Environmental Adaptation of Bacteria: Insights from Genome-Scale Metabolic Networks

Inaugural-Dissertation

zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Claus Jonathan Fritzemeier
aus Mönchengladbach

Düsseldorf, 29. Mai 2017

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

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

Referent: Prof. Dr. Martin J. Lercher
Korreferent: Prof. Dr. William F. Martin

Tag der mündlichen Prüfung: 12. September 2017

Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig 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, 29. Mai 2017

Claus Jonathan Fritzemeier

Contents

1 Abbreviations

ADP	adenosine diphosphate
ATP	adenosine triphosphate
COBRA	constraint-based reconstruction and analysis
CRAN	the Comprehensive R Archive Network
EFM	elementary flux mode
EGC	energy generating cycle
FBA	flux balance analysis
GAM	growth-associated maintenance
GEM	genome-scale model
GENRE	genome-scale network reconstruction
GPR	gene to protein (enzyme) to reaction interaction
GSM	genome-scale (metabolic) model
HGT	horizontal gene transfer
MCMC	Markov chain Monte Carlo
MMB	minimal metabolic behaviour
MTF	minimum total flux
NGAM	non-growth-associated maintenance
SBML	systems biology markup language

2 Preface

According to the "Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf vom 06.12.2013" § 6 (4) this document is written as a cumulative thesis. Four manuscripts are presented along with an accompanying text introducing the reader to the broader topic, explaining the manuscripts' findings in connection with current literature and their relation to each other. Additionally, for each manuscript the author's contributions are listed and an outlook discussing future research opportunities is given.

Manuscript 1 presents a novel software-package for constraint-based modelling written in GNU R. This new software is significantly faster than its competitors and integrates seamlessly in the environment of GNU R. This work was published as:
Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J., & Lercher, M. J. (2013). sybil - Efficient constraint-based modelling in R. BMC systems biology, 7(1), 125.

Having this novel tool at hand it was possible to perform larger *in silico* evolutionary experiments and investigate the evolution of complex innovations in microbial metabolism. A hypothesis about the influence of changing environments was proposed in *Manuscript 2* and was published as:
Szappanos, B., Fritzemeier, J., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R. A., Lercher, M. J., Pál, C., & Papp, B. (2016). Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nature Communications, 7(11607).

During the PhD project, a severe problem in published metabolic network reconstructions was discovered: some models are able to produce energy without any input of nutrients. This problem was already handled in *Manuscript 2*, but a systematic investigation and a general solution to the problem was not yet developed. *Manuscript 3* presents a novel method to detect and remove these energy-generating cycles:

Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B., & Lercher, M. J. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Computational Biology, 13(4), e1005494.

Concurrently with the work on *Manuscript 3*, a follow up on *Manuscript 2* was prepared that compares the adaptive evolution of multiple unicellular species by creating a pan-genome-scale metabolic network from 71 individual organism-specific genome-scale metabolic models. *Manuscript 4* is submitted with the following bibliography:
Fritzemeier, C. J., Lieder, F., Szappanos, B., Jarre, F., Papp, B., Pal, C., & Lercher, M. J. (2017). Differences in the adaptability of generalist and specialist bacteria: the influence of metabolic network size and structure.

3 Summary

Evolutionary biology is frequently challenged to explain how complex adaptations, *e.g.*, the mammalian eye, can arise from the purely stochastic process of evolution. These complex adaptations, which require alterations in multiple genes or even sets of completely new genes, could only arise slowly if evolution was not guided by adaptations. Cellular metabolism is without doubt a complex trait. Even the smallest unicellular organisms are able to synthesize all necessary cell components from simple nutrients.

Manuscript 1 introduces the novel software package *sybil* for constraint-based modelling with genome-scale metabolic models. Unlike most alternative software packages, *sybil* is very fast, flexible, and completely free to use. Additional packages can easily extend *sybil* and thereby add new algorithms to solve constraint-based problems.

Manuscript 2 proposes the hypothesis of stepwise metabolic niche expansion: adaptations to changing nutritional environments accelerate the evolution of complex metabolic pathways by utilizing exaptations. Unlike previous work (Barve & Wagner, 2013), the new hypothesis can explain complex adaptations without neutral mutations. In a flux balance framework, a metabolic model of *E. coli* was allowed to adapt to new environments by acquiring minimal reaction sets from a universe of reactions, a process simulating lateral gene transfer (LGT). The hypothesis is based on the result that some of these beneficial reaction sets were found to be subsets of other beneficial reaction sets that are necessary for adaptation to other environments and thus can serve as exaptations. A phylogenetic ancestor reconstruction analysis confirmed that the genes of beneficial reaction sets that serve as exaptations are frequently acquired earlier than genes depending on the exaptation. Finally, an evolutionary laboratory experiment with *E. coli* brought another piece of evidence for this hypothesis.

Manuscript 3 deals with the problem of erroneous energy-generating cycles in metabolic network reconstructions. Metabolic networks can con-

tain reaction cycles able to produce an infinite amount of energy without any nutrition uptake. Although such cycles are clearly thermodynamically infeasible, they occur in over 85% of published metabolic networks that were not extensively manually curated, such as models included in the Model SEED (Henry et al., 2010) or MetaNetX (Ganter et al., 2013) databases. *Manuscript 3* is the first work that names the problem, and presents a method to systematically identify and remove erroneous energy generating cycles from metabolic networks. The identification can efficiently be done with a modified flux balance analysis, but removal of energy generating cycles can easily disrupt the cell's energy metabolism and thereby the biomass production. Thus a modified version of the GlobalFit (Hartleb et al., 2016) algorithm was used, which calculates minimal network changes that remove the erroneous energy-generating cycles while simultaneously preserving the biomass production.

The work presented in *Manuscript 4* uses pan-genome-scale metabolic modelling to investigate the adaptability of 71 unicellular organisms to new nutrient sources. The analysis revealed a strong correlation between genome size and the number of reactions necessary for these adaptations. While the organism with the most metabolic genes, *Shigella flexneri*, is able to adapt to new environments with on average three additional reactions, the organism analysed with the smallest, reduced genome, the endosymbiont *Buchnera aphidicola*, needs at least 27 additional reactions. These results confirm the findings of an abstract toolbox model (Maslov et al., 2009); the metabolic capabilities of an organism scale approximately quadratically with the number of metabolic genes. As proposed earlier (Barve & Wagner, 2013), adaptations to one environment often go along with "inadvertent" adaptations to other non-selected environments. By quantifying this trait, it was found that organisms with large metabolic networks profit more from adaptations than those with small networks, although the latter acquire more reactions. A reason for this surprising finding might be fewer branching points in the metabolic networks of specialists. All results are consistent with the dichotomy of generalists and specialists based on the number of initially growth promoting en-

vironments; thus our results might explain the hurdle for specialists to move out of their current ecological niche.

References

Barve, A. & Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature, 500(7461), 203–6.

Ganter, M., Bernard, T., Moretti, S., Stelling, J., & Pagni, M. (2013). MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics (Oxford, England), 29(6), 815–6.

