overturned heuristically assigned constraints without cost. Overturning heuristic assignments based on corroborating evidence of $\Delta_rG'^\circ$, sequence-similarity weights, and biomass demand thus improved the predictive performance of metabolic networks. While heuristic annotations in general outperformed a purely $\Delta_rG'^\circ$ weighted approach using VBT values (Supplementary Material Section 1, available at Biophysical Journal online), selective disregard for heuristics were the predominant cause of the superior performance of RCR weighted NR. Indeed, the reaction reversibilities that were overturned by the RCR based approach were annotated with a $\Delta_rG'^\circ$ that implied a reverse poise compared to the heuristic irreversibility constraints (Supplementary Material Section 6, available at Biophysical Journal online). The $\Delta_rG'^\circ$ values were crosschecked with the Von Bertalanffy toolbox, which reported similar values for the majority of reactions (Supplementary Material Section 6, available at Biophysical Journal online).

In the *E. coli* metabolic network, a combination of weighted RA and RCR led to a significantly different reconciliation solution than weighted RA alone. By allowing RCR in the NR process, the total NR penalty to produce a unit of biomass decreased (Supplementary Material Section 6 Table 7, available at Biophysical Journal online). Overall, the total number of reactions required to produce biomass, including existing model reactions and reconciliation reactions, decreased from 489 to 476 in the dually weighted case (Figure 19, Supplementary Material Section 6 Table 7, available at Biophysical Journal online). This reconciliation algorithm shared eight reactions with the weighted RA case. For one shared reaction, no other alternatives existed, necessitating inclusion in any NR solution (*56*). Seven RA reactions differed between algorithms, and all but one of the variable reactions were sufficiently supported to incur no penalty for usage. The poorly supported reaction in the weighted RA was replaced in the dually weighted algorithm, thus reducing the number of unsupported reconciliation reactions required for biomass production (Figure 19, Supplementary Material Section 6 Table 7, available at Biophysical Journal online).

Note that the NR algorithm overturned reversibility constraints only when required for biomass production. When all disagreements between $\Delta_r G'^\circ$ and RCs were overturned, the percent correct GEP and DOR decreased (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). Similarly, penalizing the use of the same disagreeing reactions resulted in a decrease in model performance (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). This suggest that the

combination of biomass demand, sequence-similarity, and $\Delta_r G'^{\circ}$ estimates more accurately guided NR than the use of either alone.

Corroboration through alternate network quality assessment

To investigate how other quality metrics corroborate the DOR of GEP analysis for the *E. coli* network, producible metabolites and the biomass solution space were inspected for the same rsr_s range. Interestingly, high rsr_s values resulted in many more producible metabolites coinciding with a smaller biomass solution space (Figure 20). The tripling of producible metabolites must be associated with a particularly enabling RA, that is used as an alternative for RCR that solves for biomass demand during low rsr_s values. At the same time, the retentions of the RCs also led to a reduction in the biomass solution space, explaining the greater number essential gene predictions associated with higher rsr_s values (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). Both trends, the increase in producible metabolites and decrease in biomass solution space with increasing high rsr_s values were also visible for randomly constrained networks. This suggests that RA tends to create better access to all reactions. Conversely, RCR allows for more incorrect redundancy to biomass through pathways such as catabolic pathways, which are prevented from running in the anabolic direction in a correctly constrained network.

Discussion

Interpretation of gene essentiality outcomes

The comparison of predicted GE to experimentally observed GE results in one of four outcomes for each gene (represented by the confusion matrix, Figure 21). I e. each prediction can be correct or incorrect for either essential or nonessential genes: True Essential (TE), False Essential (FE), True Nonessential (TN) and, False Nonessential (FN). FE predictions can result from (i) inclusion of biomass compounds that are not vital, (ii) missing reactions (transporters), (iii) overly-constrained RC, (iv) under-annotation of promiscuous enzyme activities and, (v) an incomplete description of the experimental growth media. FN predictions are due to under-constrained aspects of the network such as (i) signaling: isozymes that are not expressed under the test conditions, (ii) lax reversibility constraints or (iii) inclusion of reactions in the model that are not present in the organism. Assigning network quality scores from the four quadrant scores is not straightforward and to some extent arbitrary. The metabolic interpretation of the scores is non-trivial, and the number of nonessential genes is much larger than the number of essential genes, which suggests that TE predictions are more valuable than TN predictions.

For an accurate network one should expect a nonzero outcome for FN because genes can be essential for reasons not captured by the network such as toxicity resulting from metabolite build-up. Consequently, the correct numbers for TE and TN are unknown, and a genome scale comparison cannot achieve a perfect score. However, an accurate network *should not* have FE outcomes, because if an organism is viable, its model must be able to produce all essential biomass components. In this work, sometimes accuracy (expressed as percent correct GEP), but mostly DOR was used to translate outcomes to a

network quality score. Although accuracy is the more straightforward interpretation, but in contrast to DOR, it does not correct for the uneven frequencies of nonessential and essential genes. By working with a scalar quality metric various NR approaches could be compared directly, but due to the uncertain expectation for FN, DOR scores are an imperfect measure and should be interpreted alongside other network quality measures. In contrast, minimization of FE outcomes is of clear importance, as FE predictions are a definite sign of shortcomings of the model.

Gene essentiality informs on network-wide parameters

Gene essentiality provides a conserved profile that is fundamental characteristic of a species (*51*). Here, GEP was used as measure of network accuracy to test the validity of informed RCs and to the usefulness of RCs in guiding NR of metabolic networks. Randomized models were used as controls to filter out random effects from true network improvements. A substantial set of genes (~1/3 experimentally essential and ~1/6 of the experimentally nonessential genes) showed variable GEP in response to network constraints, suggesting that the quality of investigated networks could be sensitively assessed with DOR of GEP.

