Bi-Level optimization algorithms for improving genome-scale metabolic models

Inaugural-Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Daniel Hartleb

aus Bergisch Gladbach

Düsseldorf, Juni 2019

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Martin Lercher

2. Prof. Dr. Oliver Ebenhöh

Tag der mündlichen Prüfung:
20.09.2021

Erklärung

Ich versichere an Eides statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf" erstellt worden ist. Die Dissertation habe ich in dieser oder in ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen oder erfolgreichen Promotionsversuche unternommen.

Düsseldorf, den 11.06.2019

Daniel Hartleb

Contents

Abbreviations

BiGG	Biochemical, Genetic and Genomic knowledge base
COBRA	Constraint-Based Reconstruction and Analysis
EC	Enzyme commission
EGC	Energy generating cycle
gDW	Dry weight of the cell in grams
FBA	Flux Balance Analysis
FNp	False negative prediction
FPp	False positive prediction
FVA	Flux Variability Analysis
GAM	Growth associated maintenance reaction
GSM	Genome-scale metabolic network
GO	Gene ontologies
GPR	Gene-Protein-Reaction association
LP	Linear programming
MCC	Matthews correlation coefficient
MILP	Mixed-integer linear programming
MOMA	Minimization of metabolic adjustments
NGAM	Non-growth associated maintenance reaction
QM	Quadratic programming
ROOM	Regulatory on/off minimization
SBML	Systems Biology Markup Language

TNp	True negative prediction
TN-seq	Transposon sequencing
TPp	True positive prediction
SyBiL	Systems Biology Library

Preface

This document was prepared according to the ‘Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf vom 15.06.2018‘. Three manuscripts are presented along with an introduction that puts them into the broader context of the current literature. Additionally, an explanation of the contributions and a short discussion of the content and potential future research directions is provided with each manuscript.

Manuscript 1 describes GlobalFit, a novel bi-level optimization algorithm, and its application to genome-scale metabolic models to improve their predictive power. It was published as *Hartleb D, Jarre F, Lercher MJ. Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. PLoS Comput Biol. 2016 Aug 2;12(8).*

Manuscript 2 follows up on the algorithm introduced in *Manuscript 1*. A pipeline is introduced, which employs information from different metabolic sources, and can help to accelerate the reconstruction of high-quality genome-scale metabolic models. This pipeline was successfully applied to three different *Streptococci* strains. *Manuscript 2* was submitted to *PNAS* as *Hartleb D, Lercher MJ. Automated high-quality reconstruction of metabolic networks from high-throughput data.* It was rejected after peer review and is currently under preparation for resubmission.

Manuscript 3 describes a novel approach to detect and remove energy generating cycles in metabolic models. These cycles can severely affect the results obtained by constraint-based methods; they are commonly found in automatically reconstructed metabolic networks. It was published as *Fritzemeier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLoS Comput Biol. 2017 Apr 18;13(4).*

Summary

Genome-scale metabolic models are reconstructed for many organisms. They are routinely used to predict metabolic behavior (O'Brien, Monk, and Palsson 2015), simulate evolutionary adaptation (Pal et al. 2006), and help to design organisms of bioengineering interest (Schirmer et al. 2010). However, the quality of metabolic models is highly variable.

Typically, metabolic models are refined by comparing *in silico* predictions to *in vivo* experiments (e.g., viability of gene knock-outs or growth in different nutritional environments) (Thiele and Palsson 2010). Several different algorithms have been developed to resolve the resulting inconsistencies between prediction and experiment (Satish Kumar, Dasika, and Maranas 2007; Zomorrodi et al. 2012; Thiele, Vlassis, and Fleming 2014; Kumar and Maranas 2009). However, these tools can iteratively correct only one inconsistency at a time. Thus, the total number of network changes may not be globally optimal, a modification introduced earlier might prevent the resolution of other inconsistencies, or a potential solution might not be found because the combination of different types of network changes is not supported.

In *Manuscript 1,* a novel bi-level optimization algorithm – GlobalFit – is introduced (Hartleb, Jarre, and Lercher 2016). GlobalFit is the first algorithm that can simultaneously solve multiple inconsistencies at a time and allows the combination of different network modifications. We applied the algorithm to the genome-scale metabolic model for *Mycoplasma genitalium* (Suthers et al. 2009), improving the overall accuracy for viability predictions from 87.3% to 97.3%.

Interestingly, solving all inconsistencies at a time resulted in the same network changes as iteratively solving each erroneous prediction together with a corresponding counter-case, while the overall time for solving decreased dramatically. Applying this subset strategy to the much better curated genome-scale metabolic model for *Escherichia coli* (Orth et al. 2011), we could again substantially improve the accuracy, from 90.8% to 95.4%.

Reconstructing metabolic models is still a laborious and time-consuming task. To accelerate this process, automatic reconstruction algorithms have been

developed. However, predictions by automatically reconstructed networks generally have low accuracy and still need to be refined manually.

In *Manuscript 2,* a novel pipeline is introduced, which gathers information from metabolic networks from closely related organisms and metabolic databases (i.e., KBase (Knowledgebase 2016), TransportDB (Ren, Chen, and Paulsen 2007), KEGG (Kanehisa et al. 2016)). At each step, the metabolic information of each gene is replaced by newer information. Finally, the draft metabolic network is refined with GlobalFit based on genome-wide gene knock-out data.

We demonstrate the applicability of this pipeline by reconstructing genome-scale metabolic models for three different *Streptococci* genomes. The predictive power of the resulting metabolic models was of the same quality as for manually curated models (e.g., *E. coli* iJO1366 (Orth et al. 2011)).

In addition to the low predictive power of automatically reconstructed metabolic models, they often contain internal energy generating cycles. These cycles can charge energy-rich metabolites such as ATP without the uptake of any nutrients. Thus, they can severely affect the energy metabolism of the model and can unrealistically inflate the maximal biomass production. However, no systematic method to eliminate those cycles had been developed previously.

In *Manuscript 3,* a variant of FBA is described to identify energy generating cycles, and a modified version of GlobalFit is subsequently used to eliminate the detected cycles (Fritzemeier et al. 2017). We could identify energy generating cycles in 65% of metabolic networks from three different databases (BiGG (Schellenberger et al. 2010), MetaNetX (Ganter et al. 2013), and ModelSEED (Henry et al. 2010)). In the following step, GlobalFit could fully eliminate energy generating cycles in 94% of the affected metabolic models.

References

Fritzemeier, C. J., D. Hartleb, B. Szappanos, B. Papp, and M. J. Lercher. 2017. 'Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal', *Plos Computational Biology*, 13: e1005494.

Ganter, M., T. Bernard, S. Moretti, J. Stelling, and M. Pagni. 2013. 'MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks', *Bioinformatics*, 29: 815-6.

Hartleb, D., F. Jarre, and M. J. Lercher. 2016. 'Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets', *Plos Computational Biology*, 12: e1005036.

Henry, C. S., M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. 2010. 'High-throughput generation, optimization and analysis of genome-scale metabolic models', *Nat Biotechnol*, 28: 977-82.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*, 44: D457-62.

Knowledgebase, Department of Energy Systems Biology. 2016. '(KBase)'. http://kbase.us.

Kumar, V. S., and C. D. Maranas. 2009. 'GrowMatch: an automated method for reconciling in silico/in vivo growth predictions', *Plos Computational Biology*, 5: e1000308.

O'Brien, E. J., J. M. Monk, and B. O. Palsson. 2015. 'Using Genome-scale Models to Predict Biological Capabilities', *Cell*, 161: 971-87.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. 2011. 'A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011', *Mol Syst Biol*, 7: 535.

Pal, C., B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver, and L. D. Hurst. 2006. 'Chance and necessity in the evolution of minimal metabolic networks', *Nature*, 440: 667-70.

Ren, Q., K. Chen, and I. T. Paulsen. 2007. 'TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels', *Nucleic Acids Research*, 35: D274-9.

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

Schellenberger, J., J. O. Park, T. M. Conrad, and B. O. Palsson. 2010. 'BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions', *BMC Bioinformatics*, 11: 213.

Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre. 2010. 'Microbial biosynthesis of alkanes', *Science*, 329: 559-62.

Suthers, P. F., M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas. 2009. 'A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189', *Plos Computational Biology*, 5: e1000285.

Thiele, I., and B. O. Palsson. 2010. 'A protocol for generating a high-quality genome-scale metabolic reconstruction', *Nat Protoc*, 5: 93-121.

Thiele, I., N. Vlassis, and R. M. Fleming. 2014. 'fastGapFill: efficient gap filling in metabolic networks', *Bioinformatics*, 30: 2529-31.

Zomorrodi, A. R., P. F. Suthers, S. Ranganathan, and C. D. Maranas. 2012. 'Mathematical optimization applications in metabolic networks', *Metab Eng*, 14: 672-86.

Zusammenfassung

Metabolische Stoffwechselmodelle, die das ganze Genom eines Organismus umfassen, wurden für viele Spezies erstellt. Sie werden regelmäßig benutzt um metabolische Eigenschaften vorherzusagen (O'Brien, Monk, and Palsson 2015), um evolutionäre Anpassung zu simulieren (Pal et al. 2006) und um Organismen mit besonderem biotechnologischen Eigenschaften zu entwerfen (Schirmer et al. 2010). Jedoch schwankt die Qualität von metabolischen Modellen stark.

Typischerweise werden metabolische Stoffwechselmodelle durch den Vergleich von *in silico* Vorhersagen und *in vivo* Experimenten verbessert (z.B.: Lebensfähigkeit nach Knockouts von Genen oder Wachstum auf verschiedenen Nährmedien) (Thiele and Palsson 2010). Es wurden verschiedene Algorithmen entwickelt um die bestehenden Unstimmigkeiten zwischen Vorhersage und Experiment zu beseitigen (Satish Kumar, Dasika, and Maranas 2007; Zomorrodi et al. 2012; Thiele, Vlassis, and Fleming 2014; Kumar and Maranas 2009). Jedoch können diese Methoden iterativ nur jeweils einen Widerspruch auflösen. Daher kann es sein, dass die absolute Anzahl an Netzwerkänderungen nicht global optimal ist. Darüber hinaus kann eine eingebaute Modifikation die Auflösung weiterer Unstimmigkeiten verhindern oder eine mögliche Lösung kann erst gar nicht gefunden werden, da die Kombination von verschiedenen Netzwerkänderungen nicht möglich ist.

In *Manuskript 1* wird ein neuartiger bi-level Optimierungsalgorithmus – GlobalFit – vorgestellt (Hartleb, Jarre, and Lercher 2016). GlobalFit ist der erste Algorithmus, der mehrere Unstimmigkeiten gleichzeitig auflösen kann und die Kombination von verschiedenen Netzwerkmodifikation erlaubt. Wir haben diesen Algorithmus auf das genomumfassende metabolische Netzwerk von *Mycoplasma genitalium* angewendet (Suthers et al. 2009) und konnten dabei die absolute Vorhersagegenauigkeit von 87,3% auf 97,3% verbessern.

Interessanterweise führte das gleichzeitige Korrigieren aller Unstimmigkeiten zu den gleichen Netzwerkmodifikation wie das iterative Korrigieren jeder einzelnen Unstimmigkeit zusammen mit einem entsprechenden Gegenfall. Hierbei sank die absolut benötigte Rechenzeit jedoch drastisch. Durch Anwendung dieser

Teilgruppenstrategie auf das viel besser ausgearbeitete genomumfassende metabolische Netzwerk von *Escherichia coli* (Orth et al. 2011) konnten wir wiederum die Vorhersagegenauigkeit wesentlich verbessern, von 90,8% auf 95,4%.

Das Erstellen von metabolischen Netzwerken ist heutzutage immer noch eine aufwendige und zeitraubende Aufgabe. Um diesen Prozess zu beschleunigen, wurden Algorithmen zur automatischen Netzwerkerstellung entwickelt. Jedoch haben automatisch erstellte Netzwerke häufig eine nur geringe Vorhersagegenauigkeit und bedürfen weiterer manueller Nachbesserungen.

In *Manuskript 2* wird eine neue algorithmische Pipeline vorgestellt, die Informationen von metabolischen Netzwerken nah verwandter Organismen und von metabolischen Datenbanken (KBase (Knowledgebase 2016), TransportDB (Ren, Chen, and Paulsen 2007), KEGG (Kanehisa et al. 2016)) sammelt. In jedem Verarbeitungsschritt werden die metabolischen Informationen von jedem Gen mit neueren Informationen überschrieben. Abschließend wird die Rohfassung des metabolischen Netzwerks mit GlobalFit auf der Basis von genomumfassenden Gen-Knockout-Datensätzen verfeinert.

Wir zeigen die Anwendbarkeit dieser Pipeline durch die Erstellung von genomumfassenden metabolischen Netzwerken für drei verschiedene *Streptococci*-Spezies. Die Vorhersagegenauigkeit der erstellten metabolischen Netzwerke ist vergleichbar mit manuell überarbeiteten metabolischen Netzwerken (z.B.: *E. coli* iJO1366 (Orth et al. 2011)).

Zusätzlich zu der niedrigen Vorhersagekraft automatisch erstellter metabolischer Modelle enthalten diese häufig energiegenerierende Zyklen. Diese internen Zyklen können ohne jegliche Aufnahme von Nährstoffen Energiemetabolite aufladen. Daher können sie den modellierten Energiestoffwechsel des simulierten Organismus erheblich beeinflussen und die maximale Biomasseproduktion unrealistisch erhöhen. Jedoch existiert bisher keine systematische Methode um solche Zyklen zu entfernen.

In *Manuskript 3* wird eine Variante von FBA beschrieben um energieerzeugende Zyklen zu identifizieren; eine modifizierte Version von GlobalFit wird vorgestellt, die anschließend die gefundenen Zyklen entfernt

(Fritzemeier et al. 2017). Wir konnten energieerzeugende Zyklen in 65% der metabolischen Netzwerke aus drei verschiedenen Datenbanken (BiGG (Schellenberger et al. 2010), MetaNetX (Ganter et al. 2013), und ModelSEED (Henry et al. 2010)) finden. Im Anschluss konnte GlobalFit energieerzeugende Zyklen in 94% der betroffenen metabolischen Netzwerke vollständig entfernen.

Literatur

Fritzemeier, C. J., D. Hartleb, B. Szappanos, B. Papp, and M. J. Lercher. 2017. 'Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal', *Plos Computational Biology*, 13: e1005494.

Ganter, M., T. Bernard, S. Moretti, J. Stelling, and M. Pagni. 2013. 'MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks', *Bioinformatics*, 29: 815-6.

Hartleb, D., F. Jarre, and M. J. Lercher. 2016. 'Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets', *Plos Computational Biology*, 12: e1005036.

Henry, C. S., M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. 2010. 'High-throughput generation, optimization and analysis of genome-scale metabolic models', *Nat Biotechnol*, 28: 977-82.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*, 44: D457-62.

Knowledgebase, Department of Energy Systems Biology. 2016. '(KBase)'. http://kbase.us.

Kumar, V. S., and C. D. Maranas. 2009. 'GrowMatch: an automated method for reconciling in silico/in vivo growth predictions', *Plos Computational Biology*, 5: e1000308.

O'Brien, E. J., J. M. Monk, and B. O. Palsson. 2015. 'Using Genome-scale Models to Predict Biological Capabilities', *Cell*, 161: 971-87.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. 2011. 'A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011', *Mol Syst Biol*, 7: 535.

Pal, C., B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver, and L. D. Hurst. 2006. 'Chance and necessity in the evolution of minimal metabolic networks', *Nature*, 440: 667-70.

Ren, Q., K. Chen, and I. T. Paulsen. 2007. 'TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels', *Nucleic Acids Research*, 35: D274-9.

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

Schellenberger, J., J. O. Park, T. M. Conrad, and B. O. Palsson. 2010. 'BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions', *BMC Bioinformatics*, 11: 213.

Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre. 2010. 'Microbial biosynthesis of alkanes', *Science*, 329: 559-62.

Suthers, P. F., M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas. 2009. 'A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189', *Plos Computational Biology*, 5: e1000285.

Thiele, I., and B. O. Palsson. 2010. 'A protocol for generating a high-quality genome-scale metabolic reconstruction', *Nat Protoc*, 5: 93-121.

Thiele, I., N. Vlassis, and R. M. Fleming. 2014. 'fastGapFill: efficient gap filling in metabolic networks', *Bioinformatics*, 30: 2529-31.

Zomorrodi, A. R., P. F. Suthers, S. Ranganathan, and C. D. Maranas. 2012. 'Mathematical optimization applications in metabolic networks', *Metab Eng*, 14: 672-86.

Introduction

5.1 Metabolic Models

Metabolism and its underlying biochemistry have been investigated for centuries, and deep knowledge has been compiled. However, metabolic components were mostly analyzed individually, which made it difficult to characterize cellular function as a system. The development and availability of genome sequencing technologies in recent years changed the focus of biological investigations from a gene-centred view of a few very well studied genes and biological processes to a genome-scale point of view. This allowed the identification of practically all enzyme encoding genes involved in the conversion of metabolites. Combining all metabolic knowledge about an organism with its genomic information led to the construction of metabolic models (O'Brien, Monk, and Palsson 2015; Palsson 2009; Cazzaniga et al. 2014).

While the first metabolic models simulated central carbon metabolism (Fell and Small 1986; van Gulik and Heijnen 1995), the first genome scale metabolic model was reconstructed for *Haemophilus influenzae* in 1999 (Edwards and Palsson 1999). Through new methods and the availability of more experimental data, metabolic models have become increasingly comprehensive and precise. By now, genome scale metabolic models exist not only for bacteria and archaea (Feist et al. 2006), but also for uni- and multicellular eukaryotes, including plants (de Oliveira Dal'Molin et al. 2010), fungi (Liu et al. 2013), and even human (Thiele et al. 2013). The number of included metabolites ranges from 274 for *Mycoplasma genitalium* (Suthers et al. 2009) to 5063 in human (Thiele et al. 2013).

Metabolic models have been successfully applied to a wide scope of biological investigations. For example, genome-scale metabolic models were used to simulate the reductive evolution of the endosymbiont *Buchnera aphidicola* (Pal et al. 2006), and organisms have been successfully engineered to overproduce desired metabolites (strain optimization) (Thakker et al. 2012). In synthetic biology, new enzymes and biosynthesis pathways to enable the production of a new component have been identified through modeling and were subsequently

introduced experimentally (Schirmer et al. 2010). Metabolic models have been used to predict the necessary gene knock-outs to overproduce a desired target metabolite (Burgard, Pharkya, and Maranas 2003; Copeland et al. 2012). Potential drug targets of the opportunistic pathogen *Vibrio vulnificus* were discovered by employing metabolic networks (Kim et al. 2011; Chavali et al. 2012). Host-pathogen interactions of human and *Mycobacterium tuberculosis* have been studied with the help of constraint-based models (Bordbar et al. 2010). With the expanding scope of synthetic biology, it is likely that metabolic models will become increasingly important for the planning of complex genetic interventions.

5.2 Mathematical Representation

Metabolic networks are formally described by a mathematical model. A stoichiometric matrix represents the metabolites (rows) and the biochemical reactions (columns) that constitute the metabolism of the organism of interest. Each metabolite is associated with a compartment, which defines its localization. Every network contains at least two compartments, the cytosol and the extracellular space (Martins Conde Pdo, Sauter, and Pfau 2016; Cazzaniga et al. 2014). The prevalent techniques to analyze metabolic models are constraint-based methods, whereof flux balance analysis (FBA) (Orth, Thiele, and Palsson 2010) is the most popular (Reed 2012). These methods employ constraints on the metabolic model that are derived from simple physical laws.

The metabolic physiological state of an organism can be mathematically described by reaction rates (fluxes) and metabolite concentrations. Reaction rates are governed by complicated, non-linear mathematical functions including metabolite concentrations and enzyme kinetics. It would be extremely challenging to compute these for whole cell models; moreover, many of the required parameters are still not known. Constraint-based models overcome this problem by assuming a steady state condition (i.e., every internal metabolite that is produced must also be consumed at the same rate) (Llaneras and Pico 2008; Durot, Bourguignon, and Schachter 2009).

Each reaction is assigned an upper and lower bound, which limit the maximal and minimal flux that can be carried by the reaction. These bounds can be related to the turnover rate and abundance of the corresponding enzyme. In general, these bounds are unknown and are fixed to a large value (e.g., -1000 $\frac{mmol}{gDW*h}$ for lower bounds and 1000 $\frac{mmol}{gDW*h}$ for upper bounds). The lower bound of reactions that are known to proceed in one direction only (irreversible reactions) are constrained to zero (Kummel, Panke, and Heinemann 2006; Llaneras and Pico 2008).

To allow metabolites to be included or excluded from the outer environment, exchange reactions are added to the metabolic network. These reactions simply convert one metabolite to nothing or vice versa. Setting the lower bounds of exchange reaction to lower than zero allows the uptake of the corresponding

metabolite, thereby simulating specific media compositions. Conversely, metabolites of a metabolic network can be mathematically eliminated by secretion (Durot, Bourguignon, and Schachter 2009).

It can be assumed that naturally occurring metabolic systems have been optimized by natural selection to fulfill a specific function, such as the production of all metabolites needed for cellular growth (biomass). The balanced production of all cell components needed for growth (including DNA, RNA, amino acids, cell wall components, lipids, sterols, essential cofactors, and secondary metabolites) can be formulated as an additional, hypothetical “biomass reaction” (Dreyfuss et al. 2013; O'Brien, Monk, and Palsson 2015). Constraint-based methods maximize this reaction. The maximal biomass production represents the maximal possible yield of biomass production from all available nutrients (Feist and Palsson 2010). Beyond the biomass reaction, other objective functions have been applied successfully with FBA, such as maximizing metabolite production, and minimizing nutrient uptake or ATP production (Llaneras and Pico 2008).

To more faithfully reflect biology, additional constraints can be imposed on metabolic networks. Conventionally, a non-growth associated maintenance reaction (NGAM) is added to the metabolic network, which reflects the ATP consumption needed for maintaining homeostasis independent from growth (e.g., turgor pressure (Feist et al. 2007), right ionic strength (Stouthamer and Bettenhaussen 1973)). The lower bound of this reaction is constrained to a value greater than zero, forcing the metabolic network to consume ATP regardless of biomass production. Additionally, the ATP requirement for growth is included in the biomass reaction (Durot, Bourguignon, and Schachter 2009).

To link genes with reactions, gene-protein-reaction associations (GPR) are included in the metabolic network. For instance, these rules record if a gene functions as an isoenzyme or in a multi-protein complex. GPRs are required to perform *in silico* gene-deletion studies (Zomorrodi et al. 2012).

Many additional constraint-based methods that employ non-linear objective functions have been introduced. For example, regulatory on/off minimization (ROOM) (Shlomi, Berkman, and Ruppin 2005) minimizes the number of regulatory changes, while minimization of metabolic adjustments (MOMA) (Segre, Vitkup, and

Church 2002) minimizes the overall flux change between a wild-type and a mutant strain. ROOM uses mixed-integer linear programming (MILP), while MOMA employs quadratic programming (QP) (Zomorrodi et al. 2012; Copeland et al. 2012).

The simplicity of constraint-based methods allows fast *in silico* computation of genome-scale metabolism, because all variables are linear and thus fast linear programming (LP) optimization techniques can be applied. But this simplicity also leads to a few drawbacks. In general, constraint-based methods ignore the influence of regulation, transcription, metabolite concentration, and enzyme kinetics, each of which can have a huge impact on the metabolism of an organism. Furthermore, metabolic models in general contain more reactions than metabolites. Thus, the mathematical equation system is underdetermined, resulting in a multidimensional solution space of the optimization problem. While the actual objective value in many cases reflects the metabolic capacities of the organism, a single optimal flux distribution is far from unique, making its interpretation difficult (Lewis, Nagarajan, and Palsson 2012; Simeonidis and Price 2015).

A number of different computational environments are available to run constraint-based analyses on metabolic networks (e.g., COBRA (Schellenberger et al. 2011; Ebrahim et al. 2013), SyBiL (Gelius-Dietrich et al. 2013)), which apply commercial (e.g., CPLEX, GUROBI) and non-commercial (e.g., Open Source Gnu Linear Programming Kit (GLPK)) mathematical solvers to find solutions for the corresponding optimization problems.

5.3 Reconstruction process of metabolic models

5.3.1 Stage one of reconstructing metabolic models

The reconstruction process can be divided into four stages. In the first stage, a draft metabolic network is generated based on the genomic content of the target organism. The genome is mapped against a database of known metabolic functions (e.g., KEGG (Kanehisa et al. 2016), MetaCyc (Caspi et al. 2016), or BiGG (Schellenberger et al. 2010)), or metabolic properties of the organism are obtained by its gene annotation (e.g., enzyme commission (EC) numbers or gene ontology classifications (GO) (Gene Ontology 2015)). This first draft network additionally contains all GPR rules associated with genes that were used for generating the metabolic model (Hamilton and Reed 2014; Thiele and Palsson 2010).

5.3.2 Stage two of reconstructing metabolic models

In the second stage, the consistency of the draft metabolic network is evaluated and where necessary curated. Each reaction must be checked for correct mass and charge balance. The GPR association of each reaction must be verified (Hamilton and Reed 2014).

Another part of this stage is to add transport reactions, which allow the uptake and secretion of metabolites and thus allow to simulate specific media compositions on which the target organism is known to grow. Identifying the correct transporters is difficult, because many transporters are highly homologous to each other, and small differences in sequence can change specificity to individual substrates (Cuevas et al. 2016; Marger and Saier 1993).

A crucial step in stage two of metabolic network reconstruction is to infer a suitable biomass reaction. Biomass components and precursors are obtained from the literature or examined experimentally and are quantified in their proportions (Cazzaniga et al. 2014). Additionally, gene knock-out data and growth assays on different growth media can be used to determine all biomass components which are really needed for growth or proliferation (Feist et al. 2009). Finally, the flux through the biomass reaction is scaled to the observed growth rate of the

investigated organism (O'Brien, Monk, and Palsson 2015). Determining an accurate biomass reaction is still a challenging task, because the organism might not have evolved to be in an optimal state and the biomass objective function is likely to be environment-dependent (Yurkovich and Palsson 2016; Feist and Palsson 2010).

Furthermore, the growth and non-growth associated maintenance reactions must be set (ATP is utilized at a non-zero rate even by non-growing cells). These parameters are usually determined by fitting *in silico* growth yields to observed experimental values (Durot, Bourguignon, and Schachter 2009; Reed, Famili, et al. 2006).

5.3.3 Stage three of reconstructing metabolic models

In stage three, the metabolic model is converted into a mathematical model, which is used for further investigations. Systems biology tools for analyzing and simulating metabolic models (e.g., COBRA (Schellenberger et al. 2011) or SybiL (Gelius-Dietrich et al. 2013)) usually can import different metabolic model formats. However, the most common format to store and distribute metabolic models is the Systems Biology Markup language (SBML) (Hucka et al. 2003; Hamilton and Reed 2014).

5.3.4 Stage four of reconstructing metabolic models

In the final stage, the reconstructed metabolic model is validated against experimental datasets to confirm its correct biological behavior and predictive capabilities; e.g., if all experimentally observed products can be secreted by the metabolic model *in silico*, if predicted and observed viability of gene knock-outs agree, and if all biomass precursor can be produced (Cazzaniga et al. 2014).

Furthermore, additional analyses can be performed, such as identifying metabolic dead ends, blocked reactions and gaps in metabolic pathways. The results can point to incomplete parts of the investigated metabolic network (Hamilton and Reed 2014).