Hartleb, D., Jarre, F., & Lercher, M. J. (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(8), e1005036.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology, 28(9), 977–82.

Maslov, S., Krishna, S., Pang, T. Y., & Sneppen, K. (2009). Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proceedings of the National Academy of Sciences of the United States of America, 106(24), 9743–8.

4 Zusammenfassung

Die Evolutionsforschung wird oft dadurch gefordert, dass sie erklären muss wie komplexe evolutionäre Anpassungen, wie z.B. das Auge eines Säugetiers, allein durch die zufälligen Prozesse der Evolution entstehen können. Diese komplexen Anpassungen benötigen oft Veränderungen in mehreren Genen oder komplett neue Gene und könnten nur langsam entstehen, wenn die Evolution nicht durch Anpassungen geleitet würde. Der Metabolismus einer Zelle ist ohne Zweifel eine komplexe Anpassung. Schon kleinste Einzeller sind in der Lage alle nötigen Zellbestandteile aus einfachsten Nährstoffen selbst zu synthetisieren.

Manuskript 1 stellt das neues Software Paket *sybil* für beschränkungsbasierte Modellierung mit metabolischen Modellen in Genomgröße vor. Im Gegensatz zu den meisten alternativen Software Paketen ist *sybil* sehr schnell, flexibel und komplett kostenlos nutzbar. *Sybil* kann einfach mit zusätzlichen Paketen erweitert werden und damit neue Algorithmen für beschränkungsbasierte Probleme integrieren.

Manuskript 2 stellt die neue Hypothese der schrittweisen Expansion über metabolische Nischen auf: Anpassungen an wechselnde Nährstoffumgebungen beschleunigen die Evolution von komplexen metabolischen Reaktionspfaden durch Exaptationen. Im Gegensatz zu einer vorherigen Arbeit (Barve & Wagner, 2013), kann die neue Hypothese die komplexen Anpassungen ohne neutrale Mutationen erklären. Mit Hilfe der Flussbilanzanalyse wurde die Anpassung eines metabolischen Modells von *E. coli* durch lateralen Gentransfer (LGT) an neue Nährstoffumgebungen simuliert. Dabei wurde dem Netzwerk je Anpassung nur die minimal nötige Anzahl an Reaktionen aus einem Reaktionsuniversum hinzugefügt. Die Hypothese basiert darauf, dass Mengen nützlicher Reaktionen Teilmengen von Reaktionsmengen sind, die Anpassungen für eine andere Umgebung sind und damit als Exaptationen dienen können. Mit einer Analyse des phylogenetischen Stammbaums konnte bestätigt werden, dass Gene der Reaktionmengen, die als Exaptation dienen, häufig vorher ins

Genom aufgenommen werden als Gene, die auf die Exaptation aufbauen. Schließlich brachte ein evolutionäres Laborexperiment mit *E. coli* einen weiteren Beweis für die Hypothese.

In *Manuskript 3* geht es um das Problem von irrtümlich Energie produzierenden Zyklen in metabolischen Netzwerk Rekonstruktionen. Metabolische Netzwerke können Reaktionszyklen enthalten, die Energie in unendlicher Menge produzieren können ohne dabei Nährstoffe aufzunehmen. Obwohl solche Zyklen offensichtlich thermodynamische Gesetze missachten, sind sie in über 85% der publizierten metabolischen Netzwerke zu finden, die nicht von Hand kuriert wurden. Solche Modelle sind in der Model SEED (Henry et al., 2010) oder MetaNetX (Ganter et al., 2013) Datenbank zu finden. *Manuskript 3* ist die erste Veröffentlichung, die das Problem benennt und eine systematische Methode zur Identifikation und Beseitigung von falschen Energie produzierenden Zyklen in metabolischen Netzwerken präsentiert. Die Identifikation kann sehr effizient mit einer Flussbilanzanalyse gemacht werden, aber die Beseitigung der Energie produzierenden Zyklen kann sehr leicht den Energiestoffwechsel der Zelle zerstören und damit auch die Biomasse Produktion. Deshalb wurde eine modifizierte Version des GlobalFit (Hartleb et al., 2016) Algorithmus benutzt, der minimale Änderungen am Netzwerk berechnet um die irrtümlich Energie produzierenden Zyklen zu entfernen aber gleichzeitig die Biomasse Produktion erhält.

Die Arbeit in *Manuskript 4* nutzt ein metabolisches Modell in Pangenomgröße und untersucht die Anpassungsfähigkeit von 71 einzelligen Organismen an neue Nährstoffquellen. Dabei wurde eine starke Korrelation zwischen Genomgröße und der Anzahl für die Anpassung nötigen Reaktionen festgestellt. Während der Organismus mit den meisten metabolischen Genen, *Shigella flexneri*, sich an neue Umgebungen mit durchschnittlich drei zusätzlichen Reaktionen anpassen kann, braucht der von den untersuchten Organismen mit dem kleinsten Genom, der Endosymbiont *Buchnera aphidicola*, mindestens 27 zusätzliche Reaktionen. Diese Ergebnisse bestätigen die Forschungsergebnisse des abstrakten Werkzeugkisten

Modells (*toolbox model*) (Maslov et al., 2009); die metabolischen Fähigkeiten eines Organismus skalieren ungefähr quadratisch mit der Anzahl seiner metabolischen Gene. Wie zuvor berichtet (Barve & Wagner, 2013), bringen Anpassungen an eine Umgebung oft „unbeabsichtigte“ Anpassungen an andere nicht selektierte Umgebungen mit sich. Die Quantifikation dieser Eigenschaft zeigte, dass Organismen mit großen metabolischen Netzwerken mehr von Anpassungen profitieren als Organismen mit kleinen Netzwerken, obwohl letztere mehr Reaktionen für Anpassungen benötigen. Ein Grund für dieses überraschende Ergebnis könnte die geringere Anzahl an Verzweigungen in metabolischen Netzwerken von Spezialisten sein. Alle Ergebnisse sind konsistent mit der Aufteilung in Generalisten und Spezialisten auf Basis der Anzahl von initial Wachstum begünstigenden Umgebungen. Deshalb könnten die Ergebnisse die Hürde erklären, die Spezialisten überwinden müssen um ihre aktuelle ökologische Nische zu verlassen.

Literatur

Barve, A. & Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature, 500(7461), 203–6.

Ganter, M., Bernard, T., Moretti, S., Stelling, J. & Pagni, M. (2013). MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics (Oxford, England), 29(6), 815–6.

Hartleb, D., Jarre, F. & Lercher, M. J. (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(8), e1005036.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. & Stevens, R. L. (2010). High-throughput generation, optimization and

analysis of genome-scale metabolic models. Nature biotechnology, 28(9), 977–82.

Maslov, S., Krishna, S., Pang, T. Y. & Sneppen, K. (2009). Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proceedings of the National Academy of Sciences of the United States of America, 106(24), 9743–8.

5 Introduction

Nothing in Biology Makes Sense Except in the Light of Evolution.