Thermodynamically informed RCs improve metabolic networks

With the aid of randomized controls, we observed that the influence of thermodynamically informed RCs on DOR of GEP by networks was clear, but modest. Randomized constraints led to networks with substantially different predictions, a small number of which outperformed rational constraints. This suggests that annotated

constraints may have room for improvement, but it may also indicate that draft networks are incorrect and/or incomplete. In the Model SEED networks, about 41% of the reactions are directionally constrained. Compared to networks with 41% randomly constrained reactions, the Model SEED RC network clearly outperformed DOR of GEP, demonstrating that the thermodynamic inferences produce a valid signal. However, completely unconstrained networks performed on par with the Model SEED RC networks in terms of percent GEP suggesting that metabolic networks may be over-constrained in their current form. In addition, all four tested metabolic networks contained gene annotations that were computationally essential, but experimentally nonessential, again suggesting that the networks were incomplete or overly constrained. Note that notwithstanding that networks were over-constrained, all networks had high dimensional null spaces (hundreds of dimensions) indicating the large degree of under-determinacy of their stoichiometric matrices and thus flux spaces.

Computational nonessential/experimental essential scores may be the result of regulatory or signaling constraints that were not modeled in the metabolic network. Errors of this nature could arise even in accurate and complete networks, and can therefore not be interpreted without reservation (*3*). Consequently, the better percent correct prediction of the experimentally essential genes by the Model SEED default model must be interpreted with caution. Note that even the unconstrained network contained more predicted essential genes than observed essential genes (157 and 107 respectively), where the above suggest that one should expect to observe more essential genes than that one should predict. The NR reaction database contained 1012 reaction with the highest level

of support for *E. coli* K12, which further corroborates the suggestion that the draft *E. coli* model of the Model SEED is incomplete. Interestingly, addition of all 1012 reactions to the model barely reduces the number of essential gene predictions and results in the best network performance in terms of percent correct GEP and DOR of GEP after weighted NR (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). Conversely, the over-prediction of essential genes could also be due to the inclusion of nonessential metabolites in the *E. coli* biomass equation, as well as an incomplete description of the experimental growth media (*56*). The latter was not the result of an incomplete set of transporters, as allowing access of all media metabolites to the cell did not alter GEP. But, because of the much heavier weighting of essential genes, the DOR of GEP interpretation of network quality strengthened the case for thermodynamically-informed RC networks (Supplementary Material Section 2 Table 2, available at Biophysical Journal online). Altogether, correct RCs appear necessary for a complete set of essential gene predictions, where further addition of supported reactions appears to improve nonessential gene predictions.

Weighted NR balances competing NR objectives

Networks reconciled that favored RA over RCR predicted gene essentiality significantly better than other approaches for the *E. coli* network. The consistent improvement observed using weighted NR in combination with large rcr_s value suggested that RA should take priority over RCR with the exception of overturning heuristic reversibility constraints. When the heuristic annotations conflicted with $\Delta_r G'^{\circ}$-based constraints, our approach preferentially overturned reversibility constraints with conflicting sources of

information. The observed improvement may be interpreted as the poor quality of these reversibility annotations. Finally, the introduced weighted linear programming approach is not limited to the set of reconciliation information that was used here but may for instance be applied to evidence pertaining to properties such as cellular compartmentation or time of expression. Parameterization of scaling factors may again be compared against randomized models to ensure that the specific selection of weightings is resulting in optimal network quality.

Figures

Figure 14

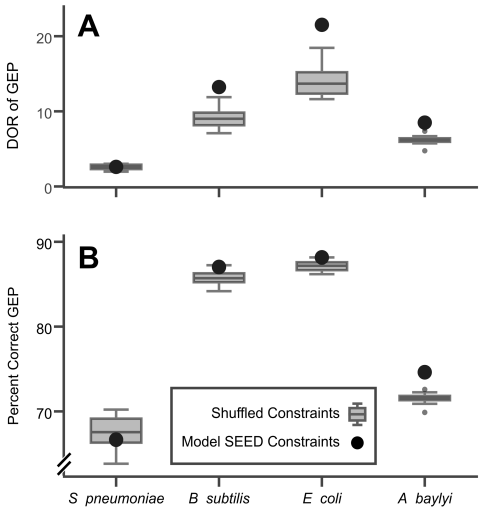


Figure 14. DOR and Percent Correct GEP for networks with randomly shuffled reaction direction constraints.

Networks for the four tested organisms were reconciled while equally penalizing both RA and RCR ($rcr_s = 1$) using RCR_uRA_u NR. DOR of GEP (plot A) and percent correct GEP (plot B) of unshuffled networks are shown as a large dot. Each box plot shows 20 randomized models for an organism that were reconciled and evaluated for GEP. The Model SEED RC consistently outperformed uninformed RC for both metrics, except for the *S. pneumoniae* network. The *E. coli* and *B. subtilis* networks outperformed the *A. baylyi* and *S. pneumonia* networks for both DOR and percent correct metrics, indicating a better quality of these draft networks.

Figure 15

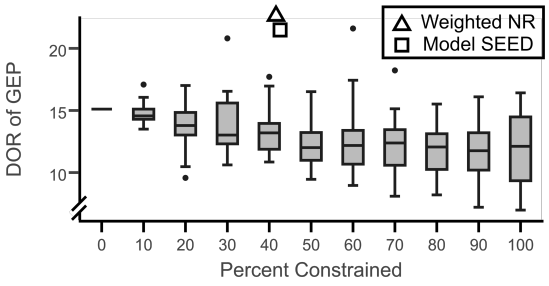


Figure 15. DOR of GEP as a function of uninformed constraints.