Comparing experimental observations and *in silico* predictions generally leads to four cases:

(i) true positive prediction (TPp), i.e., the observation and the prediction both agree that a strain is viable in the tested condition;
(ii) true negative prediction (TNp), i.e., metabolic network and experiment both predict non-growth;
(iii) false positive prediction (FPp), i.e., the metabolic network erroneously predicts viability while the experiment revealed lethality;
(iv) false negative prediction (FNp), i.e., the metabolic network falsely predicts lethality while in the experiment a viable organism was observed.

Stages two to four of the model reconstruction should be repeated until the predictions of the metabolic network reconstruction is in line with experimental observations (Hamilton and Reed 2014; Joyce and Palsson 2008).

5.3.4.1 Measurements of accuracy

The predictive power of a metabolic network for strain viability is usually measured in terms of its accuracy. Accuracy is assessed by dividing the sum of true positive and true negative predictions by the number of all considered predictions.

Alternatively, the Matthews correlation coefficient is a more balanced method to measure accuracy of binary classifications (Matthews 1975); mathematically, it is equivalent to calculating the Pearson's correlation coefficient between two binary vectors. For example, the metabolic network iPS189 (Suthers et al. 2009) of *Mycoplasma genitalium* consists almost only of essential genes. The reported accuracy of this model is 87.3%. Using a trivial model that predicts non-growth for all gene knock-outs would result in a better accuracy of 90.5%; however, the Matthews correlation coefficient would at the same time decrease from 0.56 to zero.

5.3.4.2 Reasons for inaccuracies of *in silico* predictions for *in vivo* behavior

There are many different reasons for inconsistencies between *in silico* predictions and *in vivo* observations. They range from incorrect metabolic models over experimental errors to algorithmic shortcomings:

1) Reactions are also needed for the degradation or recycling of metabolites. Removing one of these reactions does not affect the production of biomass, but violates the steady state condition required by FBA. The rise of concentration of these metabolites *in vivo* may not affect growth, or they are further metabolized or transported out of the cell by different mechanisms (Orth et al. 2011).
2) Isoenzymes or alternative pathways that can carry out the same function are missing, or the reversibility of reactions is not correctly modeled (Orth et al. 2011). Furthermore, it has been shown that enzymes can have unknown low-level side activities. This underground metabolism can contribute to the metabolic capacity of an organism, but is often not sufficiently included in metabolic networks (Notebaart et al. 2014). Additionally, many reactions in a metabolic network miss a gene association. These so-called orphan reactions are needed to allow growth, but the catalyzing enzyme is unknown (Orth and Palsson 2010). 30% to 40% of all known enzymatic functions are estimated to be processed by orphan reactions (Lespinet and Labedan 2006; Orth and Palsson 2012).
3) Metabolites are erroneously included in the biomass reaction. All genes that encode for reactions that are needed to produce or consume a metabolite that is erroneously included in the biomass reaction will then erroneously be deemed essential.
4) Isoenzymes or alternative pathways contained in the model exist *in vivo*, but do not carry sufficient flux *in vivo* (e.g., isoenzymes are not expressed, enzymes are inefficient) (Orth et al. 2011).
5) Reactions are erroneously assumed to be reversible. Directionality of a reaction depends on thermodynamics. In many cases the required parameters (in particular the concentrations of substrate and product) are not known or were measured under different conditions. Thus, many effectively irreversible reactions are labeled as bidirectional because of missing knowledge (Reed 2012).
6) Biomass components that are catalyzed by FPp are not included in the biomass reaction. Genes that encode for reactions that are needed to utilize these components are not essential. The reactions can even be

unconnected to the metabolic network and therefore be unable to carry any flux (blocked) (Tervo and Reed 2013).

7) Genes encode enzymes involved in degrading toxic metabolites. For *S. aureus* and *B. subtilis*, it has been shown that early acting genes of teichoic acid biosynthesis are non-essential, while knock-outs of genes that encode enzymes of later steps in this pathway are lethal. The reason for this counter-intuitive behavior is that deleting a late acting downstream gene will lead to the accumulation of a toxic metabolite further upstream of the pathway. If an early acting gene is removed, the metabolite cannot be produced. Consequently, a double gene knock-out of a downstream and an upstream gene is not lethal (D'Elia, Millar, et al. 2006; D'Elia, Pereira, et al. 2006).
8) Mutants did not have enough time to compensate for the gene deletion. Regulatory changes can restore the organism's capacity to produce sufficient biomass (Herring et al. 2006). For *E. coli*, it has been shown that after changing the growth media it took over 700 generations to achieve the growth yield that was predicted *in silico* (Ibarra, Edwards, and Palsson 2002). Accordingly, some genes which were deemed to be essential based on experiments are in fact unessential after regulatory compensation (O'Brien, Monk, and Palsson 2015).
9) Many microbial organisms are optimized for maximal growth rate. Faster growth can be achieved with faster, but less efficient pathways (Teusink, Bachmann, and Molenaar 2011). FBA calculates the maximum yield per input and not per time. Thus, flux predictions by FBA always use the most efficient pathways (those with the highest biomass yield per limiting nutrient), while the pathways used *in vivo* may have lower yield but allow faster growth (Schilling et al. 1999). For example, *L. plantarum* usually secretes lactate, but is also capable of mixed acid fermentation, which would produce more ATP per glucose. However, this pathway is only used under limited substrate availability. FBA predictions of the genome scale metabolic model for *L. plantarum* utilized the mixed acid fermentation in all circumstances instead of the observed lactate secretion (Teusink et al. 2006).

10) Erroneous classification of genes as essential by the experiments can also complicate the analysis or even make it impossible to reconcile *in silico* predictions with the data (D'Elia, Pereira, and Brown 2009). TN-seq methods (van Opijnen, Bodi, and Camilli 2009) do not only produce mutants that were truly unable to grow in the specified environment, but also mutated organisms that have a fitness disadvantage. Because the mutants are selected *en masse* in a competitive environment (Khatiwara et al. 2012), the less fit mutants (e.g., those with lower yield) can be underrepresented or completely disappear in the results (Le Breton et al. 2015).
11) On the other hand, genes can also be falsely identified as non-essential due to partial gene inactivation. This phenomenon frequently occurs in transposon mediated gene knock-out studies. Genes may not entirely lose their function after an insertion of a transposon, and hence the experiment does not truly represent a full knock-out (Ge and Xu 2012).
12) FBA neglects regulation. Reactions that are used by FBA can be not expressed due to regulation, or a metabolite-enzyme interaction inhibits the functioning of the enzyme (O'Brien, Monk, and Palsson 2015; Durot, Bourguignon, and Schachter 2009).
13) Furthermore, FBA does not consider dilution of metabolites. For example, in *S. cerevisiae* quinones must *in vivo* not only be recycled, but must also be replenished to compensate dilution; in contrast, *in silico* quinones are only recycled. This leads to FPp of the genes involved in quinone biosynthesis (Dreyfuss et al. 2013).
14) Constraint-based models do not consider kinetic parameters, which can significantly influence the rate of conversion of metabolites (Durot, Bourguignon, and Schachter 2009).

5.3.4.3 Algorithms for improving *in silico* predictions

Many of the above-mentioned inaccuracies between *in silico* predictions and *in vivo* observations can only be resolved through manual curation. In particular, erroneous experimental results can only be revealed by expert knowledge.

Nevertheless, algorithms have been developed which try to reconcile *in vivo* / *in silico* inconsistencies.

Several algorithms exist that reconcile FNp. These methods are generally based on gap-filling approaches. Gap-filling algorithms (e.g., Gap-Fill/Gap-Find (Satish Kumar, Dasika, and Maranas 2007), SMILEY (Reed, Patel, et al. 2006)) add reactions from a database of potential reactions (e.g., KEGG (Kanehisa et al. 2016), MetaCyc (Caspi et al. 2016), BiGG (Schellenberger et al. 2010)) or make existing non-reversible reactions reversible to ensure the viability of the metabolic network (i.e., a positive flux through the biomass reaction). Because these algorithms try to find the minimal number of reactions that needs to be added, the underlying mathematical problem is a mixed integer linear optimization problem, which is challenging to solve computationally (Zomorrodi et al. 2012). A second approach to reconcile false negative predictions is to remove metabolites from the biomass reactions (BioMog (Tervo and Reed 2013)), thereby making the genes non-essential that encode the enzymes needed for generating these metabolites.

Additionally, methods for resolving FPp have been developed. The first such approach is to remove reactions from the metabolic model. While this seems superficially similar to gap-filling, the underlying mathematical problem must be stated as a bi-level optimization problem, which has to be reformulated as a single-level optimization problem to be solved efficiently. This reformulation results again in a MILP, but is usually harder to solve than gap-filling methods, because more linear and non-linear binary variables are introduced. So far only one bi-level algorithm has been introduced *(*i.e., GrowMatch (Kumar and Maranas 2009)), with a few further methods derived from this algorithm (Zomorrodi and Maranas 2010; Henry et al. 2009). Moreover, bi-level optimization algorithms have been successfully applied to predict gene knock-outs in a target organism for overproduction of a desired metabolite (e.g., OptKnock (Burgard, Pharkya, and Maranas 2003), OptForce (Ranganathan, Suthers, and Maranas 2010)).

Similar to reconciling false negative predictions, modifying the biomass reaction can also resolve false positive predictions. In these cases, additional metabolites have to be added to the biomass reactions, thereby making the genes essential that encode the enzymes responsible for metabolizing these additional

metabolites. This approach has been successfully applied (Kumar and Maranas 2009; Tervo and Reed 2013).

5.3.5 Contribution of *Manuscript 1* to the refinement of metabolic models

Manuscript 1 presents GlobalFit, a novel bi-level optimization algorithm for refining metabolic models. While the algorithms mentioned in section 5.3.4. are greedy algorithms that only consider one erroneous case at a time, GlobalFit is capable of improving several cases simultaneously. It is thus capable of ensuring the identification of globally optimal solutions in terms of model fit to experimental data, which inspired its name. Simultaneously considering all (or a relevant subset of) growth and non-growth cases simultaneously avoids pitfalls of greedy algorithms. E.g., reconciling false positive predictions may lead to the removal of essential reactions, as removing one of these reactions would trivially lead to a non-growing metabolic network; and by reconciling false negative predictions, a true negative prediction can become a false positive prediction.

For each special type of inconsistency, a different algorithm was previously needed. GlobalFit is the first algorithm that combines several refinement strategies: removals or reversibility changes of existing reactions; additions of reactions to the model; and removals from and additions to the biomass reaction. Thus, GlobalFit allows to identify network modifications that a consecutive application of different refinement algorithms might not find.

We successfully applied GlobalFit to several manually curated metabolic networks. GlobalFit improved the overall accuracy and the Matthews correlation coefficient of the iPS189 (Suthers et al. 2009) metabolic model for *Mycoplasma genitalium* from 87.3% (MCC=0.56) to 97.9% (MCC=0.86). The small size of the metabolic network model allowed us to solve all inconsistencies simultaneously. However, a subset strategy where we solved one inconsistency simultaneously with a contrasting wild-type case resulted in the same network modifications, while requiring much less computing time.

Using this subset strategy on the much larger iJO1366 (Orth et al. 2011) metabolic model for *E. coli*, which is manually curated and represents “the” reference genome-scale metabolic network in systems biology, GlobalFit

enhanced the predictive capability of the iJO1366 metabolic model from 90.8% (MCC=0.67) to 95.4% (MCC=0.84).

5.4 Automated reconstruction

Reconstructing high-quality metabolic models is a laborious and time-consuming task. Automatic reconstruction tools, which execute the four stages described above, have been developed to accelerate the process of generating metabolic models (e.g., ModelSEED/KBase (Henry et al. 2010) (Knowledgebase 2016), RAVEN Toolbox (Agren et al. 2013), Pathway Tools (Karp et al. 2016), PyFBA (Cuevas et al. 2016), or Scrumpy (Poolman 2006)). The resulting metabolic networks generally include more genes and consist of more reactions and metabolites than manually reconstructed models. Partly as a consequence of this, automatically generated models often contain blocked reactions (i.e., reactions that cannot carry any flux) and dead-end metabolites (i.e., metabolites that cannot be produced or consumed). Automated reconstruction methods provide only draft networks, which cannot perform all metabolic capacities of the target organism. Hence, they need further manual refinement. However, with the availability of more manually curated networks and genome sequences that can serve as templates, automated reconstruction tools can generate more accurate metabolic networks and substantially reduce the amount of time needed for a high-quality metabolic model (Brandl and Andersen 2015; Notebaart et al. 2006).

5.4.1 Contribution of *Manuscript 2* to automatically generate more accurate metabolic networks

One of the main limitations of automated reconstruction tools is that they only use gap-filling methods to ensure the viability of the investigated metabolic network. This crucial gap-filling step often adds reactions without supporting evidence (Cuevas et al. 2016). Furthermore, these tools do not allow to automatically employ datasets derived from high-throughput methods (e.g., gene knock-outs, growth data on different media), which are typically utilized during manual model reconstruction. These limitations lead to less accurate genome-scale metabolic models, which require substantial manual curation to reach the prediction capabilities of careful manual reconstructions.

To accelerate the reconstruction of genome-scale metabolic models, we developed a pipeline that employs metabolic data from closely related organisms

and adds information from different metabolic databases (i.e., KBase (Knowledgebase 2016), TransportDB (Ren, Chen, and Paulsen 2007), and KEGG (Kanehisa et al. 2016)). At each step, information on the metabolic function of each gene from the previous step is superseded by newer information (which is deemed more accurate and/or reliable). In the final step, the draft metabolic network is refined with the GlobalFit algorithm introduced in *Manuscript 1* by employing genome-wide gene knock-out and nutritional environment data. Using this pipeline, we reconstructed metabolic networks for three different *Streptococci* strains. The resulting genome-scale metabolic models are of a quality comparable to that of manually curated models. Furthermore, the reconstructed models successfully predict amino acid auxotrophy, growth on different nutritional environments, and potential drug targets.

5.4.2 Energy generating cycles in metabolic network reconstructions

The solution space of constraint-based methods (e.g., FBA) can contain type-II "extreme pathways" or "elementary flux modes" (Sridharan et al. 2015; Wiback and Palsson 2002). These pathways are also called futile cycles and consume energy to drive cycles which consist only of internal metabolites (i.e., no external nutrient is exchanged with the outer environment, while energy is drained from an internal reservoir). Futile cycles are not caused by erroneous constraint-based methods, but have been shown to exist *in vivo* (e.g., organisms that live in an energy rich environment need to dissipate energy (Reidy and Weber 2002; Russell 2007)).

Energy generating cycles (EGCs) can be considered as futile cycles running in reverse. Instead of consuming energy, they are capable of generating energy without the uptake of any external metabolite. Obviously, these cycles are thermodynamically impossible and can massively distort the energy metabolism of the affected metabolic model if they occur in simulations.

5.4.2.1 Contribution of *Manuscript 3* to detect and automatically remove energy generating cycles

So far, no systematic method existed to identify and eliminate EGCs. In *Manuscript 3*, we introduce a variant of FBA to detect such cycles. Subsequently,

we systematically investigate genome-scale metabolic models from three different databases, BiGG (Schellenberger et al. 2010), MetaNetX (Ganter et al. 2013), and ModelSEED (Henry et al. 2010), for the occurrence of EGCs.

Our approach reveals that EGCs are often found in automatically reconstructed models, while only few manually curated networks suffered from EGCs. In many of the identified EGCs, the combination of several transport reactions leads to the build-up of a metabolite gradient (e.g., of H^+ ions), which can be further utilized for the generation of energy-rich metabolites (e.g., by ATP-synthase).

A modified version of the GlobalFit algorithm introduced in *Manuscript 1* was successfully applied to remove EGCs. This version solves a wild-type growth case with the biomass reaction as the objective function simultaneously with a non-growth case with the flux through EGCs as the objective function.

In many cases, GlobalFit first suggested to remove the ATP-synthase. While this network modification successfully removed all EGCs, it also blocked the ability to produce ATP by respiration. In a second run, we did not allow the removal of the ATP-Synthase. Now, typically up to five reactions needed to be removed to eliminate all EGCs. Removing such cycles led to a decreased overall biomass production, typically by about 25%. This observation is not surprising, as EGCs can produce energy without the uptake of any nutrient, thus massively distorting the energy metabolism and erroneously inflating biomass production.

5.5 Outlook

In future applications, GlobalFit cannot only be used to improve metabolic models, but it could be employed to engineer organisms that produce a desired product. GlobalFit can suggest the removal of reactions in a metabolic model that lead to a desired phenotype; these could subsequently be realized by knocking out the corresponding genes in the genome of the organism of interest.

Furthermore, integrating the pipeline for the reconstruction of high-quality metabolic models described in *Manuscript 2* – in particular the utilization of high-throughput gene knock-out or of biolog (Shea et al. 2012) data by GlobalFit – , and the procedures for EGC detection and removal described in *Manuscript 3* into automatic reconstructing tools (Henry et al. 2010; Overbeek et al. 2005; Latendresse et al. 2012; Cuevas et al. 2016) would have the potential not only to drastically accelerate the reconstruction process without extensive manual curation, but also might help to generate more reliable genome-scale metabolic models in the future.

5.6 Theses

- Previous algorithms for improving the predictive power of genome-scale metabolic networks only considered one case at a time, and for each kind of modification (i.e., addition, removal, reversibility changes of reactions; modification of the biomass reaction) a different algorithm had to be used. These limitations can lead to network changes that are not globally minimal, or the network modifications of a case solved earlier might prohibit the solution of a subsequently considered case. Solving only one case where the metabolic network should not grow requires the exclusion of the removal of essential reactions. This restricts possible network refinement steps that combine the removal of an essential reaction with the addition of a reaction to the metabolic network.

- Reconstructing genome-scale metabolic networks is still a major bottleneck of constraint-based modeling. Automatic reconstructing tools have been developed to accelerate this process. However, automatically reconstructed metabolic models often have a low predictive power. One limitation of these tools is that they only consider gap-filling methods and hence only growth cases. By employing growth and non-growth data, more accurate and reliable genome-scale metabolic models can be reconstructed automatically.

- Energy generating cycles can charge energy containing metabolites without the uptake of any nutrients. Thus, EGCs can have a huge effect on the *in silico* metabolism of metabolic models, and the elimination of EGCs is crucial for the proper functioning of the modeled energy metabolism.

5.7 References

Agren, R., L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. 2013. 'The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum', *Plos Computational Biology*, 9: e1002980.

Bordbar, A., N. E. Lewis, J. Schellenberger, B. O. Palsson, and N. Jamshidi. 2010. 'Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions', *Mol Syst Biol*, 6: 422.

Brandl, J., and M. R. Andersen. 2015. 'Current state of genome-scale modeling in filamentous fungi', *Biotechnol Lett*, 37: 1131-9.

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

Caspi, R., R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, and P. D. Karp. 2016. 'The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases', *Nucleic Acids Research*, 44: D471-80.

Cazzaniga, P., C. Damiani, D. Besozzi, R. Colombo, M. S. Nobile, D. Gaglio, D. Pescini, S. Molinari, G. Mauri, L. Alberghina, and M. Vanoni. 2014. 'Computational strategies for a system-level understanding of metabolism', *Metabolites*, 4: 1034-87.

Chavali, A. K., K. M. D'Auria, E. L. Hewlett, R. D. Pearson, and J. A. Papin. 2012. 'A metabolic network approach for the identification and prioritization of antimicrobial drug targets', *Trends Microbiol*, 20: 113-23.

Copeland, W. B., B. A. Bartley, D. Chandran, M. Galdzicki, K. H. Kim, S. C. Sleight, C. D. Maranas, and H. M. Sauro. 2012. 'Computational tools for metabolic engineering', *Metab Eng*, 14: 270-80.

Cuevas, D. A., J. Edirisinghe, C. S. Henry, R. Overbeek, T. G. O'Connell, and R. A. Edwards. 2016. 'From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model', *Frontiers in Microbiology*, 7: 907.

D'Elia, M. A., K. E. Millar, T. J. Beveridge, and E. D. Brown. 2006. 'Wall teichoic acid polymers are dispensable for cell viability in Bacillus subtilis', *J Bacteriol*, 188: 8313-6.

D'Elia, M. A., M. P. Pereira, and E. D. Brown. 2009. 'Are essential genes really essential?', *Trends Microbiol*, 17: 433-8.

D'Elia, M. A., M. P. Pereira, Y. S. Chung, W. Zhao, A. Chau, T. J. Kenney, M. C. Sulavik, T. A. Black, and E. D. Brown. 2006. 'Lesions in teichoic acid biosynthesis in Staphylococcus aureus lead to a lethal gain of function in the otherwise dispensable pathway', *J Bacteriol*, 188: 4183-9.

de Oliveira Dal'Molin, C. G., L. E. Quek, R. W. Palfreyman, S. M. Brumbley, and L. K. Nielsen. 2010. 'AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis', *Plant Physiol*, 152: 579-89.

Dreyfuss, J. M., J. D. Zucker, H. M. Hood, L. R. Ocasio, M. S. Sachs, and J. E. Galagan. 2013. 'Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus Neurospora crassa using FARM', *Plos Computational Biology*, 9: e1003126.

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

Ebrahim, A., J. A. Lerman, B. O. Palsson, and D. R. Hyduke. 2013. 'COBRApy: COnstraints-Based Reconstruction and Analysis for Python', *BMC Syst Biol*, 7: 74.

Edwards, J. S., and B. O. Palsson. 1999. 'Systems properties of the Haemophilus influenzae Rd metabolic genotype', *J Biol Chem*, 274: 17410-6.

Feist, A. M., C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson. 2007. 'A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information', *Mol Syst Biol*, 3: 121.

Feist, A. M., M. J. Herrgard, I. Thiele, J. L. Reed, and B. O. Palsson. 2009. 'Reconstruction of biochemical networks in microorganisms', *Nat Rev Microbiol*, 7: 129-43.

Feist, A. M., and B. O. Palsson. 2010. 'The biomass objective function', *Curr Opin Microbiol*, 13: 344-9.

Feist, A. M., J. C. Scholten, B. O. Palsson, F. J. Brockman, and T. Ideker. 2006. 'Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri', *Mol Syst Biol*, 2: 2006 0004.

Fell, D. A., and J. R. Small. 1986. 'Fat synthesis in adipose tissue. An examination of stoichiometric constraints', *Biochem J*, 238: 781-6.

Ganter, M., T. Bernard, S. Moretti, J. Stelling, and M. Pagni. 2013. 'MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks', *Bioinformatics*, 29: 815-6.

Ge, X., and P. Xu. 2012. 'Genome-wide gene deletions in Streptococcus sanguinis by high throughput PCR', *J Vis Exp*.

Gelius-Dietrich, G., A. A. Desouki, C. J. Fritzemeier, and M. J. Lercher. 2013. 'Sybil--efficient constraint-based modelling in R', *BMC Syst Biol*, 7: 125.

Gene Ontology, Consortium. 2015. 'Gene Ontology Consortium: going forward', *Nucleic Acids Research*, 43: D1049-56.

Hamilton, J. J., and J. L. Reed. 2014. 'Software platforms to facilitate reconstructing genome-scale metabolic networks', *Environ Microbiol*, 16: 49-59.

Henry, C. S., M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. 2010. 'High-throughput generation, optimization and analysis of genome-scale metabolic models', *Nat Biotechnol*, 28: 977-82.

Henry, C. S., J. F. Zinner, M. P. Cohoon, and R. L. Stevens. 2009. 'iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations', *Genome Biol*, 10: R69.

Herring, C. D., A. Raghunathan, C. Honisch, T. Patel, M. K. Applebee, A. R. Joyce, T. J. Albert, F. R. Blattner, D. van den Boom, C. R. Cantor, and B. O. Palsson. 2006. 'Comparative genome sequencing of Escherichia coli allows observation of bacterial evolution on a laboratory timescale', *Nat Genet*, 38: 1406-12.

Hucka, M., A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, Goryanin, II, W. J. Hedley, T. C. Hodgman, J. H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and Sbml Forum. 2003. 'The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models', *Bioinformatics*, 19: 524-31.

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

Joyce, A. R., and B. O. Palsson. 2008. 'Predicting gene essentiality using genome-scale in silico models', *Methods Mol Biol*, 416: 433-57.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. 2016. 'KEGG as a reference resource for gene and protein annotation', *Nucleic Acids Research*, 44: D457-62.

Karp, P. D., M. Latendresse, S. M. Paley, M. Krummenacker, Q. D. Ong, R. Billington, A. Kothari, D. Weaver, T. Lee, P. Subhraveti, A. Spaulding, C. Fulcher, I. M. Keseler, and R. Caspi. 2016. 'Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology', *Brief Bioinform*, 17: 877-90.

Khatiwara, A., T. Jiang, S. S. Sung, T. Dawoud, J. N. Kim, D. Bhattacharya, H. B. Kim, S. C. Ricke, and Y. M. Kwon. 2012. 'Genome scanning for conditionally essential genes in Salmonella enterica Serotype Typhimurium', *Appl Environ Microbiol*, 78: 3098-107.

Kim, H. U., S. Y. Kim, H. Jeong, T. Y. Kim, J. J. Kim, H. E. Choy, K. Y. Yi, J. H. Rhee, and S. Y. Lee. 2011. 'Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery', *Mol Syst Biol*, 7: 460.

Knowledgebase, Department of Energy Systems Biology. 2016. '(KBase)'. http://kbase.us.

Kumar, V. S., and C. D. Maranas. 2009. 'GrowMatch: an automated method for reconciling in silico/in vivo growth predictions', *Plos Computational Biology*, 5: e1000308.

Kummel, A., S. Panke, and M. Heinemann. 2006. 'Systematic assignment of thermodynamic constraints in metabolic network models', *BMC Bioinformatics*, 7: 512.

Latendresse, M., M. Krummenacker, M. Trupp, and P. D. Karp. 2012. 'Construction and completion of flux balance models from pathway databases', *Bioinformatics*, 28: 388-96.

Le Breton, Y., A. T. Belew, K. M. Valdes, E. Islam, P. Curry, H. Tettelin, M. E. Shirtliff, N. M. El-Sayed, and K. S. McIver. 2015. 'Essential Genes in the Core Genome of the Human Pathogen Streptococcus pyogenes', *Sci Rep*, 5: 9838.

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

Lewis, N. E., H. Nagarajan, and B. O. Palsson. 2012. 'Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods', *Nat Rev Microbiol*, 10: 291-305.

Liu, J., Q. Gao, N. Xu, and L. Liu. 2013. 'Genome-scale reconstruction and in silico analysis of Aspergillus terreus metabolism', *Mol Biosyst*, 9: 1939-48.

Llaneras, F., and J. Pico. 2008. 'Stoichiometric modelling of cell metabolism', *J Biosci Bioeng*, 105: 1-11.

Marger, Michael D., and Milton H. Saier. 1993. 'A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport', *Trends in Biochemical Sciences*, 18: 13-20.

Martins Conde Pdo, R., T. Sauter, and T. Pfau. 2016. 'Constraint Based Modeling Going Multicellular', *Front Mol Biosci*, 3: 3.

Matthews, B. W. 1975. 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme', *Biochim Biophys Acta*, 405: 442-51.

Notebaart, R. A., B. Szappanos, B. Kintses, F. Pal, A. Gyorkei, B. Bogos, V. Lazar, R. Spohn, B. Csorgo, A. Wagner, E. Ruppin, C. Pal, and B. Papp. 2014. 'Network-level architecture and the evolutionary potential of underground metabolism', *Proc Natl Acad Sci U S A*, 111: 11762-7.