Theodosius Dobzhansky, 1973

All life on earth evolved through adaptive evolution. Before Darwin (1859) wrote down his evolutionary theory, Paley (1802) brought up the watchmaker analogy: he argued that if you would find a mechanic watch on a meadow, the most likely explanation was, that there had to be a watchmaker who built that said watch. Naturally, the watchmaker is only a analogy for a god and the watch is meant to be a complex system like a living being. This concept is called "intelligent design" and is taught in more and more science classes around the world as replacement for evolutionary studies (Berkman et al., 2008; Watts et al., 2016). But, if not by a god, how was it even possible that complex traits have evolved without a designer, if genetic modifications, *i.e.*, mutation or recombination, are purely stochastic processes? Richard Dawkins shows in his book "The Blind Watchmaker" (1986) how something complex can be arise from random variation and selection. He utilized the infinite monkey theorem: a monkey is sitting in front of a typewriter and presses the keys randomly. After an infinite amount of time, we can be sure that the monkey has typed a certain sentence, *e.g.*, Dawkins uses the sentence "Methink it is like a weasel." from William Shakespeare's "Hamlet". However, this argument is criticized because of the immense time requirement. The process takes so long, because the monkey always starts to write the sentence from the beginning. But if we fix the letters that the monkey already typed correctly and let him only retype the wrong ones, the time to finish the sentence is strongly diminished (Dawkins, 1986). So to get something meaningful from a random input, an evaluation of fitness, *i.e.*, correctness of the letters, and a rule for selection, *i.e.*, "keep the correct letters", is needed. Additionally, the modularity of a trait, *i.e.*, the positions of the letters are independent, is important for the evolution of complex adaptations (Wagner & Altenberg, 1996).

The research field of metabolic network evolution deals with the adaptive evolution of complex systems, because metabolism is a complex trait. The metabolic network represents the entire metabolism of unicellular lifeforms and is sufficient for synthesizing basic building blocks of the cell used for reproduction. Microorganisms are central to the biotechnology industry, because of their metabolic capabilities. This is not a new innovation: fermentation of sugar has been used since thousands of years to produce alcoholic beverages. Other examples are the use of yeast in bread and the use of lactic acid bacteria to produce yoghurt. Recent examples are the production of ethanol as fuel and the decomposition of various waste materials. Much effort is put into optimizing microbial metabolism for industrial purposes (Adrio & Demain, 2014). To do so efficiently, the metabolism of the microbes has to be fully understood. These complex systems were shaped over millions of years of evolutionary fine-tuning. This long time of adaptation made them efficient in performing their tasks compared to purely chemical processes (Li et al., 2014). But the metabolism is not designed, and a coherent concept describing metabolic systems is missing. But in the light of evolution, we can try to understand how organisms have evolved and why they are as they are. This allows us to propose hypotheses about how those networks were shaped, and a growing understanding regarding the interplay of their functions may emerge.

5.1 Metabolic Modelling

Conducting experiments in the laboratory is undoubtedly an expensive and time consuming task. Thus, over the last two decades much effort was put into the development of computational methods that simulate wet lab experiments. In combination with the rising performance of computers and supercomputers (Waldrop, 2016), the complexity of simulated biological processes could be increased. Even though these *in silico* methods cannot and are not intended to replace experiments completely, they complement and reduce the amount of wet lab experiments necessary. Additionally,

simulations allow to conduct evolutionary experiments that are impossible to be done with living organisms on a reasonable time scale.

Techniques for metabolic modelling can be classified into two types: mechanistic models (also called kinetic models) and constraint-based models.

Kinetic models are created by combining biochemical reactions whose reaction kinetics are sufficiently understood to describe them quantitatively. All reactions in the model have to have known enzyme kinetics, reaction stoichiometry, and mass and electron balances. The authors of a kinetic model must carefully check every detail of a reaction before putting the reaction into the network. The resulting networks only have a small number of reactions and often only represent single pathways or parts of the cell's metabolism, because the necessary kinetic parameters are difficult to measure. Additionally, simulations with genome-scale kinetic models are computational challenging (Wiechert & Noack, 2011), because they involve solving many differential equations.

The other type of metabolic modelling is constraint-based modelling. In this approach, the space of all possible reaction rates for each reaction is constrained by the stoichiometry, predefined minimal and maximal reaction rates, and the steady-state condition that the concentrations of metabolic intermediates are constant. Then, one solution is selected from this solution space by maximizing a given objective function (Orth et al., 2010). This approach is based on linear programming and thus it is very fast to compute, *i.e.*, it requires only split seconds for a genome-scale model (Gelius-Dietrich et al., 2013).

In silico evolutionary experiments demand a high number of simulations, ideally performed on genome-scale metabolic networks. In contrast to kinetic modelling, constraint-based modelling can cope with both of those demands in acceptable computation times. Thus, this thesis focuses on constraint-based modelling with genome-scale metabolic network reconstructions.

5.1.1 Metabolic Networks

A metabolic network is the reconstruction of the metabolism of a specific organism (or sets of organisms). By applying modelling techniques, it can be used to simulate metabolic functions of the organism. The often added attribute "genome-scale" indicates that the metabolic network includes all enzymes found in an organism's genome.

The term metabolic network is generously used and is often taken as synonym for other more exact technical terms. Three of the most important and frequently used terms are: GENRE, GEM, and GSM. A genome-scale network reconstruction (GENRE) is the collection of different kinds of information linked to an organism and a specific genome (Price et al., 2004; Feist et al., 2009; Thiele & Palsson, 2010); this will be explained in detail in Section 5.1.2. The genome-scale model (GEM) is a mathematical model derived from the GENRE by applying constraint-based modelling methods, *e.g.*, flux balance analysis (FBA), which is described in Section 5.1.3. In connection with the model name the abbreviation of "genome-scale (metabolic) model" (GSM) was used earlier to stress the genome-scale size of a model, *e.g.*, in the name "*i*JR904 GSM/GPR" (Reed et al., 2003); in this example GPR indicates that the model also contains "gene to protein (enzyme) to reaction interactions" (see Section 5.1.2). Currently GSM is mostly used as stand-alone abbreviation (Ganter et al., 2013), because most of the newly published models are of genome-scale size anyway.

The first genome-scale metabolic reconstruction was published for *Haemophilus influenzae* (Schilling & Palsson, 2000), which also was the first free living organism for which the genome was completely sequenced (Fleischmann et al., 1995). From there on, many scientific groups published whole-genome metabolic network reconstructions for various organisms. Many publications in this field come from the group of Bernhard Ø. Palsson. His group also published the first model (*i*JE660) for *Escherichia coli* K-12 strain MG1655 (Edwards & Palsson, 2000) and also the three most popular metabolic models of *E. coli*: *i*JR904 (Reed et al., 2003),

*i*AF1230 (Feist et al., 2007), and *i*JO1366 (Orth et al., 2011). Each of these models is basically a refinement of the preceding model; in each step more metabolic genes were added and thus the models can predict *in vitro* growth increasingly better.