The percent of reaction reversibility constraints on the *E. coli* network was controlled by randomly constraining a portion of reactions in the *E. coli* metabolic network. The network was then reconciled while equally penalizing both RA and RCR ($rcr_s = 1$), followed by a gene essentiality evaluation. Each boxplot constitutes 20 randomly constrained networks. The average quality of predictions decreases with the number of constraints, but network instances were found that outperform the unconstrained case for all percent constrained values. For reference, 42.31% of reactions are constrained in the network reconciled by the Model SEED (black square), and 41.34% for the network reconciled using the weighted RA and weighted RCR algorithm, with a rcr_s value of 1e8 (black triangle). The weighted NR algorithm achieved the highest DOR of 22.6, while the Model SEED NR achieved a DOR of 21.5.

Figure 16

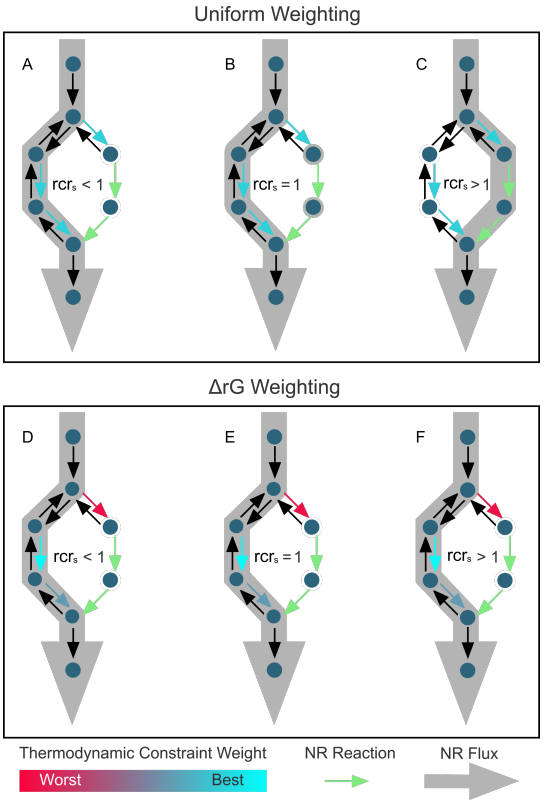


Figure 16. Network reconciliation overview.

Black arrows are reactions present in the unreconciled network. The network can be reconciled to carry flux by RCR or RA. A low scaling factor (rcr_s) preferentially causes RCR, and a high rcr_s preferentially results in RA. Using uniform weighting (A-C), reversibility constraints on reactions with a very strong thermodynamic poise may be overturned as easily as reactions that are annotated as near-reversible. Reaction specific weightings (D-F) prevent the relaxation of strongly poised reactions, which results in the preferential relaxation of multiple less strongly poised reactions (F). Note that under a low rcr_s, the poise weighting become less consequential than the number of added reactions (D).

Figure 17

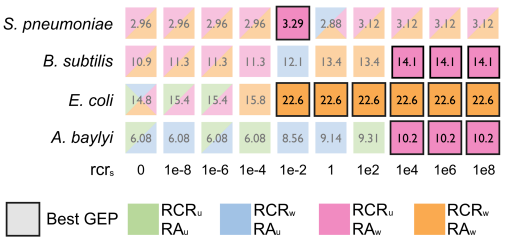


Figure 17. NR approaches with highest DOR of GEP as function of rcrs.

Four metabolic networks reconciled using four different NR approaches compared by their DOR of GEP for a range of rcr_s values (best DOR highlighted). Best performing NR approaches for a given rcr_s are color-coded. RCR_u indicates uniform reversibility constraints relaxation (RCR), RCR_w indicates weighted RCR. Similarly, RA_u indicates uniformly weighted penalties for reaction addition (RA), and RA_w indicates sequence similarity weighted reaction addition. Black borders around boxes indicate the highest DOR value for all tested cases. Note that unlike thermodynamic poise weights, sequence similarity weights consistently benefitted network GEP performance.

Figure 18

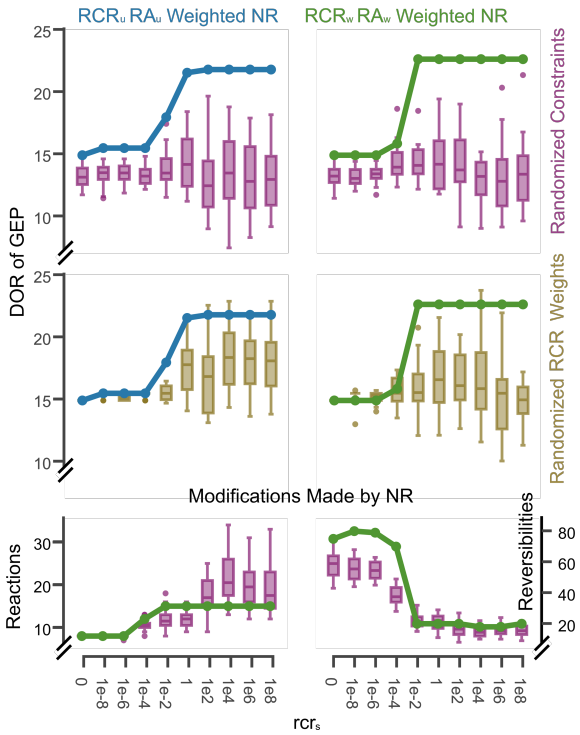


Figure 18. Sensitivity analysis of NR parameters.