Notebaart, R. A., F. H. van Enckevort, C. Francke, R. J. Siezen, and B. Teusink. 2006. 'Accelerating the reconstruction of genome-scale metabolic networks', *BMC Bioinformatics*, 7: 296.

O'Brien, E. J., J. M. Monk, and B. O. Palsson. 2015. 'Using Genome-scale Models to Predict Biological Capabilities', *Cell*, 161: 971-87.

Orth, J. D., T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson. 2011. 'A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011', *Mol Syst Biol*, 7: 535.

Orth, J. D., and B. Palsson. 2012. 'Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions', *BMC Syst Biol*, 6: 30.

Orth, J. D., and B. O. Palsson. 2010. 'Systematizing the generation of missing metabolic knowledge', *Biotechnol Bioeng*, 107: 403-12.

Orth, J. D., I. Thiele, and B. O. Palsson. 2010. 'What is flux balance analysis?', *Nat Biotechnol*, 28: 245-8.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. 2005. 'The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes', *Nucleic Acids Research*, 33: 5691-702.

Pal, C., B. Papp, M. J. Lercher, P. Csermely, S. G. Oliver, and L. D. Hurst. 2006. 'Chance and necessity in the evolution of minimal metabolic networks', *Nature*, 440: 667-70.

Palsson, B. 2009. 'Metabolic systems biology', *FEBS Lett*, 583: 3900-4.

Poolman, M. G. 2006. 'ScrumPy: metabolic modelling with Python', *Syst Biol (Stevenage)*, 153: 375-8.

Ranganathan, S., P. F. Suthers, and C. D. Maranas. 2010. 'OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions', *Plos Computational Biology*, 6: e1000744.

Reed, J. L. 2012. 'Shrinking the metabolic solution space using experimental datasets', *Plos Computational Biology*, 8: e1002662.

Reed, J. L., I. Famili, I. Thiele, and B. O. Palsson. 2006. 'Towards multidimensional genome annotation', *Nat Rev Genet*, 7: 130-41.

Reed, J. L., T. R. Patel, K. H. Chen, A. R. Joyce, M. K. Applebee, C. D. Herring, O. T. Bui, E. M. Knight, S. S. Fong, and B. O. Palsson. 2006. 'Systems approach to refining genome annotation', *Proc Natl Acad Sci U S A*, 103: 17480-4.

Reidy, S. P., and J. M. Weber. 2002. 'Accelerated substrate cycling: a new energy-wasting role for leptin in vivo', *Am J Physiol Endocrinol Metab*, 282: E312-7.

Ren, Q., K. Chen, and I. T. Paulsen. 2007. 'TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels', *Nucleic Acids Research*, 35: D274-9.

Russell, J. B. 2007. 'The energy spilling reactions of bacteria and other organisms', *J Mol Microbiol Biotechnol*, 13: 1-11.

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

Schellenberger, J., J. O. Park, T. M. Conrad, and B. O. Palsson. 2010. 'BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions', *BMC Bioinformatics*, 11: 213.

Schellenberger, J., R. Que, R. M. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and B. O. Palsson. 2011. 'Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0', *Nat Protoc*, 6: 1290-307.

Schilling, C. H., S. Schuster, B. O. Palsson, and R. Heinrich. 1999. 'Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era', *Biotechnol Prog*, 15: 296-303.

Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre. 2010. 'Microbial biosynthesis of alkanes', *Science*, 329: 559-62.

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

Shea, A., M. Wolcott, S. Daefler, and D. A. Rozak. 2012. 'Biolog phenotype microarrays', *Methods Mol Biol*, 881: 331-73.

Shlomi, T., O. Berkman, and E. Ruppin. 2005. 'Regulatory on/off minimization of metabolic flux changes after genetic perturbations', *Proc Natl Acad Sci U S A*, 102: 7695-700.

Simeonidis, E., and N. D. Price. 2015. 'Genome-scale modeling for metabolic engineering', *J Ind Microbiol Biotechnol*, 42: 327-38.

Sridharan, G. V., E. Ullah, S. Hassoun, and K. Lee. 2015. 'Discovery of substrate cycles in large scale metabolic networks using hierarchical modularity', *BMC Syst Biol*, 9: 5.

Stouthamer, A. H., and C. Bettenhaussen. 1973. 'Utilization of energy for growth and maintenance in continuous and batch cultures of microorganisms. A reevaluation of the method for the determination of ATP production by measuring molar growth yields', *Biochim Biophys Acta*, 301: 53-70.

Suthers, P. F., M. S. Dasika, V. S. Kumar, G. Denisov, J. I. Glass, and C. D. Maranas. 2009. 'A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189', *Plos Computational Biology*, 5: e1000285.

Tervo, C. J., and J. L. Reed. 2013. 'BioMog: a computational framework for the de novo generation or modification of essential biomass components', *PLoS One*, 8: e81322.

Teusink, B., H. Bachmann, and D. Molenaar. 2011. 'Systems biology of lactic acid bacteria: a critical review', *Microb Cell Fact*, 10 Suppl 1: S11.

Teusink, B., A. Wiersma, D. Molenaar, C. Francke, W. M. de Vos, R. J. Siezen, and E. J. Smid. 2006. 'Analysis of growth of Lactobacillus plantarum WCFS1

on a complex medium using a genome-scale metabolic model', *J Biol Chem*, 281: 40041-8.

Thakker, C., I. Martinez, K. Y. San, and G. N. Bennett. 2012. 'Succinate production in Escherichia coli', *Biotechnol J*, 7: 213-24.

Thiele, I., and B. O. Palsson. 2010. 'A protocol for generating a high-quality genome-scale metabolic reconstruction', *Nat Protoc*, 5: 93-121.

Thiele, I., N. Swainston, R. M. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bolling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novere, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, Sr., M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. O. Palsson. 2013. 'A community-driven global reconstruction of human metabolism', *Nat Biotechnol*, 31: 419-25.

van Gulik, W. M., and J. J. Heijnen. 1995. 'A metabolic network stoichiometry analysis of microbial growth and product formation', *Biotechnol Bioeng*, 48: 681-98.

van Opijnen, T., K. L. Bodi, and A. Camilli. 2009. 'Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms', *Nat Methods*, 6: 767-72.

Wiback, S. J., and B. O. Palsson. 2002. 'Extreme pathway analysis of human red blood cell metabolism', *Biophysical Journal*, 83: 808-18.

Yurkovich, James T., and Bernhard O. Palsson. 2016. 'Solving Puzzles With Missing Pieces: The Power of Systems Biology', *Proceedings of the Ieee*, 104: 2-7.

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

Zomorrodi, A. R., P. F. Suthers, S. Ranganathan, and C. D. Maranas. 2012. 'Mathematical optimization applications in metabolic networks', *Metab Eng*, 14: 672-86.

Manuscripts

6.1 *Manuscript 1*: Improved Metabolic Models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets

6.1.1 Details

Authors:	Daniel Hartleb, Florian Jarre, Martin J. Lercher
Authorship:	1. Author
Journal:	PLOS Computational Biology
Impact Factor:	4.62
Status:	Published: PLoS Comput Biol 12(8): e1005036, DOI: 10.1371/journal.pcbi.1005036

6.1.2 Contributions

DH conceived and designed the experiments, later refined through discussion with FJ and MJL. DH performed the experiments, developed the software, and analyzed the data. DH wrote a first draft of the manuscript, which was then iteratively refined in collaboration with FJ and MJL.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Improved Metabolic Models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets

Daniel Hartleb[1], Florian Jarre[2], Martin J. Lercher[1]*

1 Institute for Computer Science and Cluster of Excellence on Plant Sciences, Heinrich Heine University, Düsseldorf, Germany, **2** Institute for Mathematics, Heinrich Heine University, Düsseldorf, Germany

* lercher@cs.uni-duesseldorf.de

OPEN ACCESS

Citation: Hartleb D, Jarre F, Lercher MJ (2016) Improved Metabolic Models for *E. coli* and *Mycoplasma genitalium* from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. PLoS Comput Biol 12(8): e1005036. doi:10.1371/journal.pcbi.1005036

Editor: Kiran Raosaheb Patil, EMBL-Heidelberg, GERMANY

Received: November 27, 2015

Accepted: June 27, 2016

Published: August 2, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: We acknowledge financial support through the German Research Foundation DFG (IRTG 152). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Constraint-based metabolic modeling methods such as Flux Balance Analysis (FBA) are routinely used to predict the effects of genetic changes and to design strains with desired metabolic properties. The major bottleneck in modeling genome-scale metabolic systems is the establishment and manual curation of reliable stoichiometric models. Initial reconstructions are typically refined through comparisons to experimental growth data from gene knockouts or nutrient environments. Existing methods iteratively correct one erroneous model prediction at a time, resulting in accumulating network changes that are often not globally optimal. We present GlobalFit, a bi-level optimization method that finds a globally optimal network, by identifying the minimal set of network changes needed to correctly predict all experimentally observed growth and non-growth cases simultaneously. When applied to the genome-scale metabolic model of *Mycoplasma genitalium*, GlobalFit decreases unexplained gene knockout phenotypes by 79%, increasing accuracy from 87.3% (according to the current state-of-the-art) to 97.3%. While currently available computers do not allow a global optimization of the much larger metabolic network of *E. coli*, the main strengths of GlobalFit are already played out when considering only one growth and one non-growth case simultaneously. Application of a corresponding strategy halves the number of unexplained cases for the already highly curated *E. coli* model, increasing accuracy from 90.8% to 95.4%.

Author Summary

Mathematical models that aim to describe the complete metabolism of a cell help us understand cellular metabolic capabilities and evolution, and aid the biotechnological design of microbial strains with desired properties. Draft models are frequently improved through adjustments that increase the agreement of growth/non-growth predictions with observations from gene knockout experiments. Automated methods for this task typically

PLOS | COMPUTATIONAL BIOLOGY

correct one erroneous prediction after the other. We present GlobalFit, a novel method that can consider all experiments and all possible changes simultaneously to identify model modifications that are globally optimal (i.e., that correct the largest possible number of wrong predictions while introducing sets of changes that are most compatible with existing knowledge). This becomes computationally very hard when considering large metabolic models; however, a reduced application of GlobalFit that only looks at small subsets of experiments simultaneously works very well in practice. Allowing only changes that are conservative (e.g., introducing new reactions only if supported by significant genomic evidence), GlobalFit halves the number of wrong growth/non-growth predictions for the state-of-the-art metabolic models of *E. coli* and *Mycoplasma genitalium*, increasing prediction accuracy to 95.4% and 93.0%, respectively. By additionally allowing less conservative changes, we are able to improve accuracy further to 97.3% for the M. genitalium model.

Introduction

Metabolism is the best understood large cellular system. Genome-scale metabolic models that largely rely on constraints for mass balance (i.e., all internal metabolites that are produced must also be consumed) are routinely applied to predict a wide range of metabolic phenomena [1]. The most widely-used of these constraint-based methods, Flux Balance Analysis (FBA), has been successfully applied to predict a range of biological phenomena such as gene knockout effects [1] and the evolutionary adaptation of microbial strains [2–4], and has been employed to predict drug targets [5] and to design microbial strains for bioengineering [6].

Network models are reconstructed by supplementing genomic annotation with information from biochemical characterizations and the organism-specific literature [7]. The resulting draft reconstructions often contain gaps: the modeled organism or its gene knockout strain can grow *in vivo*, while the model is unable to produce biomass *in silico* in the same metabolic environment (false-negative predictions, FNp). Gap filling methods have been introduced to resolve individual FNp through a minimal number of network changes, making irreversible reactions reversible or adding reactions from a database [8–11].

A second type of inconsistencies is the erroneous prediction of growth where the experiment shows no growth (false-positive predictions, FPp). Such cases can be rectified by deleting reactions, making reversible reactions unidirectional, or adding metabolites to the biomass (all reactions necessary for the production of a given metabolite become essential once this metabolite is added to the biomass). GrowMatch [12], the current state-of-the-art in automatic network refinement, uses bi-level optimization to identify reactions that must be deleted or modified for each FPp. GrowMatch also allows to add to the biomass products and/or substrates of reactions that are experimentally essential but are blocked in the model [12].

All currently available methods for network refinement based on growth data are greedy algorithms, solving one inconsistency between model and experiment at a time [8–15]. While each individual set of network changes is minimal, the union of these sets can become larger than a minimal set of changes that solves all inconsistencies simultaneously. Reactions considered essential or model changes introduced early may make the reconciliation of FNp or FPp considered later impossible (for an example, see our application to *Mycoplasma genitalium* below). Furthermore, experimental errors that happen to be consistent with the initial model can severely bias the results. Moreover, previous methods only alter the biomass equation

independently of other network modifications [12, 16] and may miss solutions that combine biomass and network changes.

Results

An algorithm to find global rather than local optima when resolving inconsistencies

We present GlobalFit, a novel bi-level optimization method capable of comparing flux-balance analysis (FBA) [17] model predictions to measured growth across all tested environments and gene knockouts (or subsets thereof) simultaneously. Allowed model changes are (i) removals or (ii) reversibility changes of existing reactions; (iii) additions of reactions to the model from a database of potential reactions; (iv) removals of metabolites from the biomass; and (v) additions of metabolites to the biomass. GlobalFit does not change gene-protein-reaction associations (GPRs), and thus isoenzymes should be identified and included in the model as a preprocessing step.

The algorithm is first formulated as a bi-level linear problem, where each condition is represented by separate metabolites and fluxes (see the detailed method description in Methods). To ensure *in silico* growth for conditions with experimentally demonstrated growth, the biomass production for these conditions must be greater than a predefined threshold. For non-growth phenotypes, the inner optimization problem maximizes the biomass production to check whether it stays below a non-growth threshold. The outer optimization problem jointly minimizes the number of model changes and the number of experiments that are incorrectly predicted by the final model.

The penalties for individual network changes can be set independently. This allows, for example, to prefer reversibility changes over reaction additions, to preferentially remove reactions not associated with a gene, or to preferentially include additional reactions from metabolic network reconstructions of close relatives (see some suggestions for setting these penalties in the S1 Table). The bi-level problem can be re-formulated as a single-level optimization problem [18]; a corresponding implementation of GlobalFit, integrated with the *sybil* toolbox for constraint-based analyses [19], is freely available from CRAN (http://cran.r-project.org/web/packages/GlobalFit/).

While GlobalFit is designed to find globally optimal network modifications by considering all experimental data simultaneously, the corresponding MILP problem rapidly becomes prohibitively large when considering high-throughput gene knockout data. For example, simultaneously considering all possible 1366 *E. coli* knockouts [20] with 4000 allowed network modifications would result in a matrix with 13 million columns by 37 million rows, a problem size not addressable with current computing infrastructures.

However, when searching for model changes that rectify a FPp, trivial but unhelpful solutions such as the deletion of essential reactions are already avoided by simultaneously requiring growth in one or more specified true positive cases. When searching for model changes that rectify a FNp, overly generous changes (such as the removal of metabolites from the biomass) are avoided by simultaneously requiring non-growth in one or more specified true negative cases. Thus, while a globally optimal solution is only guaranteed when simultaneously considering all experimental growth data, a good approximation may be found by solving subsets of inconsistencies. We explore this "subset strategy" below in our application to the *E. coli* genome-scale model. We suggest contrasting each individual FPp with a wild-type growth case (or, if growth was assayed on different media, with a small set of wild-type growth cases). FNp may first be solved alone. However, if a suggested solution for a FNp or a FPp converts other previously correct predictions to false predictions (TPp to FNp or TNp to FPp), the originally

considered case should be solved again, this time contrasting it with the complete set of these conflicting cases. This last step must be repeated until no more additional false predictions occur (or until no solution is found).

The runtime of MILP solvers depends crucially on the number of binary variables. Importantly, this number depends only on the number of allowed changes (plus a single binary variable for the inclusion/exclusion of each growth/non-growth case). Thus, a MILP strategy that considers n possible model changes for a single growth/non-growth case solves a problem with n binary variables. In comparison, the number of binary variables in a GlobalFit run that considers n possible model changes and contrasts m growth and non-growth cases is $n+m$. The number of binary variables can be further reduced by a set of preprocessing steps (Methods).

When reconciling a metabolic network with experimental data, the most parsimonious network modifications are not always those that best describe the true metabolic system. GlobalFit can also provide a specified number of alternative optimal or sub-optimal solutions (using the integer cut method). Thus, users can choose the solution(s) that best agree with available evidence, or design additional experiments that distinguish between competing network modifications. In cases where all suggested alternatives appear excessive or unrealistic, users may also consider modifying individual GPR rules. The runtime for n alternative solutions is approximately n times the runtime for a single optimum. In the test cases reported below, we only examined a small range of alternative solutions and did not consider manual modifications.

Test case 1: Improving the iPS189 metabolic model for Mycoplasma genitalium

We first applied GlobalFit to the genome-scale metabolic network of *Mycoplasma genitalium* [21], using the same gene knockout essentiality data [22] as the initial reconstruction with GrowMatch (reported by [21] to have a global accuracy of 87.3%, corresponding to a Matthew's correlation coefficient, a more balanced measure of classification quality [23], of MCC = 0.56; Table 1). The growth medium used for the knockout experiments was chemically undefined [22]. When applying GlobalFit, we thus allowed the uptake of all nutrients for which transport reactions are included in the model. All other FBA parameters were set to the values used in [21]. The initial network obtained from [21] was not able to produce biomass; to rectify this problem, we had to convert three irreversible reactions (*ZN2t4,INSK,LYSt3*) to reversible reactions. With these modifications, the original model [21] has an accuracy of 85% and a Matthews' correlation coefficient MCC = 0.44. False predictions mainly occurred in the form of FPp, i.e., by incorrectly establishing growth *in silico* where a lethal phenotype was observed *in vivo* (Table 1).

To construct a database of potential additional reactions, we started from all reactions contained in metabolic networks provided by the BiGG database [24]. We removed globally blocked reactions, *i.e.*, those reactions of the database that were not able to carry any flux in a supernetwork containing all reactions. Reversible reactions were represented as two independent irreversible reactions, corresponding to forward and backward directions. The database is provided as S2 Database of the supplementary material.

In our first analysis, we used a very restrictive, conservative set of potential network changes: (i) addition of reactions from other network reconstructions that are catalyzed by enzymes with significant sequence similarity to the *M. genitalium* genome (BLAST e-value $<10^{-13}$); (ii) conversion of irreversible to reversible reactions for reactions that are at least classified as reversible with uncertainty in the *E.* coli model [25]; (iii) removal of reactions (separately for individual reaction directions for reversible reactions); (iv) removal of biomass components;

Table 1. Comparison of experimental and predicted viability for 187 M. genitalium gene knockouts.

	Experiment			
Predictions	**growth**	**non-growth**	**Accuracy**	**MCC**
GrowMatch (reported in [21])[1]				
growth	16	22	87.3%	0.56
no growth	2	149		
Unoptimized model[2]				
growth	12	24	85.0%	0.44
no growth	4	147		
GlobalFit, conservative				
growth	14	10	93.6%	0.68
no growth	2	161		
GlobalFit, non-conservative				
growth	14	2	97.9%	0.86
no growth	2	169		

[1] These numbers include the two genes wrongly associated with the FBA model (MG260, MG124) removed in our calculations.

[2] The initial network obtained from [21] was not able to produce biomass in any environment; to rectify this problem, we converted three irreversible reactions (ZN2t4, INSK, LYSt3) to reversible reactions. We further allowed uptake of all metabolites for which transport reactions are included (see Methods).

doi:10.1371/journal.pcbi.1005036.t001

and (v) addition of biomass components that occur in the biomass of other network reconstructions [16, 20, 24]. In this application, we assigned the same penalty (1.0) for all changes. However, as the growth medium used in the knockout experiments was undefined, we assigned a lower penalty (0.1) for the removal of exchange reactions. Thus, removal of a metabolite from the representation of the undefined medium (corresponding to the removal of an exchange reaction) was preferred to the removal of the corresponding transporter.

Solving false positive predictions (FPp). 14 out of 24 FPp could be transformed to true negatives (Tables 1 and 2), resulting in a specificity of 93.6%. Of the ten reactions that were suggested for removal, four were exchange reactions (for uracil, fructose, glycerol, and dATP), indicating the absence of these substrates from the undefined growth medium [22]; this alone solved a total of eight FPp. In each case, an alternative (though less parsimonious) solution would be the removal of the corresponding transport reaction (note, however, that the transport reactions for uracil and dATP have no associated gene).

Four of the remaining six reactions indicated for removal (NDPK1, NDPK8, NDPK9, PGAMT) were not associated with a gene; i.e., they had an empty gene-protein-reaction association (GPR). A fifth reaction, G3PD4, is associated with the gene MG260; however, this association is likely erroneous. G3PD4 is catalyzed by a glycerol-3-phosphate dehydrogenase (1.1.5.3), whereas MG260 is a lipoprotein without significant sequence similarity to any proteins with known catalytic functions. Thus, GlobalFit suggests the removal of only one reaction (URIK1) that is reliably associated with a gene.

GlobalFit finds no network modification that predicts the lethality of MG124 knockouts. The gene MG124 encodes a thioreductase (THDPO) that is presumably used by Mycoplasma to protect itself from the consequences of self-generated oxidative challenges [26]. Its metabolic function is thus to regulate metabolite concentrations and cannot be captured in FBA models.

The remaining three solved FPp cases were corrected by simultaneously adding one reaction (ACGAMPM) and removing another (PGAMT). Without PGAMT, ACGAMPM is the only reaction producing N-Acetyl-D-glucosamine 1-phosphate, a precursor of the biomass metabolites teichuronic acid and minor teichoic acid (Fig 1). ACGAMPM is associated with three

Table 2. Modifications of the *M. genitalium* network suggested by GlobalFit based on 187 gene knockout experiments (bold font indicates conservative changes).

Type	Gene	Associated reactions	Removed reactions	Added reactions	Added biomass metabolite
FPp	**MG030**	UPPRT	NDPK1^{for}, NDPK9^{for}, URIK1^{for}		
	MG052	CYTD, DCYTD	URAt2^{for} or EX_ura(e)		
	MG053	PMANM	PGAMT^{back} or G1PACT^{for}	ACGAMPM^{for}	
	MG107	DGK1, GK1, GK2	NDPK8^{for}		
	MG111	G6PI,PGI	FRUpts^{for} or EX_fru(e)^{back}		
	MG187	GLYC3Pabc	GLYCt^{back} or EX_glyc(e)^{back}		
	MG188	GLYC3Pabc	GLYCt^{back} or EX_glyc(e)^{back}		
	MG189	GLYC3Pabc	GLYCt^{back} or EX_glyc(e)^{back}		
	MG215	PFK	FRUpts^{for} or EX_fru(e)^{back}		
	MG273	PDH	DATPt^{for} or EX_datp(e)^{back}		
	MG274	PDH	DATPt^{for} or EX_datp(e)^{back}		
	MG275	NADH5	G3PD4^{for}		
	MG299	PBUTT, PTA2r, PTAr	PGAMT^{back} or G1PACT^{for}	ACGAMPM^{for}	
	MG357	ACKr, PPAK	PGAMT^{back} or G1PACT^{for}	ACGAMPM^{for}	
	MG038	GLYK			Glycerol
	MG050	DRPAr			2-Deoxy-D-ribose 5-phosphate
	MG137	UDPGALM			UDP-D-galacto-1,4-furanose
	MG259	GLNMT			S-Adenosyl-L-homocysteine
	MG356	CHOLK		EX_chol(e), CHLabc^{for}	Choline phosphate
	MG372	THZPSN			4-Hydroxy-benzyl alcohol and 4-Methyl-5-(2-phosphoethyl)-thiazole and 1-deoxy-D-xylulose 5-phosphate
	MG396	RPI			D-Ribulose 5-phosphate
	MG448	METSR-R1, METSR-R2			L methionine R oxide
FNp	**MG410**	Plabc		GLYK^{back}	
	MG411	Plabc		GLYK^{back}	

doi:10.1371/journal.pcbi.1005036.t002

isoenzymes in the *M. tuberculosis* model [27], one of which shows strong sequence similarity to the *M. genitalium* genome. Notably, PGAMT is an essential reaction in the original network reconstruction [21], and would thus not be removed by previous algorithms that consider reaction additions and removals independently [12]. An alternative to the removal of PGAMT is the deletion of G1PACT; both reactions are not associated with any genes. G1PACT and PGAMT provide an alternative pathway to metabolize actetyl-CoA. Knocking out one of these genes, PTAr (MG299) and ACKr (MG357) become the only enzymes capable of metabolizing acetyl-CoA and thus become essential. Removing only G1PACT or PGAMT would seem to resolve the FPp for MG299 and MG357, but would result in a metabolic network unable to

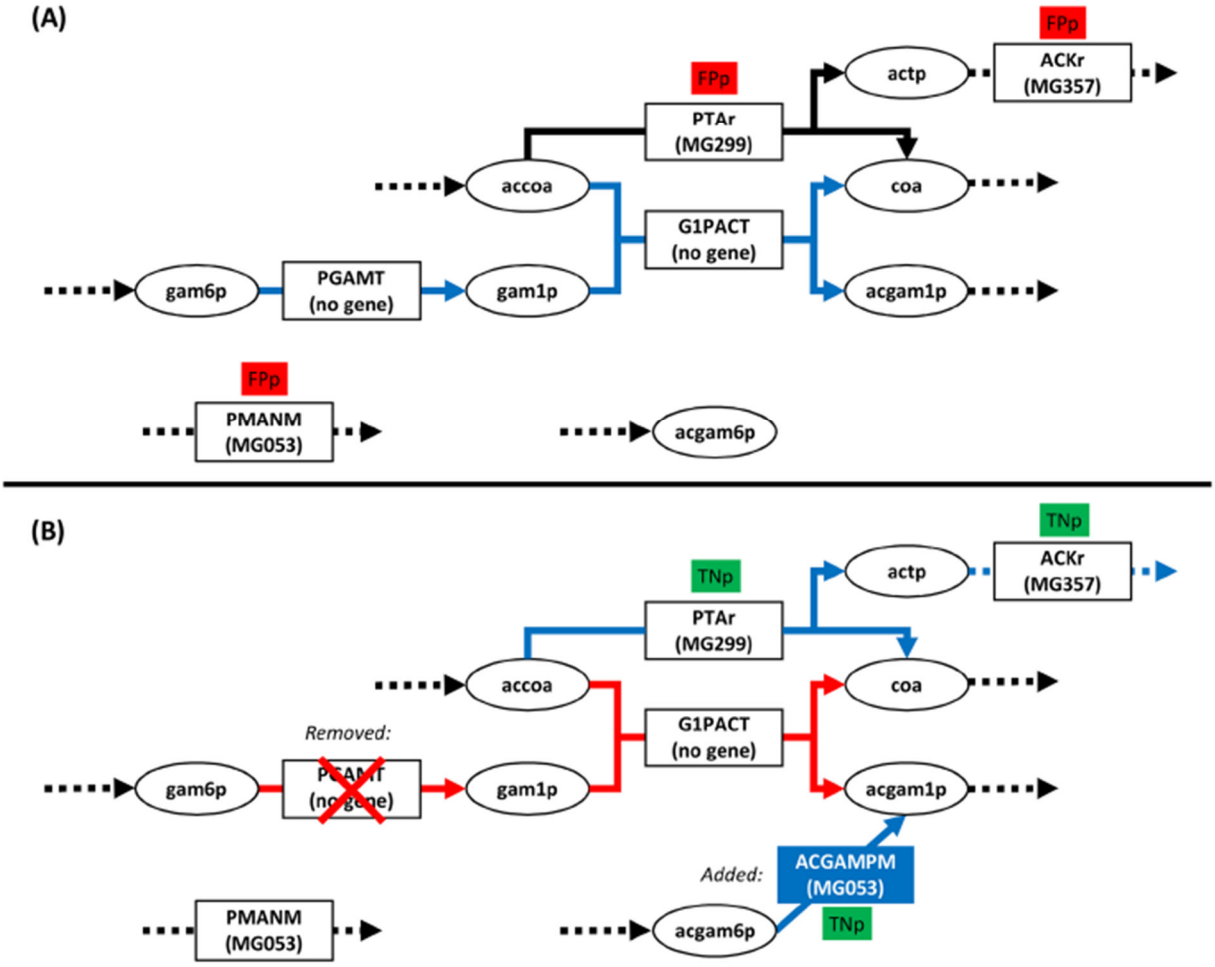

Fig 1. An example for the utility of simultaneously adding and removing reactions. Ellipses indicate metabolites, rectangles indicate reactions; abbreviations are taken from iPS189 [21]. (A) N-Acetyl-D-glucosamine 1-phosphate (acgam1p) is produced by G1PACT; MG053, MG299, and MG357 are falsely predicted to be non-essential (FPp). (B) The simultaneous removal of PGAMT (or, alternatively, G1PACT) and addition of ACGAMPM makes the genes MG053, MG299, and MG357 essential. Blue arrows mark essential pathways, while red arrows indicate blocked pathways. Note that removing either one of PGAMT or G1PACT blocks the other reaction, and that both reactions are not associated with any genes.

doi:10.1371/journal.pcbi.1005036.g001

produce the essential biomass precursor N-Acetyl-D-glucosamine 1-phosphate and would thus be unviable.