Most metabolic models are not named after the represented organism, but follow a special naming convention suggested by Reed et al. (2003). The first letter of the name is always an *i* as abbreviation for "*in silico*". This is followed by the initials of the first author and then by the number of genes contained in the reconstruction. If a model is derived from another model, it should be named by the original model, followed by an additional letter (*e.g.*, *i*AF1230b). This convention allows to easily distinguish between different models for the same species.

5.1.2 Formal Definition of a Metabolic Network

A metabolic network is the a set of multiple reactions, where reactions are connected with each other on basis of shared metabolites. Products of one reaction are the educts of another reaction. By the concatenation of multiple reactions, reaction pathways are formed. A metabolite can also be used by more than two reactions, which causes a branching of a reaction pathway.

The connections between reactions are described in the stoichiometric matrix S. If the metabolic network consists of n metabolites and m reactions, this matrix S is in $\mathbb{R}^{n \times m}$. The matrix' entries $S_{i,j}$ are defined to be the stoichiometric coefficient of the i-th metabolite in the j-th reaction. For a metabolite not participating in a reaction, the entry $S_{i,j}$ equals zero. A negative entry $S_{i,j}$ indicates that the i-th metabolite is consumed in the j-th reaction. Analogously, positive values indicate the production of a metabolite. This matrix can be read either row-wise, indicating which reaction uses the i-th metabolite, or column-wise, indicating which metabolites participate in the j-th reaction. While in nature the same metabolite can occur in different cellular compartments, in a metabolic network identical chemical compounds that occur in distinct

cell compartments are treated as different metabolites. The stoichiometric matrix S is the core information of a metabolic model and this information is sufficient for basic analyses, *e.g.*, the connectivities of reactions and metabolites.

The direction of biochemical reactions is always reversible. However, because of cellular metabolite concentrations and the enzyme kinetics, a predominant direction can often be determined, making the reaction effectively irreversible in practice. To include the directionality, upper and lower bounds are defined for each reaction. We will refer to these as two vectors $\vec{l}$ and $\vec{u}$, for lower and upper bounds, respectively. The j-th reaction flux v_j, *i.e.*, reaction direction and rate, is therewith restricted by $l_j \leq v_j \leq u_j$. A metabolic network with reaction directions is sufficient to perform an FBA and even more advanced analyses, *e.g.*, calculating minimal metabolic behaviours (MMBs) (Larhlimi & Bockmayr, 2005) or elementary flux modes (EFMs) (Schuster et al., 1996; Zanghellini et al., 2013).

The reactions in metabolic networks can be categorized into three types: internal reactions, exchange reactions, and biomass reactions. The internal reactions are mass and charge balanced and interconvert metabolites. Those reactions model enzymatically catalysed reactions, transport processes across membranes, or spontaneous reactions of metabolites. Exchange reactions can eliminate a metabolite from the network or – depending on its direction – can create the metabolite from the void. A reversible exchange reaction that exchanges metabolite M_i is defined as:

$$\mathrm{M}_i \rightleftharpoons \varnothing$$

This is necessary as input (source) and output (sink) for the model and can be interpreted as the metabolites diffusing to and from the cell in the growth medium. Hence, environmental conditions can be changed by setting the lower and upper bounds for exchange reactions appropriately. Exchange reactions are of course not mass or charge balanced. Finally, biomass reactions are a combination of the first two types and are ba-

sically a set of reactions that are coupled together, describing the total metabolic demand caused by cellular growth. This reaction set contains an energy-dissipation reaction and reactions acting as sinks for biomass precursors. The energy-dissipation reaction in the biomass reaction wastes energy by typically splitting adenosine triphosphate (ATP) into adenosine diphosphate (ADP) (Feist et al., 2007); this growth-associated maintenance (GAM) reaction models the amount of energy necessary for growth. Common biomass precursors are amino acids for protein biosynthesis, nucleotides for DNA replication and transcription, and lipids to assemble membranes. The exact composition of the biomass and therefore the biomass precursors is organism dependent and can vary, *e.g.*, by additional cell wall components.

Gene knockouts can be used to optimize microbial strains for higher product yield in the biotechnology industry (Burgard et al., 2003). In order to simulate such gene knockouts information about the connection between genes and reactions is required (*i.e.*, which gene encodes which protein and which proteins are needed for a certain enzyme). This so called "gene to protein to reaction" interaction (GPR) is defined in terms of boolean expressions. Identifiers of the genes represent the genomic presence or absence of a gene. These identifiers are connected by `AND` or `OR` operators for protein complexes or isoenzymes, respectively. The evaluation of this boolean expression for the j-th reaction with the presence and absence of the genes indicates whether the necessary enzyme for this reaction can be synthesized and thus if the reaction can be active. Otherwise the flux through reaction v_j is constrained to zero.

All the above information is typically accompanied by sum formulae of the metabolites or cross references of metabolites, reactions, and genes. This is not a necessity for constraint-based modelling, but it simplifies the assignment of data from other source to integrate the model into a larger scope. The collection of the described data, *i.e.*, the stoichiometric matrix, upper and lower bounds, and reaction and metabolite details, in genome-scale is called a GENRE. The use of a GENRE with constraint-

based modelling is described in in the next section and the creation of a GENRE is described in Section 5.1.4.

5.1.3 Constraint-Based Modelling

The field of constraint-based modelling comprises a set of over 100 methods that have evolved over time and can even be classified in a phylogeny (Lewis et al., 2012). This set of methods was given the name constraint-based reconstruction and analysis methods (COBRA methods). Central for these methods is the stoichiometric matrix. This can represent just a single reaction pathway, the full metabolism of a single cell, or the joint metabolism across multiple cells. Most studies utilizing COBRA methods simulate the metabolic phenotypes of single cellular organisms; however, applications to multicellular organisms have also been explored (Martins Conde et al., 2016). FBA is the most prominent constraint-based modelling technique.

Papoutsakis (1984) built the foundation for FBA. He developed a stoichiometric equation for butyric acid fermentation and could successfully verify the calculated data with experimental data from the literature. Later Watson (1984) was the first to use such a stoichiometric equation in connection with linear programming. The program was developed for educational purpose and was written for an Apple II computer with only 48 kB (Watson, 1984). The objective function in this first approach was to minimize free-energy dissipation. Fell and Small (1986) investigated the synthesis of triglyceride from glucose in rat adipose tissue with flux balance analysis. This was the first study applying flux balance analysis, though the explicit term was not defined yet.

Flux balance analysis (FBA) is the simplest of the commonly used constraint-based modelling methods, and a multitude of more advanced methods follow the same principles as FBA. Due to the importance of FBA for the manuscripts in this thesis a short introduction to FBA is given here, although FBA is described in textbooks and in numerous

publications (*e.g.*, Edwards et al. (2002), Price et al. (2004), Feist et al. (2007), Orth et al. (2010)).