Sensitivity of GEP to rcr_s values and NR weights were investigated through two approaches: (1) randomly shuffled RC and RCR weights (purple box plots), and (2), randomly shuffled RCR weights (yellow box plots). Shown are the uniformly weighted (RCR_uRA_u), and the doubly weighted approach, which uses $\Delta_rG'^\circ$ poise and sequence similarity weights (RCR_wRA_w). Blue lines represent DOR of GEP for the *E. coli* network reconciled with RCR_uRA_u NR, and Green lines represent DOR of GEP from the *E. coli* network reconciled with RCR_wRA_w NR. Box plots for each rcr_s value, NR method, and sensitivity analysis represent 20 randomized networks that were reconciled and evaluated for GEP. Both NR approaches outperform most networks with randomized constraints and randomized weights, but the RCR_wRA_w NR with large rcr_s values consistently

achieve the highest DOR. The number of NR modifications made using RCR_wRA_w NR and comparing to randomized constraints are shown at the bottom of the figure. Large rcr_s values predictably resulted in more RA and fewer RCR.

Figure 19

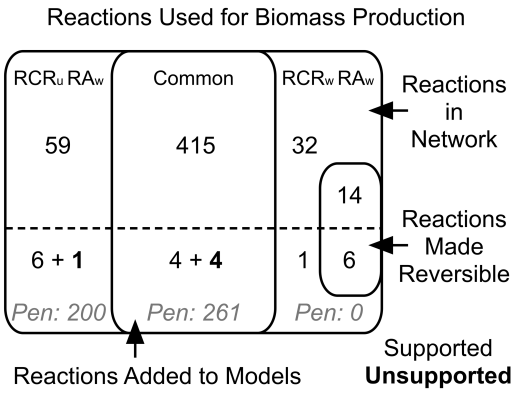


Figure 19. Reactions used in *E. coli* NR.

Compared to RCR_uRA_w NR, RCR_wRA_w NR allowed for twenty reactions to become reversible, and six of the relaxed constraints were for added reactions (rcr_s = 1e8). While both approaches added fifteen new reactions to the model, the RCR_wRA_w case led to a substantially different set of reactions required for biomass and overall thirteen fewer reactions. RCR_wRA_w NR also required the inclusion of one fewer reaction with less than perfect sequence support (bold), resulting in a smaller NR penalty contribution of the sequence similarity weighting.

Figure 20

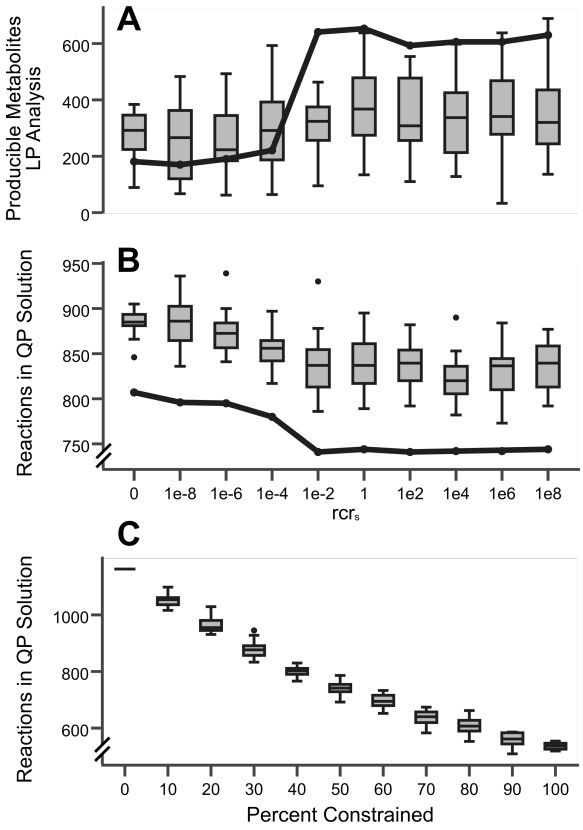


Figure 20. Alternate validating of reconciled networks.

Boxplots of various network properties for networks with randomly shuffled RC and the Model SEED RC (line). Networks were reconciled using RCR_wRA_w NR. (A) Metabolites that can be produced from available nutrients when allowing all metabolites to leave the cell, determined using LP. (B) Reactions carrying significant flux in a QP solution for biomass production, indicative for the size of the solution space. (C) QP solution space as function of percent RC (rcr_s = 1e8). Note that large rcr_s values led to networks with smaller QP solutions but that produced more metabolites.

Figure 21

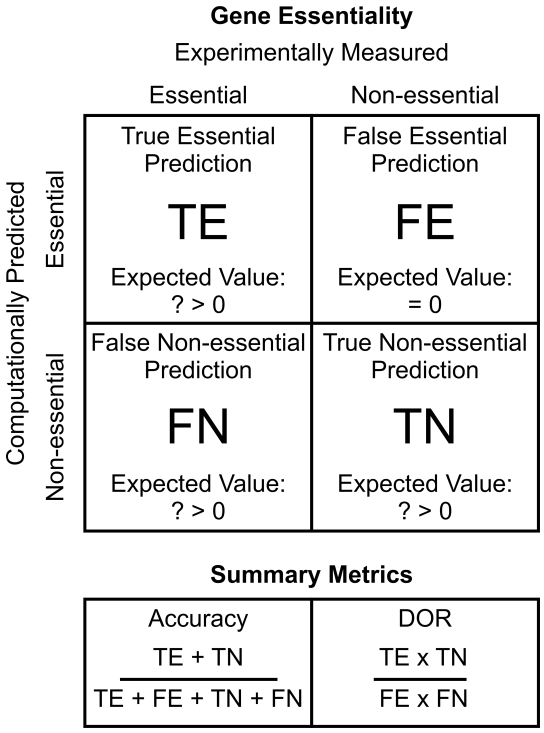


Figure 21. Gene essentiality prediction confusion matrix.