Our second application of GlobalFit to the *M. genitalium* model followed [21] by allowing changes to all reactions and biomass metabolites. The resulting model changes form a superset of those proposed by the conservative analysis. We rectified FPp for 8 further cases, resulting in a specificity of 98.3%. All eight were resolved by adding metabolites to the biomass (Table 2); in one case, a further addition of two reactions was required (EX_chol(e), CHLabcfor; Table 2). Note that these biomass changes are not conservative; while they resolve inaccuracies *in silico*, they should be confirmed through further experiments. Previous studies [10, 12, 16] have also

shown that modifying the biomass equation can improve the fit of model predictions to experimental growth data. However, estimating the correct biomass composition still remains a challenging task [7].

The two remaining unexplained FPp correspond to knocked-out genes associated with the same reaction as another gene whose knockout was a true positive prediction; thus, these predictions cannot be rectified without changing the gene-reaction associations.

The GlobalFit calculations for simultaneously solving all 11 feasible FPp cases (the number of unique enzyme complexes with FPp, Table 2) against the only FNp (two genes with FNp associated with the same reaction, Table 2) required 3h on a standard desktop computer (2 cores of an AMD Phenom 9600B 2.3GHz with 8GB RAM). However, as outlined above, the main advantage of GlobalFit is already played out when contrasting pairs of growth cases, which are much faster to solve. In the application to *M. genitalium*, we alternatively tested the subset strategy of first solving each FPp case separately against a wild-type control and each FNp alone; if the suggested solution turned the predictions for any other cases from true to false, we iteratively contrasted each case with the complete set of these negatively affected predictions. For the *M. genitalium* network, this approximate subset strategy resulted in the same proposed changes as the global analysis, while reducing the total computation time to below one minute. This result indicates that the application of GlobalFit is feasible even for very large growth datasets when employed in subset mode.

Solving false negative predictions (FNp). FNp can be due to missing isoenzymes. Thus, an important pre-processing step to the application of GlobalFit is to identify homologous genes within the genome and to make corresponding changes to the GPRs. A blast e-value threshold of 10^{-13} has been used successfully before for isoenzyme identification in *E. coli* K12 [12]; however, we could not find any close homologs for the remaining two FNp mutants at this threshold.

For FNp, the results of the conservative and non-conservative application of GlobalFit were identical. Two FNp cases (Table 2), which together act as phosphate importers, could be resolved by allowing the reversibility of the phosphorylation of glycerol. This reaction is predicted to be reversible without uncertainty in *E. coli* [25]; furthermore, the glycerol kinase of *M. genitalium* shows strong sequence similarity (BLAST e-value 10^{-136}) to the glycerol kinase of *Trypanosoma brucei*, which is known to indeed catalyze the reverse reaction [28, 29]. This single reversibility change increased sensitivity from 76.5% to 88.2%.

All modifications suggested by GlobalFit in the resolution of FPp and FNp cases were fully consistent with each other. In the highly conservative application of GlobalFit, we achieved an accuracy of 93.6% (MCC = 0.68; Table 1). If we follow previous work [21] by allowing all possible changes, GlobalFit obtains a global accuracy of 97.8%, and a Matthews correlation coefficient MCC = 0.86 (Table 1). The corresponding models differ only in their biomass reaction, and are supplied as S1 Model in SMBL format (non-conservative model: biomass reaction "Biomass"; conservative model: biomass reaction "Biomass_conservative").

Test case 2: Improving the iJO1366 metabolic model for *E. coli*

To test the applicability of GlobalFit's subset strategy to larger models, we next applied it to the most recent genome-scale metabolic reconstruction for *E. coli*, iJO1366 [20]. Again, we employed the same gene knockout essentiality data [30, 31] as used in the initial reconstruction. For all FBA simulations, we used the same parameters as described in [20]. The maximal influx of all nutrients in the defined growth media was set to 10 mmol $gDW^{-1}h^{-1}$. The lower bound of the non-growth associated maintenance reaction (ATPM) was set to 3.15 mmol $gDW^{-1}h^{-1}$. Gene essentiality was then calculated by FBA, considering any flux larger than 5%

Table 3. Comparison of experimental and predicted viability for 1366 *E. coli* gene knockouts on two different minimal media.

	Experiment			
Predictions	**growth**	**non-growth**	**Accuracy**	**MCC**
Unoptimized model (iJO1366) grown on glucose				
growth	1079	80	91.3%	0.69
no growth	39	168		
Unoptimized model (iJO1366) grown on glycerol				
growth	1073	87	90.3%	0.66
no growth	45	161		
Optimized model grown on glucose				
growth	1104	45	95.7%	0.85
no growth	14	203		
Optimized model grown on glycerol				
growth	1096	44	95.2%	0.83
no growth	22	204		

doi:10.1371/journal.pcbi.1005036.t003

of the optimal biomass core reaction as growth. For the published iJO1366 model, we obtained the same accuracies as reported originally [20]: a combined global accuracy of 90.8% calculated across knockout experiments on glucose and on glycerol media, corresponding to a Matthew's correlation coefficient MCC = 0.67 (Table 3).

In the application of GlobalFit to the iJO1366 model, we only allowed conservative network modifications (as defined for the *M. genitalium* model). However, as the growth medium used in the *E. coli* experiments was chemically defined, we did not allow the removal of exchange reactions. We constructed a database of potential new reactions as for *M. genitalium* (S2 Database).

The knockout data for *E. coli* includes growth data on two different media that contained either glucose or glycerol as carbon sources [30, 31]. Accordingly, we solved all FPp against two wild-type growth cases, one on glucose and one on glycerol. While this increases the number of continuous variables compared to using only a single wild-type growth case, the number of binary variables is still the same as in algorithms that only consider a single non-growth case at a time [12] (note that we don't allow the exclusion of any growth/non-growth case in this application). We tested if the order in which false growth/non-growth predictions are considered in GlobalFit's subset strategy affects the final result; this was not the case.

By applying the network modifications suggested by GlobalFit, we could strongly increase the quality of predictions for growth on both glycerol and glucose (Table 3); for the experiments on glucose and on glycerol combined, accuracy increased from 90.8% to 95.4%, while Matthew's correlation coefficient increased from 0.67 to 0.84. The detailed model changes are outlined below.

Solving FNp: Isoenzymes. One simple explanation for FNp is the existence of un-annotated isoenzymes. To detect such cases, we identified all FNp where the knocked-out gene has a significant bi-directional blast hit with another gene in the genome (*i.e.*, BLAST e-value $< 10^{-13}$ for the other gene when using either of the two as query). Such highly conserved homologs are likely to be functionally very similar to the knocked-out gene [12], and we updated the GPR accordingly. We only performed this analysis for those genes that were reported to be non-essential on both glucose and glycerol. In this way, we could convert six FNp to TPp (Table 4). In two cases (b0888 and b1702), the requirement for the inclusion of isoenzymes was not previously recognized, as the iJO1366 model wrongly included an alternative pathway; solving a

Table 4. Isoenzymes that resolved FNp.

Gene	Associated reactions	Isoenzyme	e-value →	e-value ←
b0888	TRDR	b0606	$2x10^{-35}$	$8x10^{-37}$
b0928	ASPTA	b4054	$2x10^{-113}$	$2x10^{-113}$
b1415	GCALDD, LCADi	b1385	$7x10^{-80}$	$1x10^{-77}$
b1702	PPS	b2383	$2x10^{-22}$	$2x10^{-22}$
b3176	PGAMT	b2048	$3x10^{-16}$	$1x10^{-18}$
b3359	SDPTA	b1748	$1x10^{-180}$	$1x10^{-180}$

doi:10.1371/journal.pcbi.1005036.t004

FPp related to the alternative pathway converted the TPp into a FNp that was then rescued by the inclusion of the newly identified isoenzymes.

Solving FNp: Removing biomass components. Removing metabolites from the biomass reaction can convert FNp to TPp, as all genes involved in the production (or, if the metabolite was a product of the biomass reaction, consumption) of a metabolite become unessential. GlobalFit suggested the removal of six metabolites from the biomass reaction, thereby resolving 19 FNp (Table 5). For example, removing Bis-molybdopterin guanine dinucleotide from the biomass reaction converted eight genes involved in the synthesis of this metabolite from essential to non-essential genes. By removing Bis-molybdopterin guanine dinucleotide and Thiamine diphosphate, two TNp become FPp (b0417 and b2530); however, because these two changes also correct 16 FNp, the overall accuracy was strongly increased.

GlobalFit further indicated the removal of calcium and copper from the biomass, which was also suggested by the BioMog algorithm based on *E. coli* growth data on different media [16]. Calcium is essential for proper functioning of *E. coli* chemotaxis [32]. However,

Table 5. Removal of biomass components from the *E. coli* model suggested by GlobalFit to remove FNp.

Gene	Associated reactions	Removed biomass metabolite
b0009	MPTAT	Bis-molybdopterin guanine dinucleotide
b0423	THZPSN3	Thiamine diphosphate
b0781	CPMPS	Bis-molybdopterin guanine dinucletide
b0783	CPMPS	Bis-molybdopterin guanine dinucletide
b0784	MOADSUx, MPTS	Bis-molybdopterin guanine dinucletide
b0785	MPTS	Bis-molybdopterin guanine dinucletide
b0826	MPTSS	Bis-molybdopterin guanine dinucletide
b0827	BMOCOS, BWCOS, MOCOS, WCOS	Bis-molybdopterin guanine dinucletide
b2103	PMPK	Thiamine diphosphate
b3040	CD2tpp, CU2tpp, FE2tpp, MN2tpp, ZN2tpp	Copper
b3196	CAt6pp	Calcium
b3807	I2FE2SS, I2FE2SS2, S2FE2SS, S2FE2SS2	[4Fe-4S] iron-sulfur cluster and [2Fe-2S] iron-sulfur cluster
b3857	BMOGDS1, BMOGDS2, BWCOGDS1, BWCOGDS2, MOGDS	Bis-molybdopterin guanine dinucletide
b3990	THZPSN3	Thiamine diphosphate
b3991	TYRL	Thiamine diphosphate
b3992	THZPSN3	Thiamine diphosphate
b3993	TMPPP	Thiamine diphosphate
b3994	AMPMS2	Thiamine diphosphate
b4407	THZPSN3	Thiamine diphosphate

doi:10.1371/journal.pcbi.1005036.t005

Table 6. Reversal of reactions of the *E. coli* network suggested by GlobalFit to remove FNp.

Gene	Associated reactions	Reversed reactions
b0159	5DOAN, AHCYSNS, MTAN	HCYSMT, CPPPGO2
b2103	PMPK	2MAHMP
b2687	RHCCE	HCYSMT
b3040	CD2tpp, CU2tpp, FE2tpp, MN2tpp, ZN2tpp	CU2abcpp
b3196	CAt6pp	CA2t3pp

doi:10.1371/journal.pcbi.1005036.t006

compromised chemotaxis will not be detected in the knockout experiments. Thus, we suggest to retain calcium in the biomass reaction when modeling *E. coli* in its natural habitat, but to remove calcium from the biomass reaction when modeling *E. coli* in cell culture.

Solving FNp: Reversing reactions. Five FNp could be resolved by reversing existing reactions in the metabolic network (Table 6). Interestingly, an alternative solution for two genes was to remove calcium or copper from the biomass reaction. For calcium, the above arguments indicate that its removal from the biomass reaction may be preferable.

Solving FNp: Adding new reactions to the network. GlobalFit could not improve the accuracy of knockout predictions by adding new reactions to the metabolic network. This may have several reasons. First, the reconstruction of the *E. coli* metabolic network iJO1366 involved extensive literature and database searches to ensure a maximal inclusion of metabolic reactions. Second, we used the BiGG database as the source for potential additional reactions. Many networks in this database are based on the *E. coli* network reconstruction; this makes it unlikely that they provide new features relevant for *E. coli*. Third, the cut-off value for the similarity of enzymes to the *E. coli* genome used in the construction of the additional reaction database might have been too strict (10^{-13}).

Solving FPp: Adding metabolites to the biomass reaction. 22 FPp could be resolved by adding metabolites as substrate or product to the biomass reaction (Table 7). 17 of these corresponded to (previously blocked) tRNA charging reactions; these were resolved by adding charged and uncharged tRNA metabolites to the two sides of the biomass reaction, respectively, similar to previous suggestions for the older iAF1260 *E. coli* model [12]. GrowMatch only considers additions to the biomass if a gene with a FPp catalyzes a blocked reaction; it then tests the addition of the metabolites consumed or produced by this reaction [12]. However, none of the genes for the remaining five FPp resolved by GlobalFit through biomass additions catalyzed blocked reactions. When allowing the addition of biomass components not included in other BiGG biomass reactions or suggested by BioMog, GlobalFit was able to resolve 4 additional FPp (for b2533, b2925, b3623, b3650); however, as these suggested modifications did not meet our strict criteria, we did not consider them further.

Solving FPp: Removing reactions. 25 FPp could be resolved by removing a total of 18 reactions from the metabolic network (Table 8). At the same time, four TPp were converted to FNp; however, two of these newly introduced FNp could subsequently be corrected through additional network modifications.

One example is the ATP synthase reaction ATPS4rpp, which is catalyzed by an enzyme complex encoded by eight genes. When *E. coli* was grown on glycerol, six of these genes were essential, while on glucose only three genes were found to be essential. Thus, overall accuracy is optimized if ATPS4rpp is essential for growth on glycerol, but non-essential for growth on glucose. We used GlobalFit to simultaneously solve a non-growth case of the ATPS4rpp knockout on glycerol, a wild-type growth case on glycerol, and a growth case of the ATPS4rpp knockout on glucose. GlobalFit found two alternative solutions that make the Phosphoglycerate kinase reaction irreversible (removing the backward direction of PGK) and also make the

Table 7. Metabolite additions to the *E. coli* biomass reaction suggested by GlobalFit to resolve FPp.

Gene	Associated reactions	Added as biomass substrate	Added as biomass product
b0194	PROTRS	L-Prolyl-tRNA(Pro)	TRNA(Pro)
b0242	GLU5K	L-Glutamate 5-phosphate	
b0526	CYSTRS	L-Cysteinyl-tRNA(Cys)	TRNA(Cys)
b0529	MTHFC, MTHFD	5-Formyltetrahydrofolate	
b0642	LEUTRS	L-Leucyl-tRNA(Leu)	TRNA(Leu)
b0680	GLNTRS	L-Glutaminyl-tRNA(Gln)	TRNA(Gln)
b0893	SERTRS, SERTRS2	L-Seryl-tRNA(Ser)	TRNA(Ser)
b0930	ASNTRS	L-Asparaginyl-tRNA(Asn)	TRNA(Asn)
b1637	TYRTRS	L-Tyrosyl-tRNA(Tyr)	TRNA(Tyr)
b1713	PHETRS	L-Phenylalanyl-tRNA(Phe)	TRNA(Phe)
b1714	PHETRS	L-Phenylalanyl-tRNA(Phe)	TRNA(Phe)
b1719	THRTRS	L-Threonyl-tRNA(Thr)	TRNA(Thr)
b1866	ASPTRS	L-Aspartyl-tRNA(Asp)	TRNA(ASP)
b1876	ARGTRS	L-Arginyl-tRNA(Arg)	TRNA(ARG)
b1912	PGSA120, PGSA140, PGSA141, PGSA160, PGSA161, PGSA180, PGSA181	Phosphatidylglycerophosphate (didodecanoyl, n-C12:0) or Phosphatidylglycerophosphate (ditetradecanoyl, n-C14:0) or Phosphatidylglycerophosphate (ditetradec-7-enoyl, n-C14:1) or Phosphatidylglycerophosphate (dihexadecanoyl, n-C16:0) or Phosphatidylglycerophosphate (dihexadec-9-enoyl, n-C16:1) or Phosphatidylglycerophosphate (dioctadecanoyl, n-C18:0) or Phosphatidylglycerophosphate (dioctadec-11-enoyl, n-C18:1)	
b2114	METTRS		TRNA(Met)
b2514	HISTRS	L-Histidyl-tRNA(His)	TRNA(His)
b2551	GHMT2r, THFAT	5-Formyltetrahydrofolate	
b2913	PGCD	3-Phosphohydroxypyruvate	
b3288	FMETTRS	N-Formylmethionyl-tRNA	
b3384	TRPTRS	L-Tryptophanyl-tRNA(Trp)	TRNA(Trp)
b4258	VALTRS	L-Valyl-tRNA(Val)	TRNA(Val)

doi:10.1371/journal.pcbi.1005036.t007

Fructose 6-phosphate aldolase reaction ($F6PA^{back}$) or the Glucose 6-phosphate dehydrogenase ($G6PDH2r^{for}$) irreversible. By applying either of these two modifications, the two TPp of ATP synthase subunits for glycerol were converted to FNp.

For two of the 25 solved FPp (b0242 and b2913), alternative solutions are provided by adding metabolites to the biomass reaction (Table 7). For example, the FPp of b2913 (encoding Phosphoglycerate dehydrogenase) could be resolved by making the Glycine hydroxymethyltransferase reaction (GHMT2r) irreversible. An alternative solution is the addition of 3-Phosphohydroxypyruvate (3php) to the biomass reaction, which was also suggested by BioMog [16]. However, only the removal of GHMT2r simultaneously resolved the FPp of b4388 (Phosphoserine phosphatase (L-serine)).

Solving FPp: Other. On glucose, 19 of the remaining 45 FPp corresponded to isoenzymes; on glycerol, 21 of the 33 remaining FPp corresponded to isoenzymes. FBA models do not account for gene regulation, and thus the corresponding reactions are assumed to remain active even when knocking out one of the isoenzymes. Thus, these FPp are due either to erroneous GPRs or to the isoenzymes not being expressed. GlobalFit does not allow changes to GPRs or inclusion of regulatory rules, and, consequently, could not find any solution for these genes.

The resulting modified model of *E. coli* metabolism is provided as S2 Model in SBML format.

Table 8. Removal of reactions of the *E. coli* network suggested by GlobalFit to correct FPp.

Gene	Associated reactions	Removed reactions
b0032	CBPS	(CBMKr^{for} and ALLTAMH^{for}) or (CBMKr^{for} and ALLTN^{for}) or (CBMKr^{for} and OXAMTC^{for}) or (CBMKr^{for} and URDGLYCD^{for}) or (CBMKr^{for} and URIC^{for})
b0033	CBPS	(CBMKr^{for} and ALLTAMH^{for}) or (CBMKr^{for} and ALLTN^{for}) or (CBMKr^{for} and OXAMTC^{for}) or (CBMKr^{for} and URDGLYCD^{for}) or (CBMKr^{for} and URIC^{for})
b0242	GLU5K	NACODA^{for}
b0243	G5SD	NACODA^{for}
b0474	ADK1, ADK3, ADK4, ADNK1, DADK	NDPK1^{for} or PRPPS^{back} or R1PK^{for} or PPM^{back} or R15BPK^{for}
b0945	DHORD2, DHORD5	DHORDfum^{for}
b0954	T2DECAI	(CTECOAI6^{back} and CTRCOAI7^{back}) or (CTECOAI6^{back} and AACPS4^{for})
b1207	PRPPS	R1PK^{for} or PPM^{back} or R15BPK^{for}
b1638	PDX5POi, PYAM5PO	PDX5PO2^{for}
b1779	GAPD	TPI^{for}
b2234	RNDR1, RNDR2, RNDR3, RNDR4	(GRXR^{for} and RNTR3c2^{for}) or (GTHOr^{for} and RNTR3c2^{for}) or (GRXR^{for} and RNTR1c2^{for}) or (GTHOr^{for} and RNTR1c2^{for})
b2235	RNDR1, RNDR2, RNDR3, RNDR4	(GRXR^{for} and RNTR3c2^{for}) or (GTHOr^{for} and RNTR3c2^{for}) or (GRXR^{for} and RNTR1c2^{for}) or (GTHOr^{for} and RNTR1c2^{for})
b2508	IMPD	HXAND or XPPT
b2913	PGCD	GHMT2r^{back}
b2926	PGK	TPI^{for}
b3731	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3733	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3734	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3735	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3736	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3738	ATPS4rpp	(F6PA^{back} and PGK^{back}) or (G6PDH2r^{for} and PGK^{back})
b3835	OPHHX	OPHHX3^{for}
b3956	PPC	FUM^{for} or MALS^{for}
b4041	G3PAT120, G3PAT140, G3PAT141, G3PAT160, G3PAT161, G3PAT180, G3PAT181	ACPPAT160^{for} or AG3PAT161^{for} or AG3PAT160^{for}
b4388	PSP_L	GHMT2r^{back}

doi:10.1371/journal.pcbi.1005036.t008

Discussion

In this work, we describe and implement a novel algorithm to automatically modify metabolic network models based on growth/non-growth data. The algorithm can utilize data from different growth environments and/or different gene knockouts. In contrast to previous approaches, the "global" mode of GlobalFit does not reconcile the network model with inconsistent experiments iteratively, but finds a globally minimal set of network changes that resolves all inconsistencies simultaneously (in so far as the inconsistencies are resolvable with the allowed model

modifications). To make GlobalFit applicable to large metabolic network reconstructions, we also explored a subset strategy, where individual false predictions are solved simultaneously with small subsets of growth/non-growth cases.

We demonstrate the utility of these approaches through applications to the previously published network models of *M. genitalium* [21] (optimizing model predictions for gene knockout data from Ref. [22]) and *E. coli* [20] (utilizing gene knockout data from Ref. [30, 31]). Allowing only highly conservative network changes (e.g., only adding reactions catalyzed by enzymes that are homologous to genes of the species studied), we were able to halve the number of false growth predictions in each case. Overall, GlobalFit improved the accuracy of growth/non-growth predictions for *M. genitalium* from 87.3% to 93.6% (MCC from 0.56 to 0.68) and for *E. coli* from 90.8% to 95.4% (MCC from 0.67 to 0.84). If we allow a much wider range of possible network modifications—which is routinely done in alternative approaches [12, 21]–even higher accuracies can be achieved. Importantly, GlobalFit can enumerate alternative optimal or sub-optimal solutions, such that expert knowledge or additional experiments can help select the biologically most realistic modifications.

For some inconsistencies, we found solutions that improved accuracy on one medium while decreasing accuracy on the other. For example, adding selenium to the biomass reaction of *E. coli* would resolve three FPp on glycerol, while converting four TPp to FNp on glucose. Thus, the accuracy achievable for one growth medium could be further improved by sacrificing the accuracy for the other medium, albeit at a likely loss of biological correctness. This observation emphasizes the utility of combining gene knockout data across different nutritional environments to avoid problems of overfitting.

In other cases, several genes whose products act together in a protein complex had contradictory experimental results: in the same medium, some were found to be essential, while the rest was declared non-essential. Such contradictions may be caused either by experimental errors, by erroneous assignment of genes to reactions (incorrect GPRs), or by a residual function of the enzyme complex even with some of its components missing. GlobalFit may suggest a solution in this case, but this will simultaneously distort one or more true predictions. For example, the FPp for the *E. coli* gene b3560 (the α-subunit of glycine tRNA synthetase) could be resolved by adding the charged and uncharged glycine tRNA to the biomass reaction as substrate and product, respectively. This modification would at the same time transform the TPp of b3559 (the β-subunit) to a FNp, and would thus not improve accuracy.

In the applications of GlobalFit, we adopted the *in silico* growth cutoffs used in the original model publications, *i.e.*, one third of the mean growth rate for *M. genitalium* [21] and 5% of the optimal biomass core reaction for *E. coli* [20]. A more general way to resolve FPp would be to treat the cutoff that distinguishes *in silico* growth from non-growth as an additional variable in the optimizations. For example, the knockout of *E. coli* ATPS4rpp reduced the biomass yield in glycerol below 10% of the wild-type yield. Such a substantial reduction in growth rate may explain why 6 out of 8 knockouts for the genes involved in the corresponding enzyme complex were labeled as essential in the experiment; however, following [20] in considering 5% biomass production as growth, we regarded these knockouts as FPp in this study. An adjustable growth threshold might have rectified these FPp cases without any model changes. It is not clear *a priori* which *in silico* cutoff corresponds best to a given set of experimental data. Thus identifying the cutoff value that minimizes the necessary model changes seems most appropriate.

In this paper, we have explored the application of GlobalFit to the improvement of existing metabolic network reconstructions and showed that it can substantially reduce the number of false growth predictions even when restricted to conservative network changes. It is conceivable that GlobalFit can also be employed for other tasks related to metabolic model refinement. One possible such application is the initial reconstruction of a metabolic network model

starting from a computer-generated template that is based on genome annotation (such as provided, e.g., by the SEED algorithm [33]). GlobalFit might also be used to remove thermodynamically impossible energy-creating cycles, which sometimes plague initial network reconstructions. While we only score growth and non-growth, GlobalFit could also be applied using yield data by choosing appropriate thresholds. Finally, we envisage future usage of *GlobalFit* for strain optimization in metabolic engineering applications that combine gene knockouts [34] with gene additions.

Methods

Formal problem definition

GlobalFit compares flux-balance analysis (FBA) [17] model predictions to measured growth across all tested environments and gene knockouts simultaneously. Allowed model changes are (i) removals or (ii) reversibility changes of existing reactions; (iii) additions of reactions to the model from a database of potential reactions; (iv) removals of metabolites from the biomass; and (v) additions of metabolites to the biomass.