The aim of FBA is to find a flux vector $\vec{v}$, such that each single reaction flux v_i satisfies the constraints of the metabolic network: Given a metabolic network as defined in Section 5.1.2 with it's stoichiometric matrix S, we can formulate the change of metabolite concentrations $\vec{c}$ over time depending on the flux vector $\vec{v}$.

$$S\vec{v} = \frac{d\vec{c}}{dt}$$

The time scales for enzymatic reactions and diffusion are magnitudes faster than cell growth, regulation, and process dynamics. Thus, we can simplify the calculation by assuming steady-state conditions, *i.e.*, no (internal) metabolite concentration changes. Practically speaking, all metabolites are instantly consumed as they are produced, and there is neither accumulation nor depletion of metabolites. Additionally, every single atom type entering the model through exchange reactions has to leave the network at the same rate. By combining the steady-state condition and the formula above, we get:

$$S\vec{v} = \frac{d\vec{c}}{dt} = 0 \iff S\vec{v} = 0$$

This equation describes the solution space of $\vec{v}$. The solution space contains all possible flux vectors that follow steady-state condition and the stoichiometry of the metabolic network. Genome-scale metabolic networks typically consist of more reactions (m) than metabolites (n), so $m > n$. When solving $S\vec{v} = 0$, we have more unknown variables ($\vec{v} = [v_1, \ldots, v_m]$) than there are equations in this system of equations. In this under-determined system of equations, $\vec{v}$ is not exactly determined, rather multiple solutions fulfil the criteria. Although we cannot calculate a single solution for $\vec{v}$, we can try to narrow down the solution space by adding the constraints on reaction directionality, *i.e.*, upper ($v_i \leq u_i$) and lower flux bounds ($v_i \geq l_i$). But even with these constraints, we still

typically have a multidimensional solution space that can be described as a steady-state flux cone (David et al., 2011; Larhlimi et al., 2012), which contains, although within certain bounds, still infinitely many solutions. We can now define an objective function $Z = \vec{z}^{\top}\vec{v}$, with $\vec{z} \in \mathbb{R}^n$, to find in this cone the solution with maximal Z. Typically, $\vec{z}$ is a vector of zeros having a one at position j, meaning the objective is to optimize the flux through the j-th reaction. The optimality criterion chosen for microbes is usually to maximize the model's biomass reaction flux, but other objectives are possible, *e.g.*, fixing the biomass reaction flux to a certain rate and minimizing glucose uptake. However, after optimization the solution for $\vec{v}$ might still not be unique. There might exist infinite equivalent solutions of $\vec{v}$ with the same value for Z. Thus, if a unique solution is needed, basic FBA is not sufficient. Alternatives like minimum total flux (MTF) (Holzhütter, 2004) or geometric FBA (Yuan et al., 2016) can extend FBA to calculate unique solutions for $\vec{v}$.

The first approaches of using FBA consisted of writing self-made scripts producing input files for the linear optimization software. At this time, there were no complete software solutions that support the handling of metabolic networks and the easy usage of constraint-based methods. This changed with the publication of the COBRA Toolbox by Becker et al. (2007). This Matlab toolbox supports easy input of metabolic networks from SBML-Files and enables the user to conduct an FBA with just a few lines of code. A later major update of this toolbox (Schellenberger, Que, et al., 2011) added more functionality, namely additional methods related to FBA like network gap filling, ^{13}C analysis, omics-guided analysis and visualization (Schellenberger, Que, et al., 2011).

Although various alternative software solutions for constraint-based modelling exist, most of them are pretty slow and thus not suitable for large-scale analyses. The computation time for analyses with single models is still acceptable, but for evolutionary experiments the complexity rises vastly. Additionally, the COBRA Toolbox is based on the commercial software Matlab, making it expensive to use.

The publication of *Manuscript 1* brought a new alternative to the existing set of available software packages in the field of constraint-based modelling. The new package *sybil* for the statistical programming language and software environment GNU R (R Development Core Team, 2014) is open source, just as R itself, and thus completely free to use. Additionally, *Manuscript 1* reported that *sybil* out-competes most alternative software packages in terms of computation speed. *Sybil* provides high-level functions that make it easy to perform simple analyses even for inexperienced users. But it is possible for developers to easily implement new constraint-based algorithms that seamlessly integrate, because *sybil* has an object-oriented architecture. These new algorithms can then be used in the top level functions without knowledge of the underlying implementation. While implementing extensions, the developer can still profit from *sybil*'s unified solver interface and the new algorithm is compatible with all solvers that are supported by *sybil*. Another advantage of the object-oriented architecture is, that experienced users can adapt their program to their needs and gain even higher speed-ups. The related package *sybilSBML* can easily read metabolic networks published as SBML files and generates from this instances of a class designed to store metabolic networks. The information stored in this class is accessible in a simple manner for other packages, *e.g.*, the package *RSeed* described in *Manuscript 1*.

5.1.4 The Reconstruction Process of Genome-Scale Metabolic Network Models

Most of the exact details of the reconstruction methodology were kept undisclosed by the groups creating the first GENREs. Although there was the necessity for a general standard of how to reconstruct GENREs, it took several years before the articles Feist et al. (2009) and, in even more detail, Thiele and Palsson (2010) set the standard for the reconstruction of metabolic networks. All earlier approaches of network reconstructions

were done without this protocol, but follow the same or similar steps (Suthers et al., 2009; AbuOun et al., 2009).

The nearly one hundred steps of the reconstruction protocol (Thiele & Palsson, 2010) can be summarized like this: first the genome of the organism of interest has to be annotated. Thereby are genes with metabolic functions identified and associations with similar genes in other organisms are established. Afterwards, for each metabolic gene the connection to the right enzyme and the right reaction has to be made. This initial set of reactions is considered a draft reconstruction. In the first place, these steps can be run automatically. Now an excessive task of manual curation starts: the reconstruction refinement phase. Each reaction has to be reviewed to ensure correct metabolite formulae and reaction stoichiometry. Importantly, the reaction direction has to be determined. Additional reactions are defined that have no evidence in the genome and are not catalysed by an enzyme, *e.g.*, transport processes via diffusion, exchange reactions, and spontaneous reactions. Finally, the biomass reaction, non-growth-associated maintenance (NGAM), *i.e.*, energy consumed by the organism for living, and growth media composition are defined. Now an evaluation phase follows that tests whether predictions of the model correlate with the known physiology of the organism.

One suggested step is called gap filling. The draft reconstruction process does not necessarily create fully connected reaction pathways. This can be due to failed annotations or missing information. An earlier published algorithm is able to detect and close these gaps by adding a minimal number of reactions from a given database to the reconstruction (Satish Kumar et al., 2007). Of course this can falsely add new metabolic functions to the network by closing gaps that actually exist in the organism's metabolic network. The algorithm GlobalFit solves this problem by contrasting growth and non-growth data sets and finding minimal changes in the network to match those (Hartleb et al., 2016). The steps of the reconstruction refinement phase are repeated until the predictions cannot be further improved. In the end the finished model has to be

saved in a file format that supports easy exchange between programs. The systems biology markup language (SBML) (Hucka et al., 2003) is established as a standard file format for constraint-based modelling.