Computational vs. experimental gene essentiality outcomes were compared using quantities associated with the confusion matrix. For metabolic networks, the expected value of TE, TN, and FN is unknown. Despite being a false prediction, TN is expected >0 because some genes are essential for non-metabolic reasons. Only FE predictions are expected to be 0 for a correct metabolic network, FE>0 indicates missing functionality in the metabolic network.

Chapter 5: Conclusions

The research presented in the previous chapters followed two major themes: synthesizing data to improve network reconstruction and reconciliation, and developing controls for metabolic networks to test network validity and underlying assumptions. The network reconciliation algorithms developed use specific data, such as sequence comparison scores, but the underlying approaches are flexible and can be modified to accommodate a wide variety of data, which will be useful as new types of high throughput data become available for more organisms. Gene essentiality is used to evaluate network reconciliation methods and compare rational approaches to randomized controls, which provide a necessary measure of success for evaluating the resulting metabolic networks.

Chapter 2 focused on the reconstruction of the metabolic networks of the picoalga *Ostreococcus tauri* and *O. lucimarinus*, and advanced a new concept for network reconciliation. Phylogenetically structured tiers of reactions were preferentially used to reconcile biomass production, minimizing the inclusion of reactions from distantly related organisms as an alternative to previous approaches that simply minimized the total number of added reactions. The reconstruction process allowed for bottom-up and top-down network reconciliation to be compared in a common computational framework, highlighting the strengths of the top-down approach. Network reconciliation also revealed missing annotations in the *O. tauri* genome, which lacked a gene annotated with an essential enzyme involved in the Calvin cycle. The phylogenetically tiered approach to network reconstruction can be advanced to include a more explicit and fine-grained

approach to estimating the phylogeny, even down to the level of individual enzymes. Future work can also explore the effects of horizontal gene transfer in metabolism and the confounding effects it may have on phylogenetic approaches to metabolic network reconstruction.

Chapters 3 and 4 dealt with the network reconciliation of well-studied bacterial metabolic networks for which experimental gene essentiality data sets were available. Chapter 3 introduced two new computational tools for probing metabolic networks and finding reconciliation solutions. An LP algorithm was used to efficiently find reconciliation reactions by minimizing metabolic flux required to produce biomass metabolites. This algorithm is further improved by weighting reactions based on the target organism's genome sequence, which was compared to a large set of enzyme sequences. A second QP algorithm was used to probe the space of possible reconciliation solutions, and revealed that the concept of defined "gaps" in metabolic networks is illusory. The QP algorithm revealed a space of reconciliation solutions spanning thousands of reactions, illustrating that missing functionality in a metabolic network can be reconciled in many ways, and selecting a "best" solution is not straightforward. This surprising result motivated the development of negative controls on the reconciliation process to ensure that the solutions discovered were more accurate than random solutions, which was indeed the case.

Chapter 4 built on chapter 3 by combining sequence similarity data with thermodynamic data, specifically a calculation of Gibbs energy for metabolic reactions under biological conditions. The combination of two data types laid a groundwork for

synthesizing multiple independent data types into a network reconciliation algorithm. Although existing reconciliation algorithms have used these types of data, this research was the first to report a systematic analysis of the network reconciliation parameters. Furthermore, the reconciliation parameters are evaluated for multiple networks using gene essentiality and compared to randomized controls to identify reconciliation approaches that select solutions from the large space of possible solutions. This approach revealed the value of prioritizing reaction addition over thermodynamic constraint relaxation, but also demonstrated that heuristic assumptions of reaction irreversibility can be usefully overturned when multiple lines of evidence disagree with heuristics.

The approaches developed in chapter 4 also allowed for the evaluation of reaction reversibility constraints in general. Even early network reconstructions were constrained by making certain reactions irreversible (*144*). While these constraints were reasonable given experimental assays, the value of irreversibility constraints in metabolic networks had not previously been demonstrated. The randomized controls and novel reconciliation approach allowed for unconstrained networks to be compared to constrained networks. Surprisingly, unconstrained networks were shown to predict gene essentiality with similar accuracy as rationally constrained metabolic networks, highlighting the importance of testing assumptions in computational models.

Randomization and experimental validation could be used to probe other aspects of metabolic networks that have not been thoroughly analyzed. Aspects of particular interest include: 1) Compartmentation and transportation of molecules in metabolic networks. 2) Classification of essential biomass metabolites while accounting for a range

of nutrient and environmental conditions. 3) Promiscuous enzyme functionalities. 4) Spontaneous chemical reactions.

While truly accurate whole-cell models of microorganisms are still far away, incremental progress can be made by steadily modeling more organisms, integrating new and more accurate data, and most importantly, carefully validating methods and models at each step.

References

1. H. Kitano, Computational systems biology. *Nature*. **420**, 206–10 (2002).
2. J. R. Karr *et al.*, A whole-cell computational model predicts phenotype from genotype. *Cell*. **150**, 389–401 (2012).
3. I. Thiele, B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121 (2010).
4. R. D. Fleischmann *et al.*, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*. **269**, 496–512 (1995).
5. J. S. Edwards, B. O. Palsson, Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–6 (1999).
6. A. M. Feist, B. O. Palsson, The biomass objective function. *Curr. Opin. Microbiol.* **13**, 344–9 (2010).
7. J. S. Edwards, B. O. Palsson, The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci.* **97**, 5528–5533 (2000).
8. J. D. Orth *et al.*, A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol. Syst. Biol.* **7**, 535 (2011).
9. B. D. Heavner, K. Smallbone, N. D. Price, L. P. Walker, *Database (Oxford).*, in press, doi:10.1093/database/bat059.
10. C. S. Henry *et al.*, High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28**, 977–82 (2010).
11. S. M. D. Seaver *et al.*, High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9645–50 (2014).
12. M. I. Sigurdsson, N. Jamshidi, E. Steingrimsson, I. Thiele, B. Ø. Palsson, A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Syst. Biol.* **4**, 140 (2010).
13. S. Seo, H. a Lewin, Reconstruction of metabolic pathways for the cattle genome.