We thus solve the following bi-level problem:

$$\min_{\delta}\Big(\sum_{y\in M}(\delta_y^{RF}+\delta_y^{RB})\times w_y^R+\sum_{x\in I}\delta_x^I\times w_x^I+\sum_{z\in D}\delta_z^{add}\times w_z^{add}+\sum_{j\in A_S}\delta_j^{AS}\times w_j^{AS}+\sum_{k\in A_P}\delta_k^{AP}\times w_k^{AP}+\sum_{l\in B_S}\delta_l^{RS}\times w_l^{RS}+\sum_{m\in A_P}\delta_m^{RP}\times w_m^{RP}+\sum_{g\in G}\delta_g^G\times w_g^G+\sum_{h\in N}\delta_h^N\times w_h^N\Big) \quad (1)$$

subject to:

$$\forall_{g\in G} S\times v_g=0 \quad (2)$$

$$\forall_{h\in G} S\times v_h=0 \quad (3)$$

$$\forall_{y\in M,\ g\in G\cup N}\ v_y^{min}\times(1-\delta_y^{RB})\le v_y^g\le v_y^{max}\times(1-\delta_y^{RF}) \quad (4)$$

$$\forall_{x\in I,\ g\in G\cup N}-1000\times\delta_x^I\le v_x^g \quad (5)$$

$$\forall_{z\in D,\ \ g\in G\cup N}\ 0\le v_z^g\le 1000\times\delta_z^{add} \quad (6)$$

$$\forall_{y\in M,\ \ g\in G\cup N}\sum_{l\in B_S}(1-\delta_l^{RS})\times c_l^{RS}+\sum_{j\in A_S}\delta_j^{AS}\times c_j^{AS}\xrightarrow{v_{Bio}^g}\sum_{m\in B_P}(1-\delta_m^{RP})\times c_m^{RP}+\sum_{k\in A_P}\delta_k^{AP}\times c_k^{AP} \quad (7)$$

$$\forall_{g\in G}\ (v_{Bio}^g+1000\times\delta_{Bio}^{IG}\ge T_g) \quad (8)$$

$$\forall_{h\in N}\ (\hat{v}_{Bio}^h-1000\times\delta_{Bio}^{IN}\le T_h) \quad (9)$$

with:

$$Inner\ Problem: \hat{v}_{Bio}^h:=\max_{v^h}v_{Bio}^h, \quad (10)$$

subject to: Eqs (3)–(7) *and to the definitions following below.*

Table 9. Definitions of the sets used in the system of equations that describes the GlobalFit algorithm.

M	The reactions included in the original (input) network reconstruction
I	All irreversible reactions that can be reversed
D	All reactions that can be added to the network (here, we consider bidirectional reactions as two separate reactions corresponding to forward and backward directions (with fluxes $\geq$0)).
B_S	All substrates that can be removed from the biomass reaction
c^{BS}	The stoichiometric coefficients of all biomass substrates
B_P	All products that can be removed from the biomass reaction
c^{BP}	The stoichiometric coefficients of all biomass products
A_S	All substrates that can be added to the biomass reaction
c^{AS}	The stoichiometric coefficients of all additional biomass substrates
A_P	All products that can be added to the biomass reaction
c^{AP}	The stoichiometric coefficients of all additional biomass products
G	All experiments with observed growth
N	All experiments with observed non-growth

doi:10.1371/journal.pcbi.1005036.t009

Line (7) defines the flux through the biomass reaction, v^g_{Bio}, for condition g. The sets used in this system of equations are listed in Table 9, while the parameters are defined in Table 10. For binary variables, 1 corresponds to TRUE (i.e., a model change is executed), while 0 corresponds to FALSE (no change compared to the initial network).

GlobalFit's logic

What is the purpose of each of the lines in the above system of equations? The network must be in a steady state (*i.e.*, no concentration changes to internal metabolites) in all conditions $g \in G$ Eq (2) and $h \in N$ Eq (3) that are to be solved simultaneously.

Lines (4)–(6) convert the binary variables for the removal or reversibility change of existing reactions, and for the addition of new reactions from the database, into constraints for the respective fluxes. In Eq (4), if $\delta_y^{RB} = 0$ (*i.e.*, no change), then the lower limit for reaction y in all conditions g (v_y^g) remains at the predefined limit v_y^{min}; setting $\delta_y^{RB} = 1$ instead sets the lower flux limit to 0, *i.e.*, removes the backwards reaction. Similarly, setting $\delta_y^{RF} = 0$ keeps the upper flux limit for reaction y at the predefined limit v_y^{max}, while setting $\delta_y^{RF} = 1$ sets the upper flux limit to 0, *i.e.*, removes the forward reaction.

Line (5) sets the lower flux limit to -1000 for reaction y in all conditions g if $\delta_x^I = 1$, i.e., it makes an irreversible reaction (with flux $v_x^g \geq 0$) reversible in this case. Line (6) allows non-zero (positive) flux for reactions that are not part of the original (input) model if $\delta_z^{add} = 1$. Note that in the database of additional potential reactions, we consider bidirectional reactions as two separate reactions corresponding to forward and backward directions (both with fluxes $\geq$0).

Metabolites can be removed from both sides of the biomass reaction (flux v^g_{Bio}), and additional metabolites can be added Eq (7) with pre-specified stoichiometric coefficients c.

To ensure *in silico* growth for conditions with experimentally demonstrated growth, the biomass flux for these conditions must be greater than a predefined threshold T_g in all conditions $g \in G$ Eq (8). Conversely, to ensure *in silico* non-growth for conditions with experimentally demonstrated non-growth, the biomass flux for these condition must be less than a predefined threshold T_h in all conditions $h \in N$ Eq (9). The thresholds T_g and T_h can be set separately for each phenotype, *e.g.*, to account for estimates of experimental errors. For non-growth phenotypes, a simple condition that forces the biomass production to be lower than a threshold is not

Table 10. The parameters of the system of equations describing the GlobalFit algorithm.

$\delta_y^{RF}, \delta_y^{RB} \in \{0,1\}$	Binary variables that indicate the removal of forward and backward reaction y, respectively
$w_y^R > 0$	Penalty for the removal of forward or backward reaction (which can be set to a different value for each reaction y)
$\delta_x^I \in \{0,1\}$	Binary variables that indicate the addition of a backward reaction for reaction x
$w_x^I > 0$	Corresponding penalties
$\delta_z^{add} \in \{0,1\}$	Binary variables that indicate the addition of reaction z
$w_z^{add} > 0$	Corresponding penalties
$\delta_j^{AS} \in \{0,1\}$	Binary variables that indicate the addition of substrate j to the biomass reaction
$w_j^{AS} > 0$	Corresponding penalties
$\delta_k^{AP} \in \{0,1\}$	Binary variables that indicate the addition of product k to the biomass reaction
$w_k^{AP} > 0$	Corresponding penalties
$\delta_l^{RS} \in \{0,1\}$	Binary variables that indicate the removal of substrate l from the biomass reaction
$w_l^{RS} > 0$	Corresponding penalties
$\delta_m^{RP} \in \{0,1\}$	Binary variables that indicate the removal of product m from the biomass reaction
$w_m^{RP} > 0$	Corresponding penalties
$\delta_g^G \in \{0,1\}$	Binary variables that indicate the exclusion of growth experiment g
$w_g^G > 0$	Corresponding penalties
$\delta_h^N \in \{0,1\}$	Binary variables that indicate the exclusion of non-growth experiment h
$w_h^N > 0$	Corresponding penalties
v_{Bio}^g	Flux through the (potentially modified) biomass reaction (see line (7))
$\hat{v}_{Bio}^g$	Optimal value for v_{Bio}^g estimated in the inner problem
$v_y^{min} \leq 0$	Minimal flux allowed through reaction y (note that we do not allow minimal fluxes >0 for non-growth cases)
$v_y^{max} \geq 0$	Maximal flux allowed through reaction y (note that we do not allow maximal fluxes <0 for non-growth cases)
$T_g > 0$	Viability threshold of growth experiment g
$T_h > 0$	Viability threshold of non-growth experiment h
$\vec{\delta}$	The vector of all δ defined above
$\vec{v}^h$	The vector of all fluxes v_i^h for experiment h

doi:10.1371/journal.pcbi.1005036.t010

sufficient, though, as a trivial solution with $\vec{v}^h = 0$ would satisfy this condition. To overcome this problem, the inner optimization problem maximizes the biomass production of non-growth cases Eq (9), and this maximum is compared against the non-growth threshold.

Line (1) describes the outer optimization problem. GlobalFit aims to find a solution that is able to correctly predict all growth and non-growth cases with a minimal number of network changes (indicated by values 1 for the binary variables):

$$\delta_y^{RF}, \delta_y^{RB}, \delta_x^I, \delta_z^{add}, \delta_j^{AS}, \delta_k^{AP}, \delta_l^{RS}, \delta_m^{RP}, \delta_g^G, \delta_h^N$$

The penalties for each type of network change, and even for each individual change, can be set independently. This allows, for example, to prefer reversibility changes over reaction additions, or to preferentially include new reactions with stronger genomic evidence, or reactions from metabolic network reconstructions of close relatives. Users should choose appropriate penalties based on the details of the network reconstruction and the proposed changes. As a starting point, we include a list of suggested penalty values in S1 Table).

To guarantee a feasible solution, even if inconsistent growth cases are used, we implemented additional binary variables that allow the exclusion of individual growth (δ_g^G Eq (8)) and non-growth cases (δ_h^N Eq (9)) from the growth threshold conditions. In our application to the *M. genitalium* network, we penalize these condition exclusions with very high values w_g^G and w_h^N; thus, any network modification that explains additional cases is preferred over the exclusion of conditions, regardless of the number of required changes. Instead, the penalties can be set to smaller values, so that the exclusion of potentially erroneous experiments is preferred over excessive network changes.

Metabolic network reconciliation with large-scale experimental data usually incorporates a manual curation stage, where experts for the physiology and biochemistry of the organism under study review network changes suggested by automated methods. To support this process, GlobalFit can put out not just one best solution, but, *e.g.*, the five best solutions that can then be reviewed to identify the changes most compatible with existing knowledge. To speed up the calculations, network changes can also be limited to a maximal number.

Re-formulation of the bi-level as a single level optimization problem

No efficient software tools for general bi-level optimization problems are available. Solving the inner problem for each possible combination of network changes would be computationally too slow. We adapt the "Reduction Ansatz" of Section 4.3.4 in [18] to eliminate the inner problem in line (9). In this approach, the optimality conditions of the inner optimization problem are expressed as equality and inequality conditions using additional "dual" variables. For fixed $\vec{\delta}$ and h, the inner problem is simply a linear program; thus, the assumptions in [18] are trivially satisfied.

Because of the use of binary variables, algorithms to solve this type of optimization problem are termed mixed integer linear programming (MILP). MILP is NP hard [35]; while no known algorithms can guarantee to find a solution efficiently, algorithms that work well for many practical problems exist in software solvers. We used the solver of IBM ILOG CPLEX 12.5; to avoid trickle flow, we implemented indicator constraints. Alternatively, our implementation of GlobalFit also allows using the GUROBI solver. Academic users can obtain both CPLEX and GUROBI free of charge.

Preprocessing

The search for a globally minimal set of network changes is a computationally very intensive task. To speed up this process, it is advisable to restrict the examined conditions to a maximal consistent ("feasible") set, *i.e.*, a maximal set of conditions that can all be correctly predicted with the same modified metabolic network (regardless of the type and number of modifications). To identify such feasible condition sets, GlobalFit provides a *simple mode*, which only minimizes the number of erroneous predictions of growth regardless of the number of network changes. To speed up the calculation of a feasible condition set, it is possible to first solve individual wrong predictions against a "control" condition, thereby identifying conditions that cannot be reconciled with the network with the allowed modifications. We applied this strategy for the pre-processing of the *M. genitalium* data (see Results).

Furthermore, the number of binary variables can be reduced by a set of additional preprocessing steps. First, binary variables for changes to the network not allowed (such as reversibility changes to reactions strictly considered irreversible) should be constrained to zero. Second, we can consider a "supermodel" that encompasses the input model with all allowed reactions converted to reversible reactions and all reactions from the database of potential additional

reactions. We can then reduce the number of binary variables further by (i) excluding all reactions that are blocked in this supermodel, (ii) constraining to zero the binary variables for the removal of reactions that are essential in this supermodel.

Enumeration of alternative solutions

GlobalFit can optionally calculate a user-defined number n of alternative optimal or suboptimal solutions. The search for alternative solutions is executed using the integer cuts method. Thus, the complexity for each additional alternative solution is only increased through a single linear constraint. Consequently, the runtime for n alternative optimal or suboptimal solutions is approximately n times the runtime for a single optimum.

Implementation and availability

We provide an implementation of GlobalFit, integrated with the *sybil* toolbox for constraint-based analyses [19], which runs in the R environment for statistical computing [36]. The source code and documentation is available free of charge from CRAN (http://cran.r-project.org/web/packages/GlobalFit/). The optimized models for *E. coli* and *M. genitalium* are provided as SBML files that can be read, e.g., by *sybil* [19] and the COBRA toolbox [37].

Supporting Information

S1 Table. Users of GlobalFit should choose appropriate penalties for proposed model changes based on the details of the network reconstruction and the proposed changes. As a starting point, this table list some suggested penalty values.
(PDF)

S1 Database. To construct a database of potential additional reactions for the conservative application of GlobalFit to *M. genitalium*, we started from all reactions contained in metabolic networks provided by the BiGG database [24]. We then restricted this dataset to reactions that are catalyzed by enzymes with significant sequence similarity to the *M. genitalium* genome (BLAST e-value $<10^{-13}$). We removed globally blocked reactions, *i.e.*, those reactions of the database that were not able to carry any flux in a supernetwork containing all reactions. Reversible reactions were represented as two independent irreversible reactions, corresponding to forward and backward directions. The database is provided as a tab-delimited text file with three columns: reaction ID; stoichiometric equation; gene-protein-reaction association (GPR).
(TSV)

S2 Database. To construct a database of potential additional reactions for the conservative application of GlobalFit to *E. coli*, we started from all reactions contained in metabolic networks provided by the BiGG database [24]. We then restricted this dataset to reactions that are catalyzed by enzymes with significant sequence similarity to the *E. coli* genome (BLAST e-value $<10^{-13}$). We removed globally blocked reactions, *i.e.*, those reactions of the database that were not able to carry any flux in a supernetwork containing all reactions. Reversible reactions were represented as two independent irreversible reactions, corresponding to forward and backward directions. The database is provided as a tab-delimited text file with three columns: reaction ID; stoichiometric equation; gene-protein-reaction association (GPR).
(TSV)

S1 Model. The *M. genitalium* iPS189 models as modified by GlobalFit are supplied as an SMBL file, which can be read, e.g., by the *sybil* toolbox for *R* [19] or the COBRA toolbox for Matlab [37]. The two models differ only by their biomass reactions: "Biomass" for

the non-conservative model; "Biomass_conservative" for the conservative model.
(XML)

S2 Model. The *E. coli* iJO1366 model as modified by GlobalFit is supplied as an SMBL file, which can be read, e.g., by the *sybil* toolbox for *R* [19] or the COBRA toolbox for Matlab [37].
(XML)

Acknowledgments

We thank Balazs Papp and Jonathan Fritzemeier for helpful discussions.

Author Contributions

Conceived and designed the experiments: DH FJ MJL. Performed the experiments: DH. Analyzed the data: DH. Wrote the paper: DH FJ MJL.

References

1. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol. 2012; 10(4):291–305. doi: 10.1038/nrmicro2737. WOS:000301780900014. PMID: 22367118
2. Ibarra RU, Edwards JS, Palsson BO. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002; 420(6912):186–9. doi: 10.1038/nature01149. WOS:000179200900048. PMID: 12432395
3. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature genetics. 2005; 37(12):1372–5. doi: 10.1038/ng1686 PMID: 16311593.
4. Pal C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. Nature. 2006; 440(7084):667–70. doi: 10.1038/nature04568 PMID: 16572170.
5. Raman K, Rajagopalan P, Chandra N. Flux balance analysis of mycolic acid pathway: Targets for antitubercular drugs. PLoS computational biology. 2005; 1(5):349–58. ARTN e46 doi: 10.1371/journal.pcbi.0010046. WOS:000234713100003.
6. Lee JW, Kim TY, Jang YS, Choi S, Lee SY. Systems metabolic engineering for chemicals and materials. Trends Biotechnol. 2011; 29(8):370–8. doi: 10.1016/j.tibtech.2011.04.001. WOS:000293485300002. PMID: 21561673
7. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols. 2010; 5(1):93–121. doi: 10.1038/nprot.2009.203 PMID: 20057383; PubMed Central PMCID: PMC3125167.
8. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC bioinformatics. 2007; 8:212. doi: 10.1186/1471-2105-8-212 PMID: 17584497; PubMed Central PMCID: PMC1933441.
9. Zomorrodi AR, Suthers PF, Ranganathan S, Maranas CD. Mathematical optimization applications in metabolic networks. Metabolic engineering. 2012; 14(6):672–86. doi: 10.1016/j.ymben.2012.09.005 PMID: 23026121.
10. Orth JD, Palsson B. Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. BMC systems biology. 2012; 6:30. doi: 10.1186/1752-0509-6-30 PMID: 22548736; PubMed Central PMCID: PMC3423039.
11. Thiele I, Vlassis N, Fleming RM. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014; 30(17):2529–31. doi: 10.1093/bioinformatics/btu321 PMID: 24812336; PubMed Central PMCID: PMC4147887.
12. Satish Kumar V, Maranas CD. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol. 2009; 5(3):e1000308. doi: 10.1371/journal.pcbi.1000308 PMID: 19282964; PubMed Central PMCID: PMC2645679.
13. Agren R, Liu LM, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. PLoS computational biology. 2013; 9(3). ARTN e1002980 doi: 10.1371/journal.pcbi.1002980. WOS:000316864200050.

14. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol Biol. 2013; 985:17–45. doi: 10.1007/978-1-62703-299-5_2 PMID: 23417797.

15. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, et al. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 2015. doi: 10.1093/bib/bbv079 PMID: 26454094.

16. Tervo CJ, Reed JL. BioMog: a computational framework for the de novo generation or modification of essential biomass components. PloS one. 2013; 8(12):e81322. doi: 10.1371/journal.pone.0081322 PMID: 24339916; PubMed Central PMCID: PMC3855262.

17. King ZA, Lloyd CJ, Feist AM, Palsson BO. Next-generation genome-scale models for metabolic engineering. Current opinion in biotechnology. 2015; 35C:23–9. doi: 10.1016/j.copbio.2014.12.016 PMID: 25575024.

18. Stein O.: Kluwer Academic Publishers; 2003. xxv, 202 p. p.Bi-level strategies in semi-infinite programming. Boston

19. Gelius-Dietrich G, Desouki AA, Fritzemeier CJ, Lercher MJ. Sybil—efficient constraint-based modelling in R. BMC systems biology. 2013; 7:125. doi: 10.1186/1752-0509-7-125 PMID: 24224957; PubMed Central PMCID: PMC3843580.

20. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Molecular systems biology. 2011; 7:535. doi: 10.1038/msb.2011.65 PMID: 21988831; PubMed Central PMCID: PMC3261703.

21. Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD. A genome-scale metabolic reconstruction of Mycoplasma genitalium, iPS189. PLoS Comput Biol. 2009; 5(2):e1000285. doi: 10.1371/journal.pcbi.1000285 PMID: 19214212; PubMed Central PMCID: PMC2633051.

22. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A. 2006; 103(2):425–30. doi: 10.1073/pnas.0510013103 PMID: 16407165; PubMed Central PMCID: PMC1324956.

23. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et biophysica acta. 1975; 405(2):442–51. PMID: 1180967.

24. Schellenberger J, Park JO, Conrad TM, Palsson BO. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics. 2010; 11:213. doi: 10.1186/1471-2105-11-213 PMID: 20426874; PubMed Central PMCID: PMC2874806.

25. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, et al. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular systems biology. 2007; 3:121. doi: 10.1038/msb4100155 PMID: 17593909; PubMed Central PMCID: PMC1911197.

26. Ben-Menachem G, Himmelreich R, Herrmann R, Aharonowitz Y, Rottem S. The thioredoxin reductase system of mycoplasmas. Microbiology. 1997; 143 (Pt 6):1933–40. doi: 10.1099/00221287-143-6-1933 PMID: 9202470.

27. Jamshidi N, Palsson BO. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC systems biology. 2007; 1:26. doi: 10.1186/1752-0509-1-26 PMID: 17555602; PubMed Central PMCID: PMC1925256.

28. Kralova I, Rigden DJ, Opperdoes FR, Michels PA. Glycerol kinase of Trypanosoma brucei. Cloning, molecular characterization and mutagenesis. Eur J Biochem. 2000; 267(8):2323–33. PMID: 10759857.

29. Balogun EO, Inaoka DK, Shiba T, Kido Y, Tsuge C, Nara T, et al. Molecular basis for the reverse reaction of African human trypanosomes glycerol kinase. Mol Microbiol. 2014; 94(6):1315–29. doi: 10.1111/mmi.12831. WOS:000346655900011. PMID: 25315291

30. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular systems biology. 2006; 2:2006 0008. doi: 10.1038/msb4100050 PMID: 16738554; PubMed Central PMCID: PMCPMC1681482.

31. Yamamoto N, Nakahigashi K, Nakamichi T, Yoshino M, Takai Y, Touda Y, et al. Update on the Keio collection of Escherichia coli single-gene deletion mutants. Molecular systems biology. 2009; 5:335. doi: 10.1038/msb.2009.92 PMID: 20029369; PubMed Central PMCID: PMCPMC2824493.

32. Tisa LS, Adler J. Calcium ions are involved in Escherichia coli chemotaxis. Proc Natl Acad Sci U S A. 1992; 89(24):11804–8. PMID: 1465403; PubMed Central PMCID: PMCPMC50645.

33. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33(17):5691–702. doi: 10.1093/nar/gki866 PMID: 16214803; PubMed Central PMCID: PMCPMC1251668.

34. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng. 2003; 84(6):647–57. doi: 10.1002/bit.10803 PMID: 14595777.

35. Hansen P J B.; Savard G;. New branch-and-bound rules for linear bilevel programming. SIAM Journal on Scientific and Statistical Computing 1992; 13(5):1194–217.

36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

37. Schellenberger J, Que R, Fleming RM, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature protocols. 2011; 6 (9):1290–307. doi: 10.1038/nprot.2011.308 PMID: 21886097; PubMed Central PMCID: PMCPMC3319681.

6.2 *Manuscript 2*: Automated high-quality reconstruction of metabolic networks from high-throughput data

6.2.1 Details

Authors:	Daniel Hartleb, Martin J. Lercher
Authorship:	1. Author
Journal:	PNAS
Impact Factor:	9.4
Status:	Submitted to PNAS, rejected after peer review; currently under preparation for resubmission

6.2.2 Contributions

DH conceived and designed the experiments, developed and implemented the software, performed the experiments, and analyzed the data. DH and MJL interpreted the results. DH wrote a first draft of the manuscript, which was then refined iteratively with MJL.

Classification: Biological Sciences – Systems Biology

Automated high-quality reconstruction of metabolic networks from high-throughput data

Daniel Hartleb[1] & Martin J. Lercher[1]*

[1] Institute for Computer Science and Cluster of Excellence on Plant Sciences, Heinrich Heine University, Universitätsstraße 1, D-40225 Düsseldorf, Germany.

* To whom correspondence should be addressed.

Short title: Automated reconstruction of metabolic networks

Corresponding author:

Martin J. Lercher, martin.lercher@hhu.de, +49 151 22964073

Institute for Computer Science, Heinrich Heine University, Universitätsstraße 1, D-40225 Düsseldorf, Germany.

Keywords: Flux Balance Analysis; FBA; GLOBALFIT; Metabolic modelling; *Streptococcus 1-3*; metabolic reconstruction

Abstract

While new genomes are sequenced at ever increasing rates, their phenotypic analysis remains a major bottleneck of biomedical research. The generation of genome-scale metabolic models capable of accurate phenotypic predictions is a labor-intensive endeavor; accordingly, such models are available for only a small percentage of sequenced species. The standard metabolic reconstruction process starts from a (semi-)automatically generated draft model, which is then refined through extensive manual curation. Here, we present a novel strategy suitable for full automation, which exploits high-throughput gene knockout or nutritional growth data. We test this strategy by reconstructing accurate genome-scale metabolic models for three strains of *Streptococcus*, a major human pathogen. The resulting models contain a lower proportion of reactions unsupported by genomic evidence than the most widely used *E. coli* model, but reach the same accuracy in terms of knockout prediction. We confirm the models' predictive power by analyzing experimental data for auxotrophy, additional nutritional environments, and double gene knockouts, and we generate a list of potential drug targets. Our results demonstrate the feasibility of reconstructing high-quality genome-scale metabolic models from high-throughput data, a strategy that promises to massively accelerate the exploration of metabolic phenotypes.

Significance statement

Reading bacterial genomes has become a cheap, standard laboratory procedure. A genome by itself, however, is of little information value – we need a way to translate its abstract letter sequence into a model that describes the capabilities of its carrier. Until now, this endeavor required months of manual work by experts. Here, we show how this process can be automated by utilizing high-throughput experimental data. We use our novel strategy to generate highly accurate metabolic models for three strains of *Streptococcus*, a major threat to human health.

Introduction

Genome-scale metabolic models have been reconstructed for a wide range of organisms (1); they have been used successfully to predict metabolic phenotypes of prokaryotic and eukaryotic cells, in applications ranging from evolutionary studies (2-4) to bioengineering designs (5). An initial draft reconstruction is typically created by mapping each gene of the organism of interest to a closely related, well-annotated relative or to a database of known metabolic functions (6), a task that can easily be automated. These draft models are often unable to produce biomass. Gap-filling methods are applied to ensure functionality of the networks, adding reactions without genomic evidence (7).

The currently most widely applied methodology to generate a reliable genome-scale metabolic network can be considered a hybrid strategy: it starts from an automatically generated draft model based on sequence similarities, followed by manual refinement of reaction content and gene-reaction mapping (6). This mode of metabolic network reconstruction is still time-consuming and laborious. To accelerate the development of metabolic models, automatic and semi-automatic algorithms have been developed (8-10). However, the resulting metabolic networks are typically unreliable and require extensive manual curation (6). For example, reactions may be missing or were added without evidence to facilitate biomass formation. Furthermore, reactions may erroneously be assumed reversible or may have been assigned to the wrong compartment, gene-reaction associations may be incorrect, or biomass components may be missing (11). Accordingly, growth/non-growth predictions for gene knockout strains from automated reconstructions typically show Matthews correlation coefficients R_M<0.5, while extensively manually curated models often agree much better with experimental data (*e.g.*, R_M=0.67 for the iJO1366 *E. coli* model (12), where R_M=1 would correspond to a perfect prediction).

Automatic reconstructions suffer from being based on only one or a few nutritional environments in which the organism is supposed to grow. They lack the capability to utilize – equally important – information about environments in which the investigated organism

cannot grow. Moreover, automatic reconstruction tools possess no mechanism to automatically incorporate genome-wide knock-out data, although such datasets are usually highly informative and thus might lead to more accurate automatic reconstructions.

Here, we develop a pipeline for the automated reconstruction of high-quality genome-scale metabolic networks. We submit an automatically generated draft network to a bi-level optimization algorithm that minimizes the deviation between model predictions and experimental growth/non-growth data from sets of gene knockout strains and/or nutritional environments. We apply our method to three species of *Streptococcus*, gram-positive bacteria that pose a sever risk to human health. *Streptococcus pyogenes* is a Lancefield group A streptococcus (GAS) (13), and is one of the two clinically most important human pathogens (14). *S. pyogenes* are responsible for at least 616 million cases of throat infection (pharyngitis, tonsillitis) worldwide per year, and 111 million cases of skin infection (primarily non-bullous impetigo) in children of less developed countries (15). *Streptococcus sanguinis* (formerly known as *S. sanguis*) is categorized as Lancefield group H (16). These oral bacteria can enter the bloodstream and may cause severe endocarditis (17, 18). Finally, *Streptococcus agalactiae* is a Lancefield group B bacterium, and is one of the major causes of pneumonia and meningitis in neonates (19, 20).