The process of creating a GENRE for a new organism involves a lot of manual work and often lasts over a year (Thiele & Palsson, 2010). This changed with the presentation of the first high-throughput pipeline for the generation of genome-scale metabolic models (Henry et al., 2010; Devoid et al., 2013; Overbeek et al., 2014). The Model SEED pipeline basically automates the steps from the established protocol (Thiele & Palsson, 2010). This pipeline runs purely as a web service with a close connection to the RAST service (Overbeek et al., 2014). First, the RAST service is used to annotate the genome, *i.e.*, to assign function to the genes of the genome on the basis of sequence homology. With the annotated genome as input, the Model SEED service reconstructs the metabolic network. Just recently a software was published that offers a similar pipeline, but is executable on a local machine (Cuevas et al., 2016).

As the extensive manual work in the reconstruction process has to be automated, the pipelines rely on a reaction database that was curated beforehand (Henry et al., 2010; King et al., 2016). The resulting reconstructions have to be considered draft reconstructions and still require human revision (Henry et al., 2010). An intermediate solution has been chosen by Monk et al. (2013). They chose to combine automated steps with manual curation and created 55 strain-specific *E. coli* GENREs. These semi-automated reconstructions are considered to have the same quality as purely manually reconstructed metabolic networks.

Assigning thermodynamically feasible reaction directions is a crucial step in the reconstruction process, because standard FBA cannot determine those during calculation. Wrongly assigned reaction directions are a likely reason for energy generating cycles (EGCs), which create an unlimited energy supply without any nutrient uptake. These EGCs are clearly thermodynamically impossible, but were occasionally discovered in published GENREs by various authors, *e.g.*, Orth et al. (2011), Arnold

et al. (2015). Although enhanced varieties of FBA include thermodynamic constraints (Henry et al., 2007; Schellenberger, Lewis, & Palsson, 2011), these methods are still unable to exclude EGCs from the solution space. Although cycles capable of generating unlimited energy can severely compromise the prediction power of a GSM, previous literature neither provided a consistent name nor a systematic method to find and remove those cycles. *Manuscript 3* deals with this problem. Additionally, it reports a systematic investigation of three big databases with metabolic reconstructions. The analysis reveals that EGCs frequently occur in GENREs, but they can easily be corrected in most cases. Especially GENREs created from the automated reconstruction pipeline Model SEED showed a high amount of EGCs. By removing one of the reactions involved, EGCs can easily be inactivated. However, if this is not done with great care, it can easily disrupt the energy metabolism; thus, the GlobalFit algorithm was used to perform network changes. This was done by slightly adapting the algorithm and then contrasting one case enforcing biomass production with a second case forbidding artificial energy production, *i.e.*, the presence of EGCs. Thus, growth of the corrected model is assured while simultaneously removing the EGC.

5.2 Evolution of Complex Innovations

Biological systems are very complex; no genome-scale system is currently fully understood. However, science can explain the function, composition, or purpose of many complexes, sequences, or interactions among them. The circadian clock, for example, is a well studied system where mathematical models precisely predict the oscillating behaviour and the minimal number of enzymes needed (Scheper et al., 1999). However, most of these systems would be constructed differently and often much simpler if an engineer had designed them. Artificial proteins can be built with the same function as the wild type protein, but with a reduced number of distinct amino acid types (Kamtekar et al., 1993; Davidson et al., 1995; Walter et al., 2005). Why has nature evolved into using more different kinds of

amino acids than necessary? And why don't organisms use even more amino acids? The answer to the second question was given by Saint-Leger et al. (2016): With a rising number of used amino acid types, the tRNAs have to differ sufficiently from each other to avoid false charging of the tRNAs. For the twenty proteinogenic amino acids, the tRNAs seem to differ enough (Saint-Leger et al., 2016). Metabolic networks were also not built by a designer, but have evolved over millions of years. We also have to ask why metabolic networks are shaped like they are today and how it was possible from purely stochastic processes to evolve complex metabolic innovations.

Adaptation is explained by Freeman and Herron (2001) as follows: "A trait, or integrated suite of traits that increases fitness of its possessor is called an adaptation and is said to be adaptive". In connection with metabolic networks in microbes, an adaptation can be the ability to utilize a novel carbon source to synthesize the essential cell compounds.

Today, several competing hypothesis exist about the driving force of metabolic network evolution (Schmidt et al., 2003; Caetano-Anollés et al., 2009). Although evolutionary theories can never be fully proven, simulations and data analyses give evidence about their certainty. Horizontal gene transfer (HGT) has not only an important role in the metabolic adaptability of prokaryotes (Pál et al., 2005), but also has a major influence on the structure of bacterial genomes (Koonin & Wolf, 2008). Thus, the majority of studies in the field of metabolic network evolution consider whole gene gains and losses (and respective reaction gains and losses) instead of point mutations.

The neutral theory of molecular evolution proposes that the majority of substitutions arise as neutral mutations, *i.e.*, the mutation does not provide a fitness benefit for the organism and is not driven by selection (Kimura, 1983). Evidence for neutral evolution can be observed in nature. For example, Schultes and Bartel (2000) found a specific RNA sequence that is able to form two ribozymes with completely distinct function. Additionally, they found minor variants of this sequence that can assemble

into only one ribozyme, and these variants can be interconverted by purely neutral mutations as regards their common function. The neutral theory has been controversial for a long time, but it was finally not able to explain key aspects of protein evolution (Kreitman, 1996). Neutral mutations do occur in nature but were found not to be the driving force in evolution (Kreitman, 1996). Wagner (2008) suggests a reconciliation theory with neutral mutations as origins of adaptive innovations. That is, neutral mutations prepare the basis for an adaptation and only have a beneficial effect in combination with another consecutive mutation. This can be equally applied to whole genes: a novel gene without beneficial effect can have a beneficial effect in conjunction with a later gene acquisition.

5.2.1 Exaptation promotes Complex Innovations

An adaptation to one condition can also serve as an inadvertent preparation for a later adaptation to another condition. This process is called exaptation. The textbook example for exaptations is the evolution of feathers for flying. At first, feathers only had the purpose of thermal insulation. Later, the feathers were exapted for flying. The term preadaptation is often used synonymously to exaptation, but preadaptation can suggest an intention in evolution and this should be avoided (Gould & Vrba, 1982). Exaptations can frequently be found as prerequisites for complex innovations (Bock, 1959; Hayden et al., 2011). The question is now if those exaptations typically arose by purely adaptive steps during evolution like the feathers, or if they evolved as non-adaptive traits.

Exaptation can also be found in metabolic networks: As adaptations to a new nutrient, *e.g.*, a new carbon source, new reactions have to evolve or have to be acquired from other organisms. These new reaction pathways might now be utilized for the adaptation to a second new nutrient. Thus, the adaptation to the first nutrient provides an exaptation for the second nutrient.