BMC Syst. Biol. **3**, 33 (2009).
14. N. C. Duarte *et al.*, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1777–82 (2007).
15. I. Thiele *et al.*, A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425 (2013).
16. J. Forster, I. Famili, B. O. Palsson, J. Nielsen, Genome-Scale Reconstruction of the Saccharomyces Cerevisie Metabolic Network. *Genome Res.*, 244–253 (2003).
17. N. C. Duarte, M. J. Herrgård, B. Ø. Palsson, Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–309 (2004).
18. J. K. Liu *et al.*, Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. *BMC Syst. Biol.* **8**, 110 (2014).
19. S. M. D. Seaver, C. S. Henry, A. D. Hanson, Frontiers in metabolic reconstruction and modeling of plant genomes. *J. Exp. Bot.* **63**, 2247–58 (2012).
20. E. J. O'Brien, J. a. Lerman, R. L. Chang, D. R. Hyduke, B. Ø. O. Palsson, Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693–693 (2014).
21. Y. Zhang *et al.*, Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. *Science (80-.).* **325**, 1544–1549 (2009).
22. R. L. Chang *et al.*, Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. *Science*. **340**, 1220–3 (2013).
23. W. R. Harcombe *et al.*, Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* **7**, 1104–15 (2014).
24. H.-C. Chiu, R. Levy, E. Borenstein, Emergent biosynthetic capacity in simple microbial communities. *PLoS Comput. Biol.* **10**, e1003695 (2014).
25. A. R. Zomorrodi, M. M. Islam, C. D. Maranas, d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth. Biol.* **3**, 247–57 (2014).
26. N. C. Kyrpides, Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat. Biotechnol.* **27**, 627–32 (2009).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).
28. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics*. **10**, 421 (2009).
29. S. R. Eddy, Multiple alignment using hidden Markov models. *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.* **3**, 114–20 (1995).
30. S. R. Eddy, A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.* **4**, e1000069 (2008).
31. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41** (2013), doi:10.1093/nar/gkt263.

32. J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29**, 2607–18 (2001).
33. F. Meyer *et al.*, GenDB - An open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187–2195 (2003).
34. R. K. Aziz *et al.*, The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. **9**, 75 (2008).
35. F. Meyer *et al.*, The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*. **9**, 386 (2008).
36. R. Overbeek *et al.*, The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206-14 (2014).
37. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
38. A. P. Arkin *et al.*, The DOE Systems Biology Knowledgebase (KBase). *bioRxiv* (2016), doi:10.1101/096354.
39. R. Agren *et al.*, The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. *PLoS Comput. Biol.* **9**, e1002980 (2013).
40. J. Schellenberger *et al.*, Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6**, 1290–307 (2011).
41. Z. A. King *et al.*, BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2016).
42. R. Caspi *et al.*, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
43. I. M. Keseler *et al.*, EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Res.* **41**, 605–612 (2013).
44. N. D. Price, J. L. Reed, B. Ø. Palsson, Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–97 (2004).
45. M. Hucka *et al.*, The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*. **19**, 524–531 (2003).
46. A. Kumar, P. F. Suthers, C. D. Maranas, MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*. **13**, 6 (2012).
47. S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **7**, 23 (2015).
48. N. M. O'Boyle, Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **4**, 22 (2012).
49. E. W. Krumholz, H. Yang, P. Weisenhorn, C. S. Henry, I. G. L. Libourel, Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *J.*

Exp. Bot. **63**, 2353–62 (2012).
50. J. Schellenberger, N. E. Lewis, B. Ø. Palsson, Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophys. J.* **100**, 544–53 (2011).
51. G. Plata, C. S. Henry, D. Vitkup, Long-term phenotypic evolution of bacteria. *Nature*. **517**, 369–372 (2014).
52. J. D. Orth, B. Palsson, Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. *BMC Syst. Biol.* **6**, 30 (2012).
53. V. Satish Kumar, M. S. Dasika, C. D. Maranas, Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*. **8**, 212 (2007).
54. J. L. Reed *et al.*, Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 17480–4 (2006).
55. N. Christian, P. May, S. Kempa, T. Handorf, O. Ebenhöh, An integrative approach towards completing genome-scale metabolic networks. *Mol. Biosyst.* **5**, 1889–903 (2009).
56. E. W. Krumholz, I. G. L. Libourel, Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks. *J. Biol. Chem.* **290**, 19197–19207 (2015).
57. M. N. Benedict, M. B. Mundy, C. S. Henry, N. Chia, N. D. Price, Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. *PLoS Comput. Biol.* **10**, e1003882 (2014).
58. E. Pitkänen *et al.*, Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput. Biol.* **10**, e1003465 (2014).
59. C. S. Henry, J. F. Zinner, M. P. Cohoon, R. L. Stevens, iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. *Genome Biol.* **10**, R69 (2009).
60. I. Thiele, N. Vlassis, R. M. T. Fleming, fastGapFill: efficient gap filling in metabolic networks. *Bioinformatics*. **30**, 2529–31 (2014).
61. J. D. Orth, B. Ø. Palsson, Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* **107**, 403–12 (2010).
62. V. S. Kumar, C. D. Maranas, GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput. Biol.* **5**, e1000308 (2009).
63. A. R. Zomorrodi, P. F. Suthers, S. Ranganathan, C. D. Maranas, Mathematical optimization applications in metabolic networks. *Metab. Eng.* **14**, 672–686 (2012).
64. C. S. Henry, L. J. Broadbelt, V. Hatzimanikatis, Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–805 (2007).
65. J. D. Orth, I. Thiele, B. Ø. Palsson, What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–8 (2010).
66. A. Kümmel, S. Panke, M. Heinemann, Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*. **7**, 512 (2006).
67. M. D. Jankowski, C. S. Henry, L. J. Broadbelt, V. Hatzimanikatis, Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487–99 (2008).