Results

Automated high-quality metabolic reconstruction strategy

Analogous to the hybrid strategy for metabolic network reconstruction currently widely employed, we start with the automated construction of a draft model based on sequence similarity. We do this by successively submitting the gene sequences of the organism under study to different annotation sources that associate them with metabolic functions (Figure 1): (i) KBase (which implements the functionality of Model SEED (8)); (ii) the existing metabolic model of *Lactococcus lactis* (21), a close relative of Streptococci; (iii) TransportDB (22); and (iv) KEGG/KAAS (23). We ordered these information sources by increasing reliability;

accordingly, at each step, information on metabolic gene functions from previous steps is superseded by newer information.

In contrast to the standard hybrid automatic/manual strategy, we continue with an automated refinement of the resulting draft model, informed through comparisons to viability data for gene knockouts and/or nutritional environments (Figure 1). For the application to *Streptococcus*, we only used gene knockout data. We identified false viability predictions of the draft model for individual gene knockouts using flux balance analysis (FBA) (24) with the biomass reaction provided by the *L. lactis* metabolic model. FPp (false positive predictions) are cases where the *in silico* gene knockout simulation predicted growth, while the *in vivo* experiment showed no growth. FNp (false negative predictions) are cases where the *in silico* analysis predicted no growth, while the corresponding knockout strain was viable in the experiment.

The automated refinement was carried out using the GLOBALFIT algorithm, which was originally developed for the further reconciliation of high-quality metabolic network reconstructions with experimental viability data (12). GLOBALFIT is a bi-level optimization program that identifies smallest sets of network modifications in order to minimize the number of FNp and FPp cases; allowed network modifications are (i) removals or (ii) reversibility changes of existing reactions; (iii) additions of reactions to the model from a database of potential reactions; (iv) removals of metabolites from the biomass; and (v) additions of metabolites to the biomass.

The strength of GLOBALFIT in comparison to previous approaches is its ability to consider several growth and/or non-growth cases simultaneously. In particular, when considering FPp, it is important to simultaneously consider a true growth case to avoid trivial solutions such as the removal of an essential reaction (12).

Generation of high-quality metabolic models for streptococci

To build the *Streptococcus* metabolic reconstructions, we initially solved each FPp simultaneously with a wild type growth case and each FNp simultaneously with a non-growth

case. If the network changes suggested by GLOBALFIT introduced new errors by converting true positive predictions (TPp) to FNp, or true negative predictions (TNp) to FPp, we solved the examined case again, this time simultaneously with all distorted cases.

To create a conservative set of potential additional enzymatic reactions, we first inferred homologs between each *Streptococcus* genome and the genes included in metabolic reconstructions in the BiGG database (25). Genes were considered homologous if bi-directional BLAST searches associated the genes with e-values $<10^{-13}$. We only allowed the addition of a reaction to the *Streptococcus* model if homologs to genes sufficient to catalyse the reaction in one of the BiGG models were present in the *Streptococcus* genome.

We allowed the potential reversal of irreversible reactions only for those reactions that were classified at least as "reversible with uncertainty" in the *E. coli* metabolic network (26). Biomass reaction changes (addition or removal of biomass components) were only introduced if no other model changes could rescue the FNp or FPp. The quantitative contribution of each biomass component to the total biomass will have to be manually adjusted to allow quantitative predictions of biomass yield, especially for biomass components added during the network refinement with GLOBALFIT.

The resulting metabolic network for *S. pyogenes* contains 653 metabolites and 661 reactions, accounting for 455 genes. The predicted network for *S. sanguinis* is substantially larger, containing 805 metabolites and 840 reactions, corresponding to 597 genes. The *S. agalactiae* model encompasses 653 metabolites and 661 reactions, accounting for 455 genes. Because streptococci are gram positive bacteria, each genome-scale metabolic network contains only two compartments: extracellular and cytosolic.

The largest set of FPp in all networks is due to the experimentally observed essentiality of the F-ATPase complex, which is encoded by 8 genes and is part of the respiratory chain of many organisms. *S. sanguinis, S. pyogenes*, and *S. agalactiae* lack a respiratory chain, and do not use this enzyme to produce ATP; instead, the F-ATPase consumes ATP to pump protons out of the cell to maintain an internal pH that is more basic than the exterior (27).

Thus, the F-ATPase does not have any metabolic function that can be accounted for in FBA models, explaining why even a perfect FBA representation of *Streptococcus* metabolism cannot predict the essentiality of the corresponding genes.

The networks predict viability of gene knockouts with high accuracy (Table 1). On average, we obtain an accuracy (percentage of true predictions) of 94.8%. Matthew's correlation coefficient (28), a more balanced measure of prediction quality, is on average R_M=0.85. When discounting the genes involved in the F-ATPase complex (which cannot be predicted correctly in any FBA model), average accuracy increases to 96.2%, with R_M=0.89.

Accuracy of the automatically generated Streptococcus models

To further investigate the metabolic capabilities of the three strains, we used GLOBALFIT to predict a minimal medium. The minimal medium for *S. sanguinis* contains nine nutrients, while *S. pyogenes* and *S. agalactiae* require a minimum of 22 metabolites, mainly because the latter two species require a larger number of externally supplied amino acids. To explore this issue further, we used GLOBALFIT to predict all amino acids that cannot be produced from a minimal medium from which all amino acids were removed; for *S. pyogenes* and *S. sanguinis*, this reduced minimal medium consisted of the eight nutrients glucose, phosphor, sulphur, iron, pyridoxal, niacin, riboflavin, and pantothenate. *S. agalactiae* additionally required thiamine. Our simulations on these reduced minimal media showed that *S. sanguinis* is only auxotrophic for Cysteine, while *S. pyogenes* and *S. agalactiae* are auxotrophic for 12 and 11 amino acids, respectively; in addition, both species require at least two out of four further amino acids (See Supplementary Table S1). In comparison, *L. lactis* is auxotrophic for Leucine, Histidine, and Methionine (21). These observations are consistent with previous experimental studies (29-31), confirming the reliability of the reconstructed metabolic networks.

We further tested the *S. sanguinis* model by comparing its predictions to growth experiments on different defined media. The metabolic model constructed by GLOBALFIT successfully predicted growth on B 48 (32) as well as on SY and M3 media (33). The study of Rogers also

provided growth information on a set of SY and M3 media lacking one metabolite (either riboflavin, pantheonate, thiamine, nicotinic acid, pyridoxal, folic acid, aminobenzoic acid, or biotin). We successfully predicted the essentiality and non-essentiality of these metabolites, except for the essentiality of pyridoxal and riboflavin. We added these components to the biomass objective function; if the growth environments had been included in the original GLOBALFIT run, this addition would have occurred automatically. Pyridoxal and riboflavin were also part of the biomass reaction of the alternative network reconstructed by KBase (see below).

In a previous study, knockout mutants for three *S. sanguinis* metabolic enzymes were unexpectedly found to be viable (34). For each of these genes, Xu *et al.* identified isozymes or paralogs in the *S. sanguinis* genome. The corresponding double-gene knockouts were unviable for two of the three reactions, but were unexpectedly viable for one enzyme (NAD(P)H-dependent glycerol-3-phosphate dehydrogenase). While we did not include these double knockouts in the training set used to derive the *S. sanguinis* network, all three double knockouts were correctly predicted by our model.

To benchmark the predictive power of our reconstructed metabolic models, we compared them to the network automatically reconstructed from the genome sequence by KBase. The KBase model for *S. agalactiae* was not able to grow anaerobically and was incapable of producing NADP; thus, we allowed the influx of oxygen and NADP for the corresponding simulations. All three networks were substantially less accurate in predicting gene knock-outs (accuracy ≤79%, R_M ≤0.42 for metabolic reactions; Table 1 and Supplementary Table S2). The F-ATP-synthase complex was not included in any metabolic network by KBase.

Could the superior performance of the GLOBALFIT models be due to overfitting to the gene knockout data? To test this, we also compared amino acid auxotrophy (31-33, 35), which was not used by either network reconstruction approach. While GLOBALFIT predictions were fully consistent with the experimental results, the KBase model required more amino acids than experimentally observed (Supplementary Table S3). For example, KBase predicted *S.*

sanguinis to be auxotrophic for asparagine, cysteine, and threonine, while only the requirement for cysteine was experimentally observed.

Prediction of potential metabolic drug targets

Streptococci have acquired resistance to major antibiotics which can make a successful treatment difficult (36, 37). Thus, new antibiotics are needed. As a starting point and to accelerate the discovery of new drugs metabolite essentiality analysis can be used (38). This approach removes *in-silico* each metabolite from the metabolic model. If the removal of the metabolites prevents the formation of biomass, it is deemed essential. Because this leads to a large list of metabolites with many unlikely candidates (e.g., currency metabolites such as ATP), subsequent filtering steps are needed.

To minimize potential side effects, we first removed all metabolites that also occur in the human metabolic model (39). In addition, we BLASTed the genes of all reactions that are associated with one of the essential metabolites against the human genome; if a distant homolog (e-value < 0.01) for at least one gene was found, the according metabolite was also discarded from further analysis.

We applied this approach to the three reconstructed *Streptococcus* metabolic networks, identifying 10 different essential metabolites likely not involved in human metabolism (Supplementary Table S4). These drug target candidates are processed in three different pathways. The reliability of our analysis is demonstrated by the prediction of PABA (4-Aminobenzoate) as a potential drug target: this substance is already targeted by many sulfonamide antibiotics, which inhibit the dihydropteroate synthase. This enzyme is essential in bacteria for producing folate, while human acquire folate as part of their nutrition.

Conclusions

At >95% (Table 1), the accuracy of gene knockout predictions from our *Streptococcus* models exceeds that of the most intensely curated other bacterial metabolic network, the iJO1366 model for *E. coli* (90.8%) (40). Despite this high accuracy, our model

reconstructions are more conservative than those of many manually curated or automatic metabolic models: at most 4% of enzymatic reactions are not associated with a gene product (Supplementary Table 5), while the corresponding number is 6% for the iJO1366 *E. coli* model and >7% for the KBase *Streptococcus* models. Less than 2% of enzymatic reactions in our models lacked an Enzyme Commission (EC) number, whereas this is the case for >7% of reactions in the KBase models and for more than one third of enzymatic reactions in the iJO1366 model (note that the KBase models did not contain any EC numbers; we obtained these values by mapping the KBase reactions to the ModelSEED Database (8)). We conclude that the general approach demonstrated here opens up the prospect to fully automate the reconstruction of high-quality genome-scale metabolic models. The reconstruction quality could be increased even further if multiple sets of growth data are employed, *e.g.*, by including high-throughput phenotyping data (41).

Methods

We devised a pipeline for metabolic network reconstruction suitable for full automation. This algorithm examines different sources of information in ascending order of reliability; at each step, information from the previous step is refined and overwritten.

Base model reconstruction

We started by uploading the three genome sequences of *S. pyogenes* NZ131 (42), *S. sanguinis* SK36 (43), and *S. agalactiae* (44) to KBase (45). KBase outputs a first draft metabolic model based on sequence similarities of the genes to its database, containing reaction IDs and associated Boolean gene-protein-reaction associations (GPR rules). We removed reactions if their GPR was invalid, *i.e.*, if the reaction required at least one unknown gene to be active. The remaining reactions were carried over to the next step only if the KBase reaction abbreviation was identical to a reaction ID in the BiGG database (25).

We updated this first draft with information from the existing metabolic model of *Lactococcus lactis* (21), a close relative of Streptococci, which uses the metabolite and reaction nomenclature of the BiGG database. For each gene in the *L. lactis* genome (46), we identified homologs in the *Streptococcus* genomes by identifying reciprocal BLAST hits with e-values $<10^{-13}$. In some cases, this strategy resulted in the mapping of several paralogous genes to one gene annotated in the *L. lactis* genome (see below). Overall, we found *L. lactis* homologs for 62% of *S. pyogenes*, 55% of *S. sanguinis*, and 61% of *S. agalactiae* genes.

We added or updated reactions of the initial KBase model with *L. lactis* reactions if the corresponding GPR could be fulfilled with *Streptococcus* genes that had a reciprocal BLAST hit with e-value $<10^{-13}$ between the two genomes. If a reaction in the *L. lactis* model was not associated with any gene, the corresponding (empty) GPR was also considered valid. We also included the biomass reaction of *L. lactis* into the base models. We adapted the relative abundance of the nucleotides to the G+C content observed in the *Streptococcus* genomes

(*L. lactis*: G+C=35.8%, *S. sanguinis*: G+C=43.4%, *S. pyogenes*: G+C=38.5%, *S. agalactiae*: G+C=35.6%).

Obtaining transport reactions from TransportDB

We replaced GPRs for transport reactions and added transport reactions with predictions from TransportDB (22). Transport reactions with empty GPRs that could not be filled through TransportDB were retained. Note that predicting transporters is still challenging and is an important source of inaccuracies in metabolic network reconstructions (6).

Initial model curation using KAAS

For some reactions, the homolog prediction through bidirectional blast hits resulted in GPRs with unrealistically large paralogous gene sets. Furthermore, some reactions in the *L. lactis* template metabolic network or in the KBase predictions may have erroneous GPR associations. To reduce these two sources of error, we refined each reaction in the extended base model by comparing the GPR with the corresponding gene function predictions made by the KEGG automated annotation server (KAAS) (23). We first re-annotated all genes in the *Streptococci* genomes using KAAS, which associated metabolic genes recognized by KAAS with EC numbers and KEGG reaction identifiers. Using KEGG (47), we additionally associated each reaction in the extended base model with an EC number and a KEGG reaction identifier. We added genes to the GPR of all reactions with the same EC number as that assigned by KAAS to the gene, and we removed genes from GPRs if the gene's EC number according to KAAS differed from that of the reaction. If KEGG contained information on protein complexes, we introduced that information into the corresponding GPRs. We dropped reactions that lost a valid GPR association in this process.

For example, the *L. lactis* metabolic network reconstruction contains a coproporphyrinogen oxidase (CPPPGO, EC: 1.3.3.3), which requires oxygen and is associated with the *L. lactis* gene iNF518 (HemN). The KAAS annotation revealed that this gene instead encodes an oxygen-independent coproporphyrinogen-III oxidase (CPPPGO2, EC: 1.3.99.22) (48). The oxygen dependent coproporphyrinogen oxidase erroneously included in the *L. lactis* model is

catalyzed by HemF, which is present neither in the genome of anaerobically living *Streptococci* nor in *L. lactis*.

ATP maintenance reactions

The non-growth associated maintenance reaction (NGAM) accounts for ATP requirements not directly related to cell growth (*e.g.,* maintenance of turgor pressure); conversely, the growth associated maintenance reaction (GAM) accounts for energy requirements directly related to cell replication (*e.g.,* synthesis of proteins, DNA, and RNA) and is part of the biomass reaction. Appropriate rates for these ATP maintenance reactions are usually determined by growth experiments (6). Because such experiments were not available for *S. sanguinis, S. pyogenes*, or *S. agalactiae*, we set the lower bounds of the according reactions to the values appropriate for *L. lactis* (GAM: 39 mM/g_{DW}/h; NGAM: 0.92 mM/g_{DW}/h) (21).

Simulated environments

The resulting draft models for *S. sanguinis* and *S. pyogenes* were used as the basis for further refinement through high-throughput gene knockout data. All analyzed gene knockout studies were performed anaerobically on undefined rich media (brain heart infusion for *S. sanguinis* (34), Todd-Hewitt Yeast medium for *S. pyogenes* (49), and TS media for *S. agalactiae* (50)). We thus allowed the uptake of all nutrients for which a transport reaction was included in the curated base model except for oxygen, constraining the lower bound of the oxygen exchange reaction to zero and the lower bound of all other exchange reactions to -5 mM/g_{DW}/h.

Identification of FPp and FNp

Using this set of parameters, we performed flux balance analysis (FBA) (24) with the biomass reaction provided by the *L. lactis* metabolic model. We identified FPp (false positive predictions) as those cases where our *in-silico* gene knockout simulation predicted growth ($v_{biomass}$>0), while the *in vivo* experiment showed no growth. Correspondingly, we identified FNp (false negative predictions) as those cases where the *in silico* gene-knockout analysis predicted no growth, while the knockout was viable in the experiment. Because no lower

threshold for growth was used in the gene knockout experiments, we also interpreted any flux through the biomass reaction $>10^{-9}$ as *in silico* growth.

Automated network refinement with GLOBALFIT

The draft networks were refined with gene knockout libraries using the previously published GLOBALFIT algorithm (12). GLOBALFIT is a bi-level optimization program that identifies smallest sets of network modifications in order to minimize the number of FNp and FPp cases; allowed network modifications are (i) removals or (ii) reversibility changes of existing reactions; (iii) additions of reactions to the model from a database of potential reactions; (iv) removals of metabolites from the biomass; and (v) additions of metabolites to the biomass.

The strength of GLOBALFIT is its ability to consider several growth or non-growth cases simultaneously. In particular, when considering FPp, it is important to simultaneously consider a true growth case to avoid trivial solutions such as the removal of an essential reaction(12). We initially solved each FPp simultaneously with a wild type growth case and each FNp simultaneously with a non-growth case. If the network changes suggested by GLOBALFIT introduced new errors by converting true positive predictions (TPp) to FNp, or true negative predictions (TNp) to FPp, we solved the examined case again, this time simultaneously with all cases converted to false predictions by the original modification.

Allowed model changes

For each network, we created a conservative set of potential additional reactions by blasting the corresponding genome against all genes annotated in networks contained in the BiGG database (25). As before, we identified homologous genes as bidirectional blast hits with e-values $<10^{-13}$. All reactions with a valid GPR of homologous genes was added to the set of potential additional reactions. Streptococci are gram positive bacteria; we thus removed all reactions that do not occur in the cytosol or the extracellular space of their original metabolic model. We allowed the potential reversal of irreversible reactions only for those reactions that were classified at least as "reversible with uncertainty" in the *E. coli* metabolic network (26).

Biomass reaction changes (addition or removal of biomass components) suggested by GLOBALFIT were manually compared to the literature and the biomass reaction generated by KBase; a fully automated approach could instead use a preformed database of potential biomass components. If a potential additional biomass component was part of the biomass reaction of the KBase model, the stoichiometric coefficient was carried over; otherwise, it was arbitrarily set to 10^{-5}. No growth experiments for *Streptococci* were available to refine the coefficients. However, the exact coefficient of each biomass component is less important than its presence for gene knockout analyses (6), and corresponding inaccuracies should not significantly bias our results. However, to allow quantitative predictions of biomass production rates, the biomass stoichiometry of the final networks will require additional curation.

Removal of exchange reactions

GLOBALFIT suggested the removal of several exchange reactions. This may indicate the absence of the corresponding transporters from the *Streptococcus* genome. However, most of these removals may simply indicate that the corresponding nutrient was missing from the undefined growth medium (12). Supplementary Table S6 lists the exchange reactions removed for each *Streptococcus* strain.

Growth on chemically defined media for S. sanguinis

The metabolic network of *S. sanguinis* was further refined by using the gene knockout library performed on a chemically defined medium (34). In contrast to the knockout experiments on undefined media, the corresponding publication lists knockout growth rates rather than simply stating growth or non-growth. For both experiment and simulations, we considered growth rates >80% of the wildtype as growth cases, and growth rates <10% of the wildtype as non-growth cases. The metabolic network of *S. sanguinis* was then further refined using the same strategy as described above.

We could successfully predict growth of the *S. sanguinis* metabolic network on B 48 (32), SY and M3 media (33). The study of Rogers also provided growth information in media lacking

one specific metabolite (riboflavin, pantheonate, thiamine, nicotinic acid, pyridoxal, folic acid, aminobenzoic acid and biotin). The *S. sanguinis* network successfully predicted the essentiality and non-essentiality of these metabolites, except the essentiality of pyridoxal and riboflavin. Therefore, we added these components to the biomass objective function. These metabolites were also part of the biomass reaction of the network reconstructed by the KBase.

The genome of *S. sanguinis* contains a gene cluster that allows the organism to produce cobalamin (vitamin B12) anaerobically; this region was presumably acquired via horizontal gene transfer (43). 10 reactions from this pathway are missing from the draft genome reconstruction. However, as cobalamin is not part of the biomass reaction and all enzymes involved in cobalamin production are non-essential, GLOBALFIT did not attempt to complete the pathway. We used GLOBALFIT to identify a minimal set of missing reactions for cobalamin production by using a growth case with a cobalamin-consuming reaction as the objective function and allowing the addition of all reactions in the BiGG database. The predicted complete pathway was identical to the one in the metabolic network of *Clostridium ljungdahlii* (51) and *Geobacter metallireducens* (52). This manual refinement did not affect the knock-out predictions.

Acknowledgements

We acknowledge financial support through the German Research Foundation DFG (GRK 1525 supporting DH; EXC 1028 and CRC 680 to MJL).

Author contributions

DH conceived of the study and performed all analyses. MJL guided the study. DH and MJL wrote and edited the manuscript.

References

1. O'Brien EJ, Monk JM, & Palsson BO (2015) Using Genome-scale Models to Predict Biological Capabilities. *Cell* 161(5):971-987.
2. Ibarra RU, Edwards JS, & Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420(6912):186-189.
3. Pal C, Papp B, & Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37(12):1372-1375.
4. Pal C, *et al.* (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440(7084):667-670.
5. Lee JW, Kim TY, Jang YS, Choi S, & Lee SY (2011) Systems metabolic engineering for chemicals and materials. *Trends Biotechnol* 29(8):370-378.
6. Thiele I & Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 5(1):93-121.
7. Reed JL, *et al.* (2006) Systems approach to refining genome annotation. *Proc Natl Acad Sci U S A* 103(46):17480-17484.
8. Overbeek R, *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33(17):5691-5702.
9. Latendresse M, Krummenacker M, Trupp M, & Karp PD (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics* 28(3):388-396.
10. Cuevas DA, *et al.* (2016) From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model. *Front Microbiol* 7.
11. Hamilton JJ & Reed JL (2014) Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environ Microbiol* 16(1):49-59.
12. Hartleb D, Jarre F, & Lercher MJ (2016) Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. *PLoS Comput Biol* 12(8):e1005036.
13. Cunningham MW (2000) Pathogenesis of group A streptococcal infections. *Clin Microbiol Rev* 13(3):470-511.
14. Bessen DE (2009) Population biology of the human restricted pathogen, Streptococcus pyogenes. *Infect Genet Evol* 9(4):581-593.
15. Carapetis JR, Steer AC, Mulholland EK, & Weber M (2005) The global burden of group A streptococcal diseases. *Lancet Infect Dis* 5(11):685-694.
16. Bridge PD & Sneath PH (1983) Numerical taxonomy of Streptococcus. *J Gen Microbiol* 129(3):565-597.
17. Mylonakis E & Calderwood SB (2001) Infective endocarditis in adults. *N Engl J Med* 345(18):1318-1330.
18. Ahmed R, Hassall T, Morland B, & Gray J (2003) Viridans streptococcus bacteremia in children on chemotherapy for cancer: an underestimated problem. *Pediatr Hematol Oncol* 20(6):439-444.
19. Liu GJ, Zhang W, & Lu CP (2013) Comparative genomics analysis of Streptococcus agalactiae reveals that isolates from cultured tilapia in China are closely related to the human strain A909. *Bmc Genomics* 14.
20. Campbell JR, *et al.* (2000) Group B streptococcal colonization and serotype-specific immunity in pregnant women at delivery. *Obstet Gynecol* 96(4):498-503.
21. Flahaut NA, *et al.* (2013) Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. *Appl Microbiol Biotechnol* 97(19):8729-8739.
22. Ren Q, Chen K, & Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 35(Database issue):D274-279.

23. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, & Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(Web Server issue):W182-185.
24. Orth JD, Thiele I, & Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245-248.
25. King ZA, *et al.* (2015) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*.
26. Feist AM, *et al.* (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.
27. Kuhnert WL & Quivey Jr RG, Jr. (2003) Genetic and biochemical characterization of the F-ATPase operon from Streptococcus sanguis 10904. *J Bacteriol* 185(5):1525-1533.
28. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405(2):442-451.
29. Milligan TW, Doran TI, Straus DC, & Mattingly SJ (1978) Growth and Amino-Acid Requirements of Various Strains of Group-B Streptococci. *J Clin Microbiol* 7(1):28-33.
30. Mickelson MN (1964) Chemically Defined Medium for Growth of Streptococcus Pyogenes. *Journal of Bacteriology* 88(1):158-&.
31. Carlsson J (1972) Nutritional requirements of Streptococcus sanguis. *Arch Oral Biol* 17(9):1327-1332.
32. Carlsson J (1971) Growth of Streptococcus-Mutans and Streptococcus-Sanguis in Mixed Culture. *Archives of Oral Biology* 16(8):963-&.
33. Rogers AH (1973) Vitamin Requirements of Some Oral Streptococci. *Archives of Oral Biology* 18(2):227-232.
34. Xu P, *et al.* (2011) Genome-wide essential gene identification in Streptococcus sanguinis. *Sci Rep* 1:125.
35. Willett NP, Morse GE, & Carlisle SA (1967) Requirements for Growth of Streptococcus Agalactiae in a Chemically Defined Medium. *Journal of Bacteriology* 94(4):1247-&.
36. Mousavi SM, Nasaj M, Hosseini SM, & Arabestani MR (2016) Survey of strain distribution and antibiotic resistance pattern of group B streptococci (Streptococcus agalactiae) isolated from clinical specimens. *GMS Hyg Infect Control* 11:Doc18.
37. Passali D, Lauriello M, Passali GC, Passali FM, & Bellussi L (2007) Group A streptococcus and its antibiotic resistance. *Acta Otorhinolaryngol Ital* 27(1):27-32.
38. Kim HU, Kim TY, & Lee SY (2010) Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen Acinetobacter baumannii AYE. *Mol Biosyst* 6(2):339-348.
39. Duarte NC, *et al.* (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* 104(6):1777-1782.
40. Orth JD, *et al.* (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol Syst Biol* 7:535.
41. Miller JM & Rhoden DL (1991) Preliminary Evaluation of Biolog, a Carbon Source Utilization Method for Bacterial Identification. *J Clin Microbiol* 29(6):1143-1147.
42. McShan WM, *et al.* (2008) Genome sequence of a nephritogenic and highly transformable M49 strain of Streptococcus pyogenes. *J Bacteriol* 190(23):7773-7785.
43. Xu P, *et al.* (2007) Genome of the opportunistic pathogen Streptococcus sanguinis. *J Bacteriol* 189(8):3166-3175.
44. Tettelin H, *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial 'pan-genome' (vol 102, pg 13950, 2005). *P Natl Acad Sci USA* 102(45):16530-16530.
45. Knowledgebase DoESB (2016) (KBase).
46. Wegmann U, *et al.* (2007) Complete genome sequence of the prototype lactic acid bacterium Lactococcus lactis subsp. cremoris MG1363. *J Bacteriol* 189(8):3256-3270.