Barve and Wagner (2013) argue that exaptations can arise non-adaptively as a side product of adaptive evolution. The experimental basis

for this theory is given by laboratory evolution with RNA enzymes that were performed by the same group (Hayden et al., 2011). Additionally many computational studies support that theory (Barve & Wagner, 2013; Hosseini et al., 2015; Hosseini & Wagner, 2016; Hosseini et al., 2016). Barve and Wagner (2013) explain the importance of non-adaptive mutations as exaptation for evolutionary innovations. For this, the authors use a Markov chain Monte Carlo (MCMC) algorithm to generate random samples of metabolic networks. The evolving networks are demanded to remain viable on a certain carbon source. Starting from an *E. coli* network, in each step of the MCMC simulation, one reaction is added from a database and another reaction is removed from the network. Thus the generated random networks remain constant in size, but sample the space of available reactions from the database. Importantly, the random walk may add reactions (or, in a stepwise fashion whole pathways) to the metabolic network that are disconnected from the rest of the network and thus do not have a beneficial effect. After the random walk, the resulting random network is tested for growth on other carbon sources. The number of new usable carbon sources is interpreted as a measurement of exaptation. Their main finding is that although the networks are adapted to a certain carbon source, they are typically also viable on multiple other carbon sources, emphasizing the importance of exaptation. Other work from mostly the same authors (Hosseini et al., 2015; Hosseini & Wagner, 2016; Hosseini et al., 2016) follow the same strategy and come to similar conclusions. It is important to note, however, that the simulation procedure is biologically unrealistic: bacteria tend to loose genes that don't contribute to fitness (Mira et al., 2001) and generally do not contain large numbers of randomly sampled surplus genes. Another strong argument against non-adaptive origins of exaptations is that this process is expected to be too slow: the neutral mutations are unlikely to be fixed in populations. Furthermore, no direct empirical support exists for this theory in bacteria.

Manuscript 2 provides evidence that supports the hypothesis that the adaptation to changing environments accelerates the development

of metabolic innovations in bacteria. The proposed hypothesis demands every gene acquisition to be adaptive, *i.e.*, to have a direct fitness benefit; neutral gene acquisitions are not allowed. Although neutral gene acquisitions frequently happen, they are unlikely to be fixed in the population. Evidence for this hypothesis is provided by three complementary analyses. First, a computational study with a genome-scale metabolic network shows the importance of exaptation and changing environments for complex metabolic innovations in *E. coli*. The analysis revealed that purely adaptive mutations serve as exaptations for a following adaptation. Second, a comparative genomic analysis of 943 bacterial genomes was performed that supports the hypothesis and the findings from the first analysis. To investigate the history of gene gains and co-gains, the ancestral states of orthologous enyzmes were reconstructed along the phylogenetic tree. Patterns of enzyme gains that correspond to the hypothesis were significantly over-represented in this tree. Third, a wet laboratory experiment shows how an *E. coli* mutator strain develops an exaptation for growing on ethylene glycol (EG) while adapting to propylene glycol (PG). The wild type *E. coli* can grow on neither of them (EG-, PG-). With the high mutation rate of the mutator strain, adaptation to propylene glycol (PG+) is possible, but still growth on ethylene glycol remains impossible (EG-). Subsequently, the PG+ cells can adapt to ethylene glycol (EG+). Further investigation discovered the necessity of two over expressed genes for growth on ethylene glycol (EG+), while only one of those genes is necessary for growth on propylene glycol (PG+). Thus, adaptation to PG provides an exaptation for growth on EG, providing direct evidence for the stepwise metabolic niche expansion hypothesis.

5.2.2 Ability for Metabolic Adaptation is Genome-Size Dependent

Prokaryotes can obtain genes via HGT and expand their metabolic network in order to utilize additional nutrients. The key analysis in *Manuscript 2* dealt with the adaptability of *E. coli* to new environments. To

simulate this, a set of minimal nutritional environments was created. For each environment that did not support growth of *E. coli*, was the number of additional reactions calculated that are necessary to make the model viable in this environment. Astonishingly, in the vast majority (74%) of environments, only one to three additional reactions were necessary for adaptation. Although *E. coli* is known to be a generalist, *i.e.*, able to grow in many different environments, the mechanisms behind this remained unclear. Is this ease of adaptation specific to *E. coli*, or is it a general feature of bacterial metabolism? *Manuscript 4* explores this central question of metabolic adaptability across different species and examines which factors influence adaptability.

Only very limited literature exists on this topic. One theoretical approach addressing this issue is the toolbox model from Maslov et al. (2009), which can be described as follows. Every metabolic gene or enzyme of an organism is a tool in a toolbox. Multiple tools can be combined into a metabolic pathway. If a new metabolic gene evolves in the organism or is acquired via HGT, the toolbox is enlarged. The crucial point is that a large toolbox compared to a small one gains more functionality by adding one tool, because the novel tool can be combined with more already existing tools. Maslov et al. (2009) use simulations of abstract networks to explain with this model the quadratic scaling between transcription factors and gene count (Van Nimwegen, 2003).

Wolf and Koonin (2013) propose a biphasic model of genome evolution based on genome sequence analysis. Their phylogenetic analyses revealed that short phases of rising genome complexity alternate with long phases of genome reduction. Hence, specialists arise from genome reductions. Nevertheless, the reasons for this biphasic evolution and hurdles detaining specialists to evolve into generalists remain unclear.

Manuscript 4 shows how the metabolic adaptabilities of 71 unicellular organisms differ significantly depending on their metabolic gene content. Organisms with a high number of metabolic genes tend to adapt more easily to new nutrients than small organisms with reduced genomes.

These findings, which were obtained in a flux balance analysis framework, correspond to the earlier proposed abstract toolbox model (Maslov et al., 2009). Additionally, is in this paper shown, that organisms can be partitioned into groups of generalists and specialists based on the number of viable environments. Metabolic models categorized as generalists tend to develop multiple collateral additional metabolic capabilities by adapting to one new environment, while the specialist's adaptations can mainly serve only the specific purpose favoured by natural selection. The metabolic network structure reflects this partitioning as well. Specialists tend to have more linear metabolic pathways, *i.e.*, they have less branching points than generalists.

5.3 References

AbuOun, M., Suthers, P. F., Jones, G. I., Carter, B. R., Saunders, M. P., Maranas, C. D., Woodward, M. J., & Anjum, M. F. (2009). Genome Scale Reconstruction of a Salmonella Metabolic Model: COMPARISON OF SIMILARITY AND DIFFERENCES WITH A COMMENSAL Escherichia coli STRAIN. Journal of Biological Chemistry, 284(43), 29480–29488.

Adrio, J. L. & Demain, A. L. (2014). Microbial enzymes: tools for biotechnological processes. Biomolecules, 4(1), 117–139.

Arnold, A., Sajitz-Hermstein, M., & Nikoloski, Z. (2015). Effects of varying nitrogen sources on amino acid synthesis costs in Arabidopsis thaliana under different light and carbon-source conditions. PLoS ONE, 10(2), e0116536.

Barve, A. & Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature, 500(7461), 203–6.

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., & Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols, 2(3), 727–738.