68. E. Noor *et al.*, An integrated open framework for thermodynamics of reactions that combines accuracy and coverage. *Bioinformatics*. **28**, 2037–44 (2012).
69. R. M. T. Fleming, I. Thiele, von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. *Bioinformatics*. **27**, 142–3 (2011).
70. A. Varma, B. O. Palsson, Parametric sensitivity of stoichiometric flux balance models applied to wild-typeEscherichia coli metabolism. *Biotechnol. Bioeng.* **45**, 69–79 (1995).
71. J. S. Edwards, M. Covert, Minireview Metabolic modelling of microbes : the flux-balance approach. **4**, 133–140 (2002).
72. A. Varma, B. O. Palsson, Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731 (1994).
73. R. U. Ibarra, J. S. Edwards, B. O. Palsson, Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. **420**, 20–23 (2002).
74. R. Adadi, B. Volkmer, R. Milo, M. Heinemann, T. Shlomi, Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput. Biol.* **8**, e1002575 (2012).
75. B. Vandersluis *et al.*, Broad metabolic sensitivity profiling of a prototrophic yeast deletion collection. *Genome Biol.* **15**, R64 (2014).
76. R. Zarecki *et al.*, Maximal sum of metabolic exchange fluxes outperforms biomass yield as a predictor of growth rate of microorganisms. *PLoS One*. **9**, e98372 (2014).
77. A. R. Joyce *et al.*, Experimental and computational assessment of conditionally essential genes in Escherichia coli. *J. Bacteriol.* **188**, 8259–71 (2006).
78. A. P. Burgard, P. Pharkya, C. D. Maranas, Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–57 (2003).
79. J. M. Dreyfuss *et al.*, Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM. *PLoS Comput. Biol.* **9**, e1003126 (2013).
80. T. Baba *et al.*, *Mol. Syst. Biol.*, in press, doi:10.1038/msb4100050.
81. T. van Opijnen, K. L. Bodi, A. Camilli, Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods*. **6**, 767–72 (2009).
82. J. M. Peters *et al.*, A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. *Cell*. **165**, 1493–1506 (2016).
83. H. Yang *et al.*, Genome-scale metabolic network validation of Shewanella oneidensis using transposon insertion frequency analysis. *PLoS Comput. Biol.* **10**, e1003848 (2014).
84. E. V. Koonin, Y. I. Wolf, Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–719 (2008).
85. C. S. Henry *et al.*, Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochim. Biophys. Acta*. **1810**, 967–77 (2011).

86. X. Wang *et al.*, The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* **43**, 1035–9 (2011).
87. F. Zhu, R. Massana, F. Not, D. Marie, D. Vaulot, Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005).
88. E. Derelle *et al.*, Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11647–52 (2006).
89. B. Palenik *et al.*, The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7705–10 (2007).
90. F. Corellou *et al.*, Clocks in the green lineage: comparative functional analysis of the circadian architecture of the picoeukaryote ostreococcus. *Plant Cell*. **21**, 3436–49 (2009).
91. N. Grimsley, B. Péquin, C. Bachy, H. Moreau, G. Piganeau, Cryptic sex in the smallest eukaryotic marine green alga. *Mol. Biol. Evol.* **27**, 47–54 (2010).
92. C. Courties *et al.*, Smallest eukaryotic organism. *Nature*. **370** (1994), pp. 255–255.
93. M.-J. Chrétiennot-Dinet *et al.*, A new marine picoeucaryote: Ostreococcus tauri gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia*. **34**, 285–292 (1995).
94. M. Kanehisa *et al.*, From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354-7 (2006).
95. S. S. Merchant *et al.*, The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*. **318**, 245–50 (2007).
96. M. G. Poolman, L. Miguet, L. J. Sweetlove, D. A. Fell, A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol.* **151**, 1570–81 (2009).
97. C. G. de Oliveira Dal'Molin, L.-E. Quek, R. W. Palfreyman, S. M. Brumbley, L. K. Nielsen, AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol.* **152**, 579–89 (2010).
98. P. May, J.-O. Christian, S. Kempa, D. Walther, ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. *BMC Genomics*. **10**, 209 (2009).
99. P. May *et al.*, Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics*. **179**, 157–66 (2008).
100. S. M. Keating, B. J. Bornstein, A. Finney, M. Hucka, SBMLToolbox: An SBML toolbox for MATLAB users. *Bioinformatics*. **22**, 1275–1277 (2006).
101. T. F. Smith, M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–7 (1981).
102. M. L. Mavrovouniotis, Estimation of standard Gibbs energy changes of biotransformations. *J. Biol. Chem.* **266**, 14440–5 (1991).
103. R. G. Forsythe, P. D. Karp, M. L. Mavrovouniotis, Estimation of equilibrium constants using automated group contribution methods. *Comput. Appl. Biosci.* **13**, 537–43 (1997).