47. Kanehisa M, *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42(Database issue):D199-205.
48. Homuth G, Heinemann M, Zuber U, & Schumann W (1996) The genes of lepA and hemN form a bicistronic operon in Bacillus subtilis. *Microbiology* 142 (Pt 7):1641-1649.
49. Le Breton Y, *et al.* (2015) Essential Genes in the Core Genome of the Human Pathogen Streptococcus pyogenes. *Sci Rep* 5:9838.
50. Hooven TA, *et al.* (2016) The essential genome of Streptococcus agalactiae. *Bmc Genomics* 17.
51. Nagarajan H, *et al.* (2013) Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of Clostridium ljungdahlii. *Microb Cell Fact* 12:118.
52. Feist AM, *et al.* (2014) Constraint-based modeling of carbon fixation and the energetics of electron transfer in Geobacter metallireducens. *PLoS Comput Biol* 10(4):e1003575.

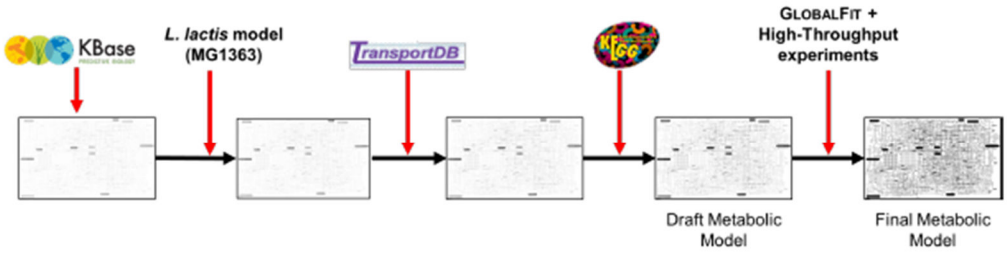

Figure 1. Automated workflow used for the reconstruction of the *Streptococcus* metabolic models. Information added at each step supersedes information from previous steps.

Table 1. Comparison of experimental and predicted viability for Streptococci gene knockouts

Predictions		Experiment growth	Experiment non-growth	GlobalFit Accuracy[1]	GlobalFit R_M [2]	KBase Accuracy[3]	KBase R_M [4]
S. sanguinis	growth	479	17	95.3%	0.83	78.6%	0.33
	no growth	12	89	(96.6%)	(0.88)		
S. pyogenes	growth	283	15	95.4%	0.88	78.1%	0.42
	no growth	3	90	(96.6%)	(0.91)		
S. agalactiae	growth	304	17	93.8%	0.84	77.6%	0.39
	no growth	10	103	(95.3%)	(0.88)		

[1] percentage of correct viability predictions. Values in parentheses are calculated excluding the F-ATPase complex.

[2] Matthew's correlation coefficient (28). Values in parentheses are calculated excluding the F-ATPase complex.

[3] percentage of correct viability predictions (excluding genes associated with non-metabolic generic reactions, *i.e.*, protein biosynthesis, DNA replication, and RNA transcription).

[4] Matthew's correlation coefficient (28) (excluding genes associated with non-metabolic generic reactions, *i.e.*, protein biosynthesis, DNA replication, and RNA transcription).

Supplementary Tables

Supplementary Table S1. Essential amino acids

S. sanguinis	***S. pyogenes***	***S. agalactiae***
Cysteine	Arginine	Arginine
	Cysteine	Cysteine
	Histidine	Histidine
	Isoleucine	Isoleucine
	Leucine	Leucine
	Lysine	Lysine
	Methionine	Methionine
	Phenylalanine	Phenylalanine
	Threonine	Tryptophan
	Tryptophan	Tyrosine
	Tyrosine	Valine
	Valine	
		Glutamine or Glutamate
	Glutamine or Glutamate	Glycine or Serine
	Glycine or Serine	

Supplementary Table S2. Comparison of experimental and predicted viability for Streptococci gene knockouts with metabolic models reconstructed by KBase.

Predictions		Experiment growth	Experiment non-growth	Accuracy[1]	R_M ([2])
S. sanguinis	growth	390	46	78.6%	0.33
	no growth	98	110	(77.6%)	(0.46)
S. pyogenes	growth	233	44	78.1%	0.42
	no growth	44	89	(78.5%)	(0.51)
S. agalactiae[3]	growth	335	63	77.6%	0.39
	no growth	52	69	(78.7%)	(0.50)

[1] percentage of correct viability predictions. Values in parentheses are calculated including genes associated with non-metabolic generic reactions (*i.e.*, protein biosynthesis, DNA replication, and RNA transcription).

[2] Matthew's correlation coefficient (28). Values in parentheses are calculated including genes associated with non-metabolic generic reactions (*i.e.*, protein biosynthesis, DNA replication, and RNA transcription).

[3] S. agalactiae could not grow anaerobically, thus we allowed the influx of oxygen

Supplementary Table S3. Essential amino acids for KBase models[1]

S. sanguinis	*S. pyogenes*	*S. agalactiae*
Glycine-Asparagine	Alanine-Glutamine or Alanine-Aspartate or Alanine-Glutamate or Alanylglycine or Alanine-Leucine or Alanine-Histidine or Alanine-Threonin	Alanine-Glutamine or Alanine-Aspartate or Alanine-Glutamate or Alanylglycine or Alanine-Leucine or Alanine-Histidine or Alanine-Threonin
Cysteine or Cysteine-Glycine or Glycine-Cysteine	Arginine	Arginine
Glycine-Tyrosine	Cysteine or Glycine-Cysteine or Cysteine-Glycine	Cysteine or Glycine-Cysteine or Cysteine-Glycine
	Alanine-Histidine	Alanine-Histidine
	Isoleucine	Isoleucine
	Leucine or Glycine-Leucine	Leucine or Glycine-Leucine
	Lysine	Lysine
	Methionine or Methionine-Alanine or Glycine-Methionine	Methionine or Methionine-Alanine or Glycine-Methionine
	Glycine-Phenylalanine	Glycine-Phenylalanine
	Alanine-Threonine	Tryptophan
	Tryptophan	Glycine-Tyrosine
	Glycine-Tyrosine	Valine
	Valine	
		Glutamine or Glycine-Glutamine or Glycine-Glutamate
	Glycine-Asparagine or Glycine-Aspartate	
	Glutamine or Glycine-Glutamine or Glycine-Glutamate	

[1] Note that for several amino acids, the networks automatically reconstructed by KBase possessed only exchange reactions for di-peptides. For example, the observed auxotrophy of glycine or serine for *S. pyogenes* could not be tested, because glycine was already imported as glycine-tyrosine.

Supplementary Table S4 Essential metabolites of the three *Streptococci*

Organism[1]	Metabolite	Reaction	Pathway	E.C. Number
SSA, SAK	**2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine**	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	Folate biosynthesis	2.7.6.3
		dihydroneopterin aldolase	Folate biosynthesis	4.1.2.25
SSA, SAK	**4-Aminobenzoate**	aminodeoxychorismate lyase	Folate biosynthesis	4.1.3.38
		Dihydroperoate synthase	Folate biosynthesis	2.5.1.15
SSA, SAK, SPY	**D-glutamate**	glutamate racemase	D-Glutamine and D-glutamate metabolism	5.1.1.3
		UDP-N-acetylmuramoyl-L-alanine---D-glutamate ligase	Peptidoglycan biosynthesis	6.3.2.9
SPY	**Undecaprenyl-diphospho-N-acetylmuramoyl-L-alanyl-D-glutamyl-L-lysyl-D-alanyl-D-alanine**	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	Peptidoglycan biosynthesis	2.4.1.227
SSA, SAK, SPY	**UDP-N-acetyl-3-O-(1-carboxyvinyl)-D-glucosamine**	UDP-N-acetylenolpyruvoylglucosamine reductase	Peptidoglycan biosynthesis	1.3.1.98
SPY	**UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-L-lysyl-D-alanyl-D-alanine**	phospho-N-acetylmuramoyl-pentapeptide-transferase (alpha-glutamate)	Peptidoglycan biosynthesis	2.7.8.13
SSA, SAK, SPY	**UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-L-lysyl-D-alanyl-D-alanine synthetase (alpha-glutamate)**	UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-L-lysyl-D-alanyl-D-alanine synthetase (alpha-glutamate)	Peptidoglycan biosynthesis	6.3.2.10
SSA, SAK, SPY	**UDP-N-acetylmuramoyl-L-alanine**	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase	Peptidoglycan biosynthesis	6.3.2.9
SSA, SAK, SPY	**UDP-N-acetylmuramoyl-L-alanyl-D-glutamate**	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate:meso-2,6-diaminoheptanedioate ligase	Peptidoglycan biosynthesis	6.3.2.13
		UDP-N-acetylmuramoyl-L-alanyl-D-glutamate---L-lysine ligase	Peptidoglycan biosynthesis	6.3.2.7
SSA, SAK, SPY	**UDP-N-acetylmuramate**	UDP-N-acetylmuramoyl-L-alanine synthetase	Peptidoglycan biosynthesis	6.3.2.8

[1] SSA=*Streptococcus sanguinis*, SAK=*Streptococcus agalactiae*, SPY=*Streptococcus pyogenes*

Supplementary Table S5. Non-transport reactions associated with genes and EC number

	Reactions associated with genes	Reactions with EC Number
S. sanguinis	95.9%	98.4%
S. pyogenes	97.1%	98.0%
S. agalactiae	97.8%	98.2%
***S. sanguinis* KBase Model**	93.0%	92.8%
***S. pyogenes* KBase Model**	87.7%	91.5%
***S. agalactiae* KBase Model**	90.0%	92.3%

Supplementary Table S6. Removal of nutrients from the growth medium suggested by GLOBALFIT

S. sanguinis	***S. pyogenes***	***S. agalactiae***
N-Acetyl-D-Glucosamine	Citrate	Acetaldehyde
Fructose	Malate	N-Acetyl-D-mannosamine
Asparagine[1]	Pyruvate	N-Acetyl-D-Glucosamine
Aspartate[1]	Serine	Dihydroxyacetone
Cysteine[1]	Glycerol	Deoxyribose
Glycine[1]	Fructose	Fructose
Threonine[1]	Mannose	Glycerol 3-phosphate
	N-Acetyl-D-mannosamine	Glycerol
	N-Acetyl-D-Glucosamine	Inosine
		Mannose
		Malate
		Thymidine
		Hypoxanthine
		Xanthine
		Coenzyme A

[1] GLOBALFIT suggested to constrain the corresponding influx to -0.1

6.3 *Manuscript 3*: Erroneous energy generating cycles in published genome-scale metabolic networks: Identification and removal

6.3.1 Details

Authors:	Claus Jonathan Fritzemeier*, Daniel Hartleb, Balázs Szappanos, Balázs Papp, Martin J. Lercher
Authorship:	2nd Author
Journal:	PLOS Computational Biology
Impact Factor:	4.62
Status:	Published: PLoS Comput Biol 13(4): e1005494, DOI: 10.1371/journal.pcbi.1005494

6.3.2 Contributions

DH developed and implemented the modified version of GlobalFit. Together with CJF and MJL, DH designed the experiments, analyzed the data and wrote the paper.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal

Claus Jonathan Fritzemeier[1], Daniel Hartleb[1], Balázs Szappanos[2], Balázs Papp[2], Martin J. Lercher[1]*

1 Institute for Computer Science and Cluster of Excellence on Plant Sciences, Heinrich Heine University, Düsseldorf, Germany, **2** Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Szeged, Hungary

* lercher@cs.uni-duesseldorf.de

OPEN ACCESS

Citation: Fritzemeier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ (2017) Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLoS Comput Biol 13(4): e1005494. https://doi.org/10.1371/journal.pcbi.1005494

Editor: Costas D. Maranas, The Pennsylvania State University, UNITED STATES

Received: October 19, 2016

Accepted: April 1, 2017

Published: April 18, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: We gratefully acknowledge financial support by the DFG (http://www.dfg.de/; grants EXC 1028 and CRC 680 to MJL; GRK 1525 fellowship to DH) and through an ENORM graduate school (http://www.e-norm.hhu.de/) fellowship to CJF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Energy metabolism is central to cellular biology. Thus, genome-scale models of heterotrophic unicellular species must account appropriately for the utilization of external nutrients to synthesize energy metabolites such as ATP. However, metabolic models designed for flux-balance analysis (FBA) may contain thermodynamically impossible energy-generating cycles: without nutrient consumption, these models are still capable of charging energy metabolites (such as ADP→ATP or $NADP^+$→NADPH). Here, we show that energy-generating cycles occur in over 85% of metabolic models without extensive manual curation, such as those contained in the ModelSEED and MetaNetX databases; in contrast, such cycles are rare in the manually curated models of the BiGG database. Energy generating cycles may represent model errors, *e.g.*, erroneous assumptions on reaction reversibilities. Alternatively, part of the cycle may be thermodynamically feasible in one environment, while the remainder is thermodynamically feasible in another environment; as standard FBA does not account for thermodynamics, combining these into an FBA model allows erroneous energy generation. The presence of energy-generating cycles typically inflates maximal biomass production rates by 25%, and may lead to biases in evolutionary simulations. We present efficient computational methods (i) to identify energy generating cycles, using FBA, and (ii) to identify minimal sets of model changes that eliminate them, using a variant of the GlobalFit algorithm.

Author summary

Genome-scale metabolic models are routinely used to simulate the growth of unicellular organisms, and are likely to become an important tool in the medical sciences. The most popular method employed for this task is flux balance analysis (FBA), a simplified mathematical description able to describe the simultaneous activity of hundreds of biochemical reactions. Cellular functions are often dependent on the availability of sufficient energy, and thus a correct representation of energy metabolism appears crucial to metabolic

Competing interests: The authors have declared that no competing interests exist.

modeling. However, we found that the majority of FBA models generated directly from genome sequences, as well as a minority of carefully curated models, are capable of generating energy out of thin air. These models charge energy metabolites such as ATP without any nutrient uptake. We named the corresponding sets of reactions "erroneous energy generating cycles" (EGCs) and developed a high-throughput algorithm for their identification. We found EGCs in 238 (68%) of 350 metabolic models from three different databases. We developed a second, fully automated method for EGC removal. Simulations on the corrected models typically showed growth rates that were 25% slower than in the original models, demonstrating the importance of checking metabolic model reconstructions for EGCs.

Introduction

Constraint-based analysis, in particular flux-balance analysis (FBA), is the current state of the art in genome-scale metabolic modeling [1]. Constraint-based modeling assumes a steady state (*i.e.*, every internal metabolite that is produced must be consumed at the same rate) and imposes lower and upper bounds on metabolic fluxes. However, constraint-based analyses typically do not explicitly consider thermodynamics. As a result, the mathematical solution of constraint-based problems is often thermodynamically infeasible [2, 3]. Specifically, internal cycles (sometimes called type-III pathways [4]), which consist only of internal reactions and do not exchange metabolites with the environment, violate the second law of thermodynamics. The thermodynamic driving forces around a biochemical reaction cycle must add up to zero; hence, there cannot be a flux in a closed cycle [5–7].

These thermodynamically infeasible type-III pathways [4] have to be distinguished from futile cycles (type-II pathways; Fig 1), which additionally consume cofactors to generate a driving force around the cycle [8, 9]. Futile cycles are not an artifact of metabolic modeling, but have been experimentally observed [10]; *e.g.*, some prokaryotes that live in very energy-rich environments need to dissipate energy by converting ATP to ADP [11].

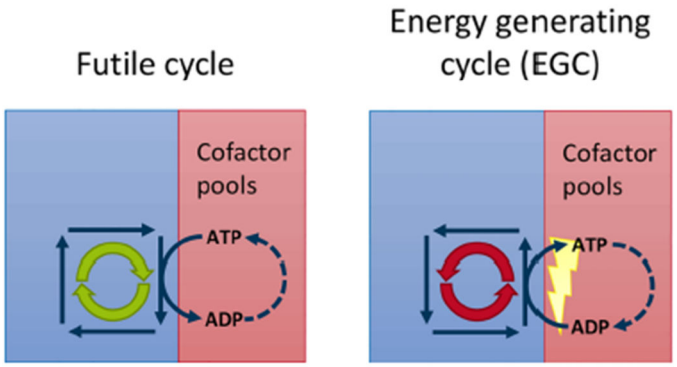

Fig 1. A futile cycle that consumes energy drawn from a cofactor pool (left) and an energy generating cycle (EGC) (right), which is thermodynamically impossible but occurs in some metabolic network models (figure extended from [12]). We can convert the type-II pathways to type-III pathways by closing the cycles in the cofactor pools (dashed arrows).

https://doi.org/10.1371/journal.pcbi.1005494.g001

A futile cycle running in reverse would charge energy metabolites such as ATP without an external source of energy (Fig 1). Accordingly, we classify type-II pathways [12] into two subgroups by taking the directionality of cofactor utilization into account: (a) futile cycles, which consume energy and are thus thermodynamically feasible, and (b) energy generating cycles (EGC), which charge energy metabolites without a source of energy.

While such EGCs are thermodynamically impossible, they can—and, as we show below, do—occur in constraint-based models. Futile cycles will rarely occur in FBA solutions, as they dissipate energy and hence divert metabolic investment away from biomass production. EGCs, in contrast, can have a substantial effect on the predictions of constraint-based analyses, as they generate energy out of nothing that then supports *in silico* growth. A simple example illustrating a (hypothetical) EGC is shown in Fig 2.

Eliminating EGCs is crucial for the correct modeling of energy metabolism, as has been recognized earlier (see, *e.g.*, [13–15]). While thermodynamically infeasible type-III pathways (internal cycles) can be easily removed through a simple post-processing step [5, 16], the same strategy cannot be used to suppress EGCs. In principle, EGCs could be excluded from the solution space by systematically assigning sufficiently detailed thermodynamic constraints. Thermodynamics-Based Metabolic Flux Analysis (TMFA) [17], for example, searches for a set of feasible metabolite concentrations such that all reactions proceed in the direction of negative free energy change ($\Delta G<0$) or, equivalently, a ratio of product to substrate concentrations below the reaction's equilibrium constant, K_{eq}. However, it can be shown mathematically that for any flux distribution without type-III pathways, there exists a distribution of metabolite concentrations such that the flux distribution is thermodynamically feasible, *i.e.*, all fluxes proceed in the direction of negative free energy change (see the theorem in [16]). This theoretical result respects the fact that metabolite concentrations must have a single value for all reactions they participate in; Supplementary S1 Text shows a small example network that illustrates the inability of ll-COBRA and TMFA to reliably exclude EGCs.

The simplest EGC could be established through an ATP energy dissipation reaction ($ATP + H_2O \rightarrow ADP + Pi + H^+$) that is allowed to proceed in the backwards direction. An energy-generating backward flux can be achieved as long as the concentration ratio $([ATP][H_2O]) / ([ADP][Pi][H^+])$ is smaller than the corresponding equilibrium constant $K_{eq} = 2\times10^{-5}M^{-1}$. If we treat the concentration of H_2O (55M) as constant, this cannot occur within the physiological concentration bounds assumed by Henry *et al.* [17], 10^{-5}M and 0.02M, showing that TMFA's metabolite concentration bounds avoid the utilization of at least some EGCs. Note, however, that reactions central to an EGC may have equilibrium constants compatible with

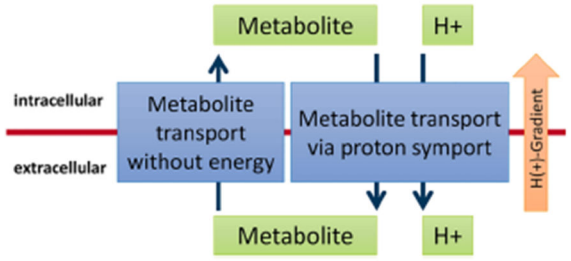

Fig 2. A simple (hypothetical) example of an energy generating cycle (EGC). A symporter that exports a metabolite and a proton acts together with a transporter that takes the same metabolite up without a proton. A combination of both reactions builds up a proton gradient that can then be utilized to generate energy (*e.g.*, via an ATP synthase).

https://doi.org/10.1371/journal.pcbi.1005494.g002

the concentration bounds, especially if the total free energy change is spread over several individual reactions. Moreover, the TMFA strategy relies on the availability of equilibrium constants for all reactions central to the EGC.

In contrast to TMFA, several alternative thermodynamically informed constraint-based methods only consider chemical potentials, which do not incorporate information on reaction specifics such as the equilibrium constant K_{eq} [7, 12, 18, 19]. For every flux distribution free of type-III pathways, it is possible to find a distribution of chemical potentials such that all fluxes proceed in the direction of chemical potential reduction [16]. This means that potentials capable of driving energy dissipation reactions towards the high-energy metabolite can always be found. Thus, constrained-based methods designed to ensure thermodynamic feasibility based on freely variable chemical potentials do not guarantee the elimination of EGCs.

The detection and removal of EGCs is currently not part of established metabolic network reconstruction pipelines [2]. In particular, automatic reconstructions algorithms [20, 21] currently do not test for EGCs. Sometimes, EGCs are identified in the manual reconstruction process, and parts of the cycles are constrained to zero flux as a makeshift correction [13]. Accordingly, as we demonstrate below, the problem of erroneous free energy generation occurs in a majority of automated and a subset of manual network reconstructions.

Results and discussions

Erroneous energy-producing cycles occur in many published reconstructions

EGCs can be identified through a variant of FBA [14]. To efficiently identify the existence of diverse EGCs, we first add a dissipation reaction to the metabolic network for each metabolite used to transmit cellular energy; *e.g.*, for ATP, the irreversible reaction ATP + H_2O → ADP + P + H^+ is added. These dissipation reactions close any existing energy-generating cycles, thereby converting them to type-III pathways. Fluxes through any of the dissipation reactions at steady state indicate the generation of energy through the metabolic network. Second, all uptake reactions are constrained to zero. The sum of the fluxes through the energy dissipation reactions is now maximized using FBA. For a model without EGCs, these reactions cannot carry any flux without the uptake of nutrients.

We used this approach to identify the presence of EGCs for 14 different energy metabolites (ATP, CTP, GTP, UTP, ITP, NADH, NADPH, Flavin adenine dinucleotide, Flavin mononucleotide, Ubiquinol-8, Ubiquinol-8, 2-Demethylmenaquinol 8, Acetyl-CoA, L-Glutamate) and for proton exchange between periplasm and cytosol (for simplicity counted as a 15th "energy metabolite" below); see Suppl. S1 Table for the corresponding dissipation reactions. We did not require the energy dissipation reactions to be charge-balanced; e.g., in the reaction NADH → NAD^+ + H^+, we omitted the molecule that acts as the acceptor of the two electrons. Adding the electron acceptor to the dissipation reaction would not dissipate the energy stored in NADH, as this energy could then potentially be re-used by internal cofactor regeneration reactions; in this case, the dissipation reaction could be active even in the absence of EGCs. FBA models do not keep track of metabolite charges, and thus the general problem posed by charge unbalanced reactions is not that they affect constraint-based simulations directly; instead, they are a sign of incorrect reaction stoichiometry, which is especially severe in the case of electron imbalances. It is important that models are mass and electron balanced [2] before conducting the EGC analysis. While EGCs induced by mass or electron unbalanced reactions may be detected by our method, they cannot be removed properly without fixing the reaction stoichiometries.

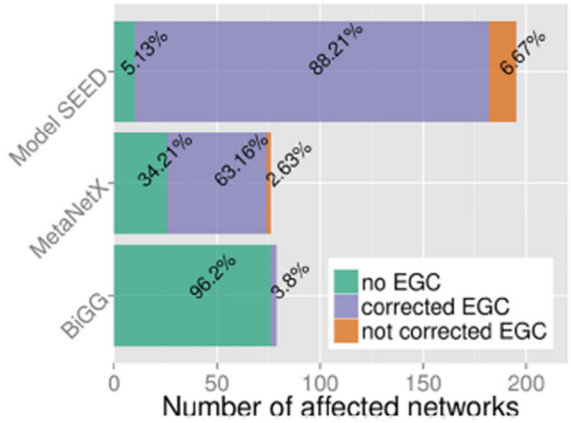

Fig 3. The majority of metabolic network reconstructions in two of the examined databases (ModelSEED and MetaNetX) contain erroneous internal EGCs that generate energy. In contrast, most models in BiGG do not contain EGCs. Total bar size reflects the number of models contained in each database. Green: models without EGCs; purple: models with EGCs that could be corrected through GlobalFit; orange: models with EGCs that cannot be corrected through reaction removals.

https://doi.org/10.1371/journal.pcbi.1005494.g003

We analyzed all models in three large databases of constraint-based metabolic networks: BiGG [22], ModelSeed [23], and MetanetX [24]. Overall, we found that over two thirds (68%) of tested models supported a non-zero flux through at least one of the 15 energy dissipation reactions, although this percentage differed drastically between databases (Fig 3).

The BiGG database contains high-quality manual [2] and, in the case of 54 *E. coli* strains, semi-automated [25] genome scale metabolic reconstructions. We found EGCs in only 3 out of the 79 BiGG models (3.8%; Suppl. S2 Table). The ModelSEED database is connected to a service for high-throughput reconstruction and analysis of metabolic networks. A special feature is the fully automated reconstruction of genome-scale networks from genome sequences. Due to the fully automated reconstruction, models created by this service should be considered as draft models, and manual steps for model improvement are recommended [23]. Consistent with this recommendation, we identified EGCs in 95% of ModelSEED models (185 out of 195; Suppl. S2 Table). Finally, MetaNetX is a Meta-Database for metabolic network models, gathering metabolic networks from different databases (including The ModelSEED and the BiGG database) and mapping them to one common namespace. This allows easy meta-analysis, manipulation, and comparison of those models [24]. Our FBA strategy found EGCs in 66% of MetaNetX models (50 out of 76; Suppl. S2 Table).

GlobalFit can eliminate >90% of EGCs by removing reactions

For each network with EGCs, we then used a slightly modified version of GlobalFit [26] to suggest a minimal number of reaction removals that eliminate all EGCs, allowing independent removals of forward and backward directions for reversible reactions. GlobalFit was originally designed to reconcile inconsistencies between FBA model predictions and measured growth/non-growth data, *e.g.*, from gene knockouts. GlobalFit uses a bi-level optimization method to identify the minimal set of network changes needed to correctly predict all experimentally observed growth and non-growth cases (or a subset thereof) simultaneously. We slightly altered the original algorithm, now simultaneously contrasting one growth case (the network with the biomass reaction as the objective function, ensuring that the suggested modifications

do not interfere with biomass production), and one non-growth case (the network with the sum of energy dissipation reactions as the objective function, ensuring that the modified network contains no EGCs; for details see Materials and methods). It can be argued that some types of reactions should be preferentially removed; e.g., reactions only weakly supported by genomic evidence may be removed first, and it may be more likely that one direction of a reaction labeled as reversible represents a network error than that an irreversible reaction is erroneous. While the modified GlobalFit algorithm allows such differential weighing of different reaction types, we considered all reaction removals as equally likely in the application detailed below. Moreover, reactions could be preferentially removed depending on the estimated equilibrium constant (or standard Gibb's free energy change ΔG_0).

For 94% of metabolic models with EGCs (223 out of 238), GlobalFit found a set of reaction removals that eliminated all EGCs while maintaining the ability to produce biomass. In many cases, GlobalFit suggested the removal of the ATP synthase reaction. While this will indeed remove most ATP-producing cycles, it will also abolish the model's natural ability to produce ATP through respiration. To avoid this undesired side effect, we performed a second search for reaction removals that eliminated all EGCs, this time forcing the algorithm to retain the ATP synthase reaction. This step could be adapted to the physiology of the studied organism by selecting a different reaction set to be retained. In each case, we could identify an alternative set of reaction removals; below, we only consider these alternative sets of suggested network changes. Note that GlobalFit does not actually remove the offending reactions, but constrains their fluxes to zero. This allows their reactivation in conditions where they are deemed thermodynamically feasible, although alternative measures must then be taken to avoid EGCs.