Berkman, M. B., Pacheco, J. S., & Plutzer, E. (2008). Evolution and creationism in America's classrooms: A national portrait. PLoS Biology, 6(5), 0920–0924.

Bock, W. (1959). Preadaptation and multiple evolutionary pathways. Evolution, 13(2), 194–211.

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. Biotechnology and Bioengineering, 84(6), 647–657.

Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., & Mittenthal, J. E. (2009). The origin and evolution of modern metabolism. The International Journal of Biochemistry & Cell Biology, 41(2), 285–297.

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

Darwin, C. R. (1859). On the Origin of Species by means of Natural Selection; or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.

David, L., Marashi, S.-A., Larhlimi, A., Mieth, B., & Bockmayr, A. (2011). FFCA: a feasibility-based method for flux coupling analysis of metabolic networks. BMC bioinformatics, 12(1), 236.

Davidson, A. R., Lumb, K. J., & Sauer, R. T. (1995). Cooperatively folded proteins in random sequence libraries. Nature Structural Biology, 2(10), 856–864.

Dawkins, R. (1986). The Blind Watchmaker. New York: W. W. Norton & Company, Inc.

Devoid, S., Overbeek, R., DeJongh, M., Vonstein, V., Best, A. A., & Henry, C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. Methods in molecular biology (Clifton, N.J.) 985, 17–45.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. The American Biology Teacher, 35(3), 125–129.

Edwards, J. S. & Palsson, B. Ø. (2000). Robustness analysis of the Escherichia coli metabolic network. Biotechnology Progress, 16(6), 927–939.

Edwards, J. S., Covert, M., & Palsson, B. Ø. (2002). Metabolic modelling of microbes: The flux-balance approach. Environmental Microbiology, 4(3), 133–140.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., & Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Molecular systems biology, 3, 121.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. Nature reviews. Microbiology, 7(2), 129–143.

Fell, D. A. & Small, J. R. (1986). Fat synthesis in adipose tissue - An examination of stoichiometric constraints. Biochem. J, 238, 781–786.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., … Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly of Haemophilus-Influenzae Rd. Science, 269(5223), 496–512.

Freeman, S. & Herron, J. C. (2001). Evolutionary Analysis.

Ganter, M., Bernard, T., Moretti, S., Stelling, J., & Pagni, M. (2013). MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. Bioinformatics (Oxford, England), 29(6), 815–6.

Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J., & Lercher, M. J. (2013). sybil - Efficient constraint-based modelling in R. BMC systems biology, 7(1), 125.

Gould, S. & Vrba, E. S. (1982). Exaptation — A Missing Term in the Science of Form. Paleobiology, 8(1), 4–15.

Hartleb, D., Jarre, F., & Lercher, M. J. (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(8), e1005036.

Hayden, E. J., Ferrada, E., & Wagner, A. (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. Nature, 474(7349), 92–95.

Henry, C. S., Broadbelt, L. J., & Hatzimanikatis, V. (2007). Thermodynamics based metabolic flux analysis. Biophysical journal, 92(5), 1792–1805.

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology, 28(9), 977–82.

Holzhütter, H. G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. European Journal of Biochemistry, 271(14), 2905–2922.

Hosseini, S.-R., Barve, A., & Wagner, A. (2015). Exhaustive Analysis of a Genotype Space Comprising 1015 Central Carbon Metabolisms Reveals an Organization Conducive to Metabolic Innovation. PLoS Computational Biology, 11(8), e1004329.

Hosseini, S.-R., Martin, O., & Wagner, A. (2016). Phenotypic innovation through recombination in genome-scale metabolic networks. Proceedings of Royal Society B, 283(1839).

Hosseini, S.-R. & Wagner, A. (2016). The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. BMC Systems Biology, 10(1), 97.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., … Wang, J. (2003). The systems biology markup language (SBML): a medium

for representation and exchange of biochemical network models. Bioinformatics, 19(4), 524–531.

Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M., & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. Science (New York, N.Y.) 262(5140), 1680–5.

Kimura, M. (1983). The neutral theory of molecular evolution (M. Kimura, Ed.). Cambridge University Press.

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., & Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research, 44(D1), D515–D522.

Koonin, E. V. & Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic acids research, 36(21), 6688–719.

Kreitman, M. (1996). The neutral theory is dead. Long live the neutral theory. Bioessays, 18(8), 678–683.

Larhlimi, A. & Bockmayr, A. (2005). Minimal metabolic behaviors and the reversible metabolic space. In Preprint 299, dfg research center matheon.

Larhlimi, A., David, L., Selbig, J., & Bockmayr, A. (2012). F2C2: a fast tool for the computation of flux coupling in genome-scale metabolic networks. BMC bioinformatics, 13(1), 57.

Lewis, N. E., Nagarajan, H., & Palsson, B. O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nature reviews. Microbiology, 10(4), 291–305.

Li, C. W., Ciston, J., & Kanan, M. W. (2014). Electroreduction of carbon monoxide to liquid fuel on oxide-derived nanocrystalline copper. Nature, 508(7497), 504–7.

Martins Conde, P. d. R., Sauter, T., & Pfau, T. (2016). Constraint Based Modeling Going Multicellular. Frontiers in molecular biosciences, 3(2012), 3.

Maslov, S., Krishna, S., Pang, T. Y., & Sneppen, K. (2009). Toolbox model of evolution of prokaryotic metabolic networks and their

regulation. Proceedings of the National Academy of Sciences of the United States of America, 106(24), 9743–8.

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. Trends in Genetics, 17(10), 589–596.

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., Feist, A. M., & Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. Proceedings of the National Academy of Sciences of the United States of America, 110(50), 20338–43.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. a., Nam, H., Feist, A. M., & Palsson, B. Ø. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. Molecular systems biology, 7(535), 535.

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? Nature biotechnology, 28(3), 245–248.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research, 42(D1), D206–14.

Pál, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature Genetics, 37(12), 1372–1375.

Paley, W. (1802). Natural theology.

Papoutsakis, E. T. (1984). Equations and calculations for fermentations of butyric acid bacteria. Biotechnology and Bioengineering, 26(2), 174–187.

Price, N. D., Reed, J. L., & Palsson, B. Ø. (2004). Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature reviews. Microbiology, 2(11), 886–97.

R Development Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Reed, J. L., Vo, T. D., Schilling, C. H., & Palsson, B. O. (2003). An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome biology, 4(9), R54.

Saint-Leger, A., Bello, C., Dans, P. D., Torres, A. G., Novoa, E. M., Camacho, N., Orozco, M., Kondrashov, F. A., & Ribas de Pouplana, L. (2016). Saturation of recognition elements blocks evolution of new tRNA identities. Science Advances, 2(4), e1501860–e1501860.

Satish Kumar, V., Dasika, M. S., & Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. BMC bioinformatics, 8(8), 212.

Schellenberger, J., Lewis, N. E., & Palsson, B. Ø. (2011). Elimination of Thermodynamically Infeasible Loops in Steady-State Metabolic Models. Biophysical