104. A. Flamholz, E. Noor, A. Bar-Even, R. Milo, eQuilibrator--the biochemical thermodynamics calculator. *Nucleic Acids Res.* **40**, D770-5 (2012).
105. F. D. Ciccarelli *et al.*, Toward automatic reconstruction of a highly resolved tree of life. *Science*. **311**, 1283–7 (2006).
106. M. A. Larkin *et al.*, Clustal W and Clustal X version 2.0. *Bioinformatics*. **23**, 2947–8 (2007).
107. D. Posada, jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.* **25**, 1253–1256 (2008).
108. D. J. Zwickl, thesis, University of Texas at Austin (2006).
109. M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, T. Ideker, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. **27**, 431–2 (2011).
110. J. P. Faria, M. Rocha, R. L. Stevens, C. S. Henry, in *Advances in Bioinformatics*, M. P. Rocha, F. F. Riverola, H. Shatkay, J. M. Corchado, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010), pp. 209–215.
111. N. R. Boyle, J. A. Morgan, Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst. Biol.* **3**, 4 (2009).
112. R. L. Chang *et al.*, Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.* **7**, 518 (2011).
113. J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, E. Gilles, Metabolic network structure determines key aspects of functionality and regulation. *Nature*. **420**, 190–193 (2002).
114. P. Khatri, M. Sirota, A. J. Butte, Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **8** (2012), doi:10.1371/journal.pcbi.1002375.
115. M. B. Biggs, J. A. Papin, Metabolic Network-Guided Binning of Metagenomic Sequence Fragments. *Bioinformatics*, 1–8 (2015).
116. R. Caspi *et al.*, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, 459–471 (2014).
117. N. E. Lewis, H. Nagarajan, B. O. Palsson, Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305 (2012).
118. A. Bordbar, J. M. Monk, Z. a King, B. O. Palsson, Constraint-based models predict metabolic and associated cellular functions. *Nat. Rev. Genet.* **15**, 107–20 (2014).
119. E. J. O'Brien, J. M. Monk, B. O. Palsson, Using Genome-scale Models to Predict Biological Capabilities. *Cell*. **161**, 971–987 (2015).
120. T. J. Mueller, B. M. Berla, H. B. Pakrasi, C. D. Maranas, Rapid construction of metabolic models for a family of Cyanobacteria using a multiple source annotation workflow. *BMC Syst. Biol.* **7**, 142 (2013).
121. B. Szappanos *et al.*, An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat. Genet.* **43**, 656–62 (2011).
122. A. Typas *et al.*, High-throughput , quantitative analyses of genetic interactions in E

. coli. **5**, 781–787 (2008).
123. M. Durot *et al.*, Iterative reconstruction of a global metabolic model of Acinetobacter baylyi ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst. Biol.* **2**, 85 (2008).
124. O. Folger *et al.*, Predicting selective drug targets in cancer through metabolic networks. *Mol. Syst. Biol.* **7**, 501 (2011).
125. Y.-K. Oh, B. O. Palsson, S. M. Park, C. H. Schilling, R. Mahadevan, Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.* **282**, 28791–9 (2007).
126. R. Schuetz, L. Kuepfer, U. Sauer, Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Mol. Syst. Biol.* **3**, 119 (2007).
127. C. H. Schilling, D. Letscher, B. O. Palsson, Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–48 (2000).
128. J. S. Edwards, R. U. Ibarra, B. O. Palsson, In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–30 (2001).
129. N. E. Lewis *et al.*, Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 1–13 (2010).
130. H. Qian, D. A. Beard, S. D. Liang, Stoichiometric network theory for nonequilibrium biochemical systems. *Eur. J. Biochem.* **270**, 415–421 (2003).
131. V. S. Martínez, L.-E. Quek, L. K. Nielsen, Network thermodynamic curation of human and yeast genome-scale metabolic models. *Biophys. J.* **107**, 493–503 (2014).
132. A. M. Feist *et al.*, A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, 121 (2007).
133. M. L. Mavrovouniotis, Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. *Biotechnol. Bioeng.* **36**, 1070–82 (1990).
134. R. M. T. Fleming, I. Thiele, H. P. Nasheuer, Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. *Biophys. Chem.* **145**, 47–56 (2009).
135. E. Noor, H. S. Haraldsdóttir, R. Milo, R. M. T. Fleming, Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* **9**, e1003098 (2013).
136. J. J. Hamilton, V. Dwivedi, J. L. Reed, Quantitative assessment of thermodynamic constraints on the solution space of genome-scale metabolic models. *Biophys. J.* **105**, 512–22 (2013).
137. D. A. Beard, S. Liang, H. Qian, Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83**, 79–86 (2002).
138. A. Hoppe, S. Hoffmann, H.-G. Holzhütter, Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux

distributions in metabolic networks. *BMC Syst. Biol.* **1**, 23 (2007).
139. H.-G. Holzhütter, The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *Eur. J. Biochem.* **271**, 2905–22 (2004).
140. S. Holzhütter, H.-G. Holzhütter, Computational design of reduced metabolic networks. *Chembiochem*. **5**, 1401–22 (2004).
141. K. Kobayashi *et al.*, Essential Bacillus subtilis genes. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 4678–83 (2003).
142. J. A. Thanassi, S. L. Hartman-Neumann, T. J. Dougherty, B. A. Dougherty, M. J. Pucci, Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. *Nucleic Acids Res.* **30**, 3152–62 (2002).
143. J.-H. Song *et al.*, Identification of essential genes in Streptococcus pneumoniae by allelic replacement mutagenesis. *Mol. Cells*. **19**, 365–374 (2005).
144. D. A. Fell, J. R. Small, Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* **238**, 781–786 (1986).