Most erroneous models can be corrected by making up to five originally reversible reactions irreversible (Fig 4). The removal of irreversible reactions was only rarely suggested by the algorithm (Fig 4), while the complete removal of reversible reactions was never observed. In the remaining unsolved models, EGCs could in principle be eliminated by adding reactions to the metabolic networks. The addition of reactions not directly connected to an EGC may be needed to restore biomass production in case no solution exists that preserves viability after EGC removal. While the modified version of GlobalFit is capable of suggesting such

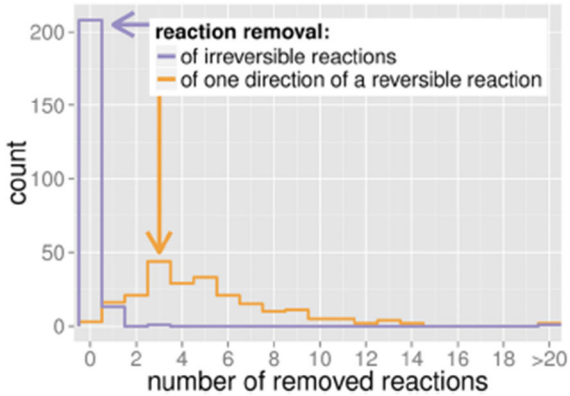

Fig 4. Most erroneous models can be corrected by making up to 5 originally reversible reactions irreversible. Purple: histogram of the number of irreversible reactions removed in each model to eliminate EGCs. Orange: histogram of the number of reversible reactions made irreversible to eliminate EGCs.

https://doi.org/10.1371/journal.pcbi.1005494.g004

additions, the application of this strategy would require manual revision, as it might incorrectly add new metabolic capabilities.

While bi-level mixed integer optimization algorithms such as the one used by GlobalFit typically require long computation times, GlobalFit resolved most solvable EGCs in under 10s, and all but one EGC within one minute (Supplementary S1 Fig). The only calculation that required over one minute was for the yeastnet 7.6 model [27], for which the CPLEX solver did not find an optimal solution within the set limit of 60 hours on 16 CPUs. The best set of changes found for this yeast model eliminated all EGCs by removing 76 reactions (or reaction directions). According to CPLEX, there is no alternative elimination of EGCs with fewer than 33 reactions; thus, this model contains at least 33 EGCs. As many EGCs include transport reactions across cellular membranes, the large number of EGCs found in this eukaryotic model (and the resulting increased computation time) may be caused by the existence of several intracellular compartments and the associated transport processes.

Freely available energy may boost biomass production. Accordingly, the elimination of EGCs through the reaction removals suggested by GlobalFit resulted in biomass reductions in 92% of cases (206 out of 223), typically by more than 25% (Fig 5). This indicates that the *in-silico* biomass yield may be unrealistically high in a majority of automatically generated models.

Examples for network corrections suggested by GlobalFit

One of the simplest EGCs we identified is displayed in Fig 6(A). This cycle is contained in only two metabolic models from The ModelSEED database, *Klebsiella pneumoniae* MGH 78578 (Seed272620.3) and *Flavobacterium johnsonia johnsoniae* UW101 (Seed376686.6). In this EGC, a malate symporter (rxn10153) transports malate together with two protons out of the cell. The exported malate molecule is then re-imported together with a sodium ion via the malate/Na+ symporter (rxn05207). The sodium is in turn exported by a Na+/Proton antiporter (rxn05209) in exchange for the import of only one of the protons of the first reaction. Thus, the second exported proton from the first reaction is free to drive an ATP-synthase reaction, generating ATP from ADP without access to an external energy source. To eliminate this EGC, the cost of either malate or sodium transport in terms of translocated protons must be corrected. This option was not given to GlobalFit, which instead suggests to remove the export direction of the malate symporter (rxn10153).

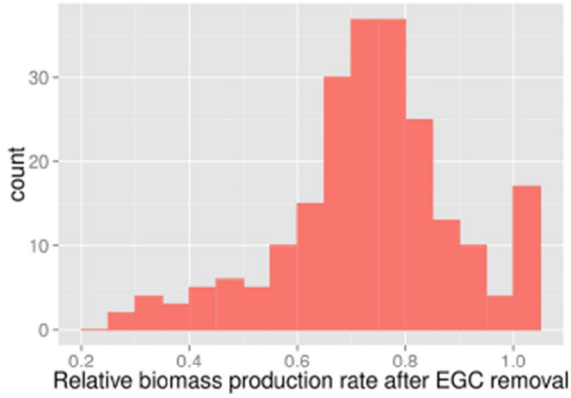

Fig 5. Removal of EGCs led to substantially reduced maximal biomass yield in most models. Histogram of the ratio between maximal biomass production rate before and after EGC removal.

https://doi.org/10.1371/journal.pcbi.1005494.g005

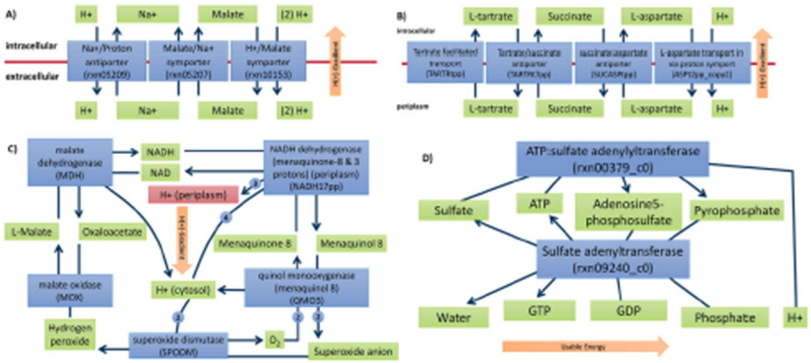

Fig 6. Examples of EGCs found in published genome-scale models. Green/red: metabolites; blue: reactions, linking substrates and products; orange: direction of the energy gradient utilized by the energy dissipation reaction. (A) The simplest identified cycle, which links a Na+/proton antiporter (exporting Na+ in exchange for a single proton) and a Malate/proton symporter (importing Malate together with two protons) via a Malate/Na+ symporter. (B) A cycle involving two antiporters and one symporter, driven by a transporter that translocates tartrate from the periplasm to the cytosol. (C) A NADH:menaquinone oxidoreductase, which translocates protons in the process of transferring electrons from NADH to Menaquinone 8, driven by a chain of four enzymes. (D) rxn00379 creates Adenosine 5'-phosphosulfate from ATP and sulfate. The equivalent sulfate adenyltransferase rxn09240 catalyzes the backward reaction, but charges a GTP in addition to the ATP.

https://doi.org/10.1371/journal.pcbi.1005494.g006

In the manually curated model iJO1366, six reactions (SPODM, SPODMpp, SUCASPtpp, SUCFUMtpp, SUCMALtpp, and SUCTARTtpp) were inactivated in the published model to avoid unrealistic energy generating loops by constraining their flux to zero [13]. We could identify two distinct EGCs (Fig 6B and 6C) by reactivating these reactions in the iJO1366 model and in the 54 other *E. coli* models derived from this reconstruction [25]. One of these EGCs is the rather simple cycle (Fig 6B) based on tartrate facilitated transport (TARTRtpp), found in 45 of the 55 *E. coli* strain reconstructions in Ref. [25]. This reaction spontaneously imports tartrate from the periplasm into the cell, while the tartrate/succinate antiporter (TARTRt7pp) exports tartrate, but simultaneously imports succinate. The cycle continues with the succinate/aspartate antiporter and then the aspartate/proton symporter, so that eventually a proton gradient between periplasm and cytosol is established. GlobalFit suggests to remove the utilized direction of the tartrate/succinate antiporter.

The other EGC found in the unconstrained *E. coli* models is a more complicated cycle (Fig 6C) that occurs in 46 of the 55 *E. coli* reconstructions [25], including the manually curated iJO1366 model [13]. A proton gradient across the periplasmic membrane is established by a NADH:menaquinone oxidoreductase (NADH17pp), which translocates protons in the process of transferring electrons from NADH to Menaquinone 8, driven by a chain of four enzymes, including superoxide dismutase (SPODM). In order to deactivate the cycle, GlobalFit removes the backward direction of the Malate oxidase (MOX) or the forward reaction of the Superoxide dismutase (SPODM). In this case, removal of the Malate oxidase would also be suggested by an analysis of standard free energy changes, at it is highly energetically unfavourable.

The EGC shown in Fig 6D was found in 99 out of 195 metabolic models from the ModelSEED database [20]. rxn00379 creates Adenosine 5'-phosphosulfate from ATP and sulfate. The sulfate adenyltransferase rxn09240 catalyses the backward reaction (and has the same

EC-number assigned), but charges not only an ATP, but additionally a GTP in the process. To eliminate this EGC, GlobalFit suggests removing either one of the participating reactions.

Conclusions

EGCs are a major issue in FBA modeling—they are able to produce energy out of thin air, thereby severely affecting the appropriate representation of energy metabolism and of biomass yield. EGCs not only affect the accurate representation of existing metabolic systems. They will be particularly problematic in evolutionary simulations that involve the incorporation of foreign metabolic reactions from other species [28–30]. Such mixing of reactions from disparate model reconstructions may easily introduce EGCs, and may thus lead to erroneous phenotype predictions. We have recently suggested a protocol for evolutionary simulations that avoids this problem [31]. Here, we present an improved computational method for the high-throughput identification of EGCs.

EGC identification is currently not a recognized step in model reconstruction, although some authors have eliminated EGCs from their manually curated models before publication. While constraint-based methods may avoid the utilization of EGCs based on thermodynamic considerations [17], such methods are computationally expensive and require careful analysis of the EGCs and the bounds on metabolite concentrations to guarantee the absence of EGCs from the resulting flux distributions. Instead, we propose to correct the metabolic model itself, and present a modified version of the previously published GlobalFit algorithm to eliminate EGCs through the removal of minimal reaction sets. The resulting model can then be used with the full suite of standard constraint- based methods.

We found EGCs in the majority of automatically generated models and in a small subset of manually curated networks. Many of the identified EGCs—in particular those that occurred most frequently—involved the erroneous maintenance of proton gradients across cellular membranes. The simplest EGCs would consist of two reversible reactions that catalyze the same biochemical conversion using different amounts of energy metabolites (Figs 2 and 6D). Such trivial EGCs are easily recognizable and are consequently rarely included in published metabolic networks; most EGCs in published models are more complex, and not easily identified by eye. We note that automatically reconstructed models often contain other types of errors as well [2]. For example, charge and mass imbalanced reactions appear to be common in automatic reconstructions and can lead to erroneous FBA predictions. Such reactions can potentially introduce EGCs, and we thus suggest to correct them as a preprocessing step.

The inclusion of reaction sets that are capable of forming an EGC into a metabolic network reconstruction is not necessarily erroneous. It is conceivable that one part of the cycle is thermodynamically feasible in one condition, whereas the other part is thermodynamically feasible in another condition, while both are not thermodynamically feasible simultaneously. Accordingly, modeling algorithms that respect thermodynamic constraints do not utilize potential EGCs [3, 6, 7, 17, 32–34]. FBA, however, does not consider thermodynamics; instead, optimization of its objective function (*e.g.*, biomass production rate) will usually lead to the exploitation of EGCs. One possible solution would be to constrain the fluxes through thermodynamically impossible sections of EGCs to zero in each simulated environment; this, however, would require a detailed understanding of environment-specific thermodynamics (or, alternatively, environment-specific gene regulation).

Our algorithms are suitable to guide a manual curation of draft networks, and should be included in the standard toolbox used for metabolic network reconstruction. GlobalFit can enumerate alternative solutions to eliminate EGCs, which can then be used as a basis for expert curation. In the context of automated network reconstruction pipelines such as ModelSEED

or kBase, our methods could be applied without human interaction, albeit at the risk of removing reactions that might be thermodynamically feasible in particular environments.

Materials and methods

Dataset and EGC detection

We started from 350 genome-scale metabolic networks (GSMs) that were downloaded from three databases: BiGG [22]–mostly manually created GSMs (accessed July 2015); ModelSeed [23]–GSMs created automatically from genome sequences (accessed July 2015); and MetaNetX [24]–a meta-database containing metabolic models from various sources (accessed January 2016). We removed networks that were unable to produce biomass in a maximally rich environment. We checked the correct direction of exchange reactions, and set the lower bound of the ATP maintenance reaction (ATPM) to zero, *i.e.*, we did not require a non-growth-related production of ATP.

To each GSM, we added 15 energy dissipation reactions (Supplementary S1 Table), where the namespace for metabolite names had to match the source of the network, *i.e.*, BiGG, ModelSEED, or MetaNetX. Because not every metabolic network covers the full range of metabolites used in the energy dissipation reactions (EDR), we checked the integration of the reactions in the network, defined as the fraction of the reaction's metabolites also present in the remainder of the model (*i.e.*, a reaction with an integration of 1 is completely integrated, whereas reactions with an integration < 1 cannot carry any flux). Because EGCs tend to run with maximal fluxes, all network reactions except the newly added ones (those in energy dissipation reactions) are restricted to fluxes in the range $[-1, 1]$ for reversible and $[0, 1]$ for irreversible reactions.

To establish the presence of EGCs for different energy metabolites, we maximized one energy dissipation reaction flux v_d at a time while prohibiting all influx into the model:

$$\max\left(v_d\right)$$

subject to:

$$\begin{aligned} &S\boldsymbol{v} = 0 \\ &\forall i \notin E: \quad v_i^{min} \leq v_i \leq v_i^{max} \\ &\forall i \in E: \quad v_i = 0 \end{aligned}$$

Here, S is the stoichiometric matrix, $\boldsymbol{v}$ the vector of fluxes, d the index of one of the energy dissipation reactions, $\boldsymbol{v}^{min}$ and $\boldsymbol{v}^{max}$ the vector of lower and upper reaction bounds, respectively, and E is the set of indices of all exchange reactions.

An optimal value v_d^* for this optimization with $v_d^* > 0$ indicates the presence of at least one cycle that is able to generate a specific type of energy metabolite (corresponding to the index d) in the network. Because $0 \leq |v_i| \leq 1$ for all reactions other than dissipation reactions, the value of v_d^* is a lower bound for the number of non-overlapping EGCs for the tested energy metabolite in the network.

The modified GlobalFit algorithm

Once a GSM was identified to contain at least one EGC, GlobalFit was used to eliminate all EGCs from the network. GlobalFit was developed to find globally minimal sets of model changes that simultaneously reconcile sets of experimental growth and non-growth observations with model predictions; a detailed description of the original GlobalFit algorithm can be found in [26]. We modified GlobalFit for the efficient removal of EGCs as outlined below.

In this modified version, the only allowed type of model change is the removal of unidirectional reactions, where reversible reactions are treated as two independent unidirectional reactions. We contrast a single growth with a single non-growth case. The non-growth case reflects the removal of all EGCs: with no nutrient uptake allowed, the maximal sum of fluxes through the energy dissipation reactions must be zero, $\max(\Sigma_d|v_d|) = 0$. To ensure that reaction removals do not abolish biomass production by eliminating EGCs, we set up a growth case with a minimal biomass production rate in a rich medium that allows uptake of all nutrients.

Formally, we solve the following bi-level optimization problem, which is a variation of the original GlobalFit problem [26] (Variable definitions are listed in Table 1):

$$\min_{\delta}\left(\sum_{y\in M}\left(\delta_y^{RF}+\delta_y^{RB}\right)\right) \tag{1}$$

subject to:

$$S_g \times \boldsymbol{v}^g = 0 \tag{2}$$

$$\forall_{y\in M}\, v_y^{min}\times(1-\delta_y^{RB}) \le v_y^g \le v_y^{max}\times(1-\delta_y^{RF}) \tag{3}$$

$$v_{Bio}^g \ge T_g \tag{4}$$

$$S_{ng} \times \boldsymbol{v}^{ng} = 0 \tag{5}$$

$$\forall_{y\in M}\, v_y^{min}\times(1-\delta_y^{RB}) \le v_y^{ng} \le v_y^{max}\times(1-\delta_y^{RF}) \tag{6}$$

$$\min_{v^{ng}}(c^t \times \boldsymbol{v}^{ng}) = 0 \tag{7}$$

δ_y^{RB} and δ_y^{RF} are binary variables. Setting one of these variables to 1 will constrain the corresponding flux of the growth—Eq (3)—and non-growth case—Eq (6)–to zero. The total number of reaction removals is minimized, where the removal of forward and backward reaction is treated separately in Eq (1)–*i.e.*, δ_y^{RB} and δ_y^{RF} are independent. Both the growth and the non-growth case must be in steady state, Eqs (2) and (5). The biomass production of the growth case has to be greater than a predefined threshold T_g, Eq (4). All entries in c^t are 0, except for the positions of the energy dissipation reactions, which are 1. The maximal summed flux

Table 1. Definitions of the variables used in the system of equations that describes the modified GlobalFit algorithm.

M	The set of reactions included in the original (input) network reconstruction
S	Stoichiometric matrix of the original (input) network reconstruction
$\boldsymbol{v}$	Flux vector
g	Growth case
ng	Non-growth case
v_y^{min}	Lower bound of reaction y
v_y^{max}	Upper bound of reaction y
v_{Bio}^g	Biomass reaction of the growth case
T_g	Growth threshold of the growth case
c^t	Vector containing ones and zeros. All entries are zero, except for the positions of the energy dissipation reactions

https://doi.org/10.1371/journal.pcbi.1005494.t001

through all energy dissipation reactions must be zero, Eq (7). We convert this bi-level optimization problem into a single level optimization problem as described in [26].

All calculations were run in GNU R with the SyBiL library [35] and a modified GlobalFit library [26] under linux. We used IBM ILOG CPLEX as the solver for the mixed integer linear optimizations. Each calculation was run on 8 CPU cores and 50GB main memory.

Supporting information

S1 Fig. Computation time. Distribution of computation (wall-clock) times for the application of GlobalFit to the metabolic models containing EGCs. While almost all computations finished in under a minute on a PC (8 CPUs, 50Gb RAM), the search for model corrections requires considerably more time for the yeast 7 metabolic network (data points off scale; the "simple" calculations were stopped after 61.27 hours and the "synthase" run needed 25.03 minutes). The red line ("simple") is for runs allowing all reaction removals; the blue line ("synthase") is for runs not allowing removal of the ATP synthase reaction.
(TIFF)

S1 Table. Energy dissipation reactions. Energy dissipation reactions (EDRs) for each of the 15 different types of energy metabolites in the cell.
(XLSX)

S2 Table. EGC occurrences in models. For each model, this table shows whether biomass production was possible at all (hasGrowth), whether energy generating cycles are present (hasEGCs), the identified types of EGCs (e.g., generates.ATP), and the reactions (or reaction directions) removed by GlobalFit for the "simple" run (all removals allowed) and with removal of the ATP synthase forbidden ("synthase").
(XLSX)

S1 Text. Toy model. A small example network that illustrates the inability of ll-COBRA and TMFA to reliably exclude EGCs.
(PDF)

Acknowledgments

We are grateful for computational support through the Zentrum für Informations- und Medientechnologie (ZIM) at Heinrich Heine University Düsseldorf.

Author Contributions

Conceptualization: CJF DH BS BP MJL.

Formal analysis: CJF DH BS.

Funding acquisition: MJL.

Investigation: CJF DH BS.

Methodology: CJF DH BS.

Software: CJF DH.

Supervision: BP MJL.

Validation: CJF DH.

Visualization: CJF DH.

Writing – original draft: CJF DH MJL.

Writing – review & editing: CJF DH BS BP MJL.

References

1. O'Brien EJ, Monk JM, Palsson BO. Using Genome-scale Models to Predict Biological Capabilities. Cell. 2015; 161(5):971–87. https://doi.org/10.1016/j.cell.2015.05.019 PMID: 26000478
2. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010; 5(1):93–121. https://doi.org/10.1038/nprot.2009.203 PMID: 20057383
3. Price ND, Thiele I, Palsson BO. Candidate states of Helicobacter pylori's genome-scale metabolic network upon application of "loop law" thermodynamic constraints. Biophys J. 2006; 90(11):3919–28. https://doi.org/10.1529/biophysj.105.072645 PMID: 16533855
4. Schilling CH, Schuster S, Palsson BO, Heinrich R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnol Prog. 1999; 15(3):296–303. https://doi.org/10.1021/bp990048k PMID: 10356246
5. Muller AC, Bockmayr A. Fast thermodynamically constrained flux variability analysis. Bioinformatics. 2013; 29(7):903–9. https://doi.org/10.1093/bioinformatics/btt059 PMID: 23390138
6. Schellenberger J, Lewis NE, Palsson BO. Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophys J. 2011; 100(3):544–53. https://doi.org/10.1016/j.bpj.2010.12.3707 PMID: 21281568
7. Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. Biophys J. 2002; 83(1):79–86. https://doi.org/10.1016/S0006-3495(02)75150-3 PMID: 12080101
8. Wiback SJ, Palsson BO. Extreme pathway analysis of human red blood cell metabolism. Biophys J. 2002; 83(2):808–18. https://doi.org/10.1016/S0006-3495(02)75210-7 PMID: 12124266
9. Sridharan GV, Ullah E, Hassoun S, Lee K. Discovery of substrate cycles in large scale metabolic networks using hierarchical modularity. BMC Syst Biol. 2015; 9:5. https://doi.org/10.1186/s12918-015-0146-2 PMID: 25884368
10. Reidy SP, Weber JM. Accelerated substrate cycling: a new energy-wasting role for leptin in vivo. Am J Physiol Endocrinol Metab. 2002; 282(2):E312–7. https://doi.org/10.1152/ajpendo.00037.2001 PMID: 11788362
11. Russell JB. The energy spilling reactions of bacteria and other organisms. J Mol Microbiol Biotechnol. 2007; 13(1–3):1–11. https://doi.org/10.1159/000103591 PMID: 17693707
12. Price ND, Famili I, Beard DA, Palsson BO. Extreme pathways and Kirchhoff's second law. Biophys J. 2002; 83(5):2879–82. https://doi.org/10.1016/S0006-3495(02)75297-1 PMID: 12425318
13. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Mol Syst Biol. 2011; 7:535. https://doi.org/10.1038/msb.2011.65 PMID: 21988831
14. Quek LE, Dietmair S, Hanscho M, Martinez VS, Borth N, Nielsen LK. Reducing Recon 2 for steady-state flux analysis of HEK cell culture. J Biotechnol. 2014; 184:172–8. https://doi.org/10.1016/j.jbiotec.2014.05.021 PMID: 24907410
15. Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics. 2016; 12:109. https://doi.org/10.1007/s11306-016-1051-4 PMID: 27358602
16. Desouki AA, Jarre F, Gelius-Dietrich G, Lercher MJ. CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. Bioinformatics. 2015; 31(13):2159–65. https://doi.org/10.1093/bioinformatics/btv096 PMID: 25701569
17. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. Biophysical Journal. 2007; 92(5):1792–805. https://doi.org/10.1529/biophysj.106.093138 PMID: 17172310
18. Beard DA, Babson E, Curtis E, Qian H. Thermodynamic constraints for biochemical networks. J Theor Biol. 2004; 228(3):327–33. https://doi.org/10.1016/j.jtbi.2004.01.008 PMID: 15135031
19. Nigam R, Liang S. Second Law of Thermodynamics Applied to Metabolic Networks. 2003.
20. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014; 42(Database issue):D206–14. https://doi.org/10.1093/nar/gkt1226 PMID: 24293654
21. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods Mol Biol. 2013; 985:17–45. https://doi.org/10.1007/978-1-62703-299-5_2 PMID: 23417797

22. Schellenberger J, Park JO, Conrad TM, Palsson BO. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics. 2010; 11:213. https://doi.org/10.1186/1471-2105-11-213 PMID: 20426874

23. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 2010; 28(9):977–82. https://doi.org/10.1038/nbt.1672 PMID: 20802497

24. Ganter M, Bernard T, Moretti S, Stelling J, Pagni M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics. 2013; 29(6):815–6. https://doi.org/10.1093/bioinformatics/btt036 PMID: 23357920

25. Monk JM, Charusanti P, Aziz RK, Lerman JA, Premyodhin N, Orth JD, et al. Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. Proc Natl Acad Sci U S A. 2013; 110(50):20338–43. https://doi.org/10.1073/pnas.1307797110 PMID: 24277855

26. Hartleb D, Jarre F, Lercher MJ. Improved Metabolic Models for E. coli and Mycoplasma genitalium from GlobalFit, an Algorithm That Simultaneously Matches Growth and Non-Growth Data Sets. PLoS Comput Biol. 2016; 12(8):e1005036. https://doi.org/10.1371/journal.pcbi.1005036 PMID: 27482704

27. Aung HW, Henry SA, Walker LP. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. Ind Biotechnol (New Rochelle N Y). 2013; 9(4):215–28.

28. Matias Rodrigues JF, Wagner A. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol. 2009; 5(12):e1000613. https://doi.org/10.1371/journal.pcbi.1000613 PMID: 20019795

29. Barve A, Wagner A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature. 2013; 500(7461):203-+. https://doi.org/10.1038/nature12301 PMID: 23851393

30. Hosseini SR, Martin OC, Wagner A. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc Biol Sci. 2016; 283(1839).

31. Szappanos B, Fritzemeier J, Csorgo B, Lazar V, Lu XW, Fekete G, et al. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun. 2016; 7.

32. Kummel A, Panke S, Heinemann M. Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics. 2006; 7:512. https://doi.org/10.1186/1471-2105-7-512 PMID: 17123434

33. De Martino D, Capuani F, Mori M, De Martino A, Marinari E. Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. Metabolites. 2013; 3(4):946–66. https://doi.org/10.3390/metabo3040946 PMID: 24958259

34. Hoppe A, Hoffmann S, Holzhutter HG. Including metabolite concentrations into flux balance analysis: Thermodynamic realizability as a constraint on flux distributions in metabolic networks. Bmc Systems Biology. 2007; 1.

35. Gelius-Dietrich G, Desouki AA, Fritzemeier CJ, Lercher MJ. Sybil—efficient constraint-based modelling in R. BMC Syst Biol. 2013; 7:125. https://doi.org/10.1186/1752-0509-7-125 PMID: 24224957

Acknowledgement

Firstly, I would like to express my sincere gratitude to my advisor Prof. Martin Lercher for giving me the opportunity to obtain my Ph.D. degree in his working group. Besides that, I would also like to thank him for his continuous support, fruitful discussions, great ideas, and at least 1000 cups of coffee.

Also, I would like to thank Prof. Oliver Ebenhöh for being my second referee.

Furthermore, I thank Prof. Florian Jarre for explaining the exciting world of mathematical optimization.

I would also thank all members of the bioinformatics working group for a great time, particularly Dr. Jonathan Fritzemeier for fruitful collaborations.

I would like to thank Prof. Shin-Han Shiu and the whole Shiu lab for hosting my research stay at the MSU in East Lansing.

I would also thank Al Bay for being a great host during my stay in the United States.

I am grateful for funding through the German Research Foundation and the International Graduate School iGRAD-Plant.

Last but not least, I want to thank Ina, my friends and family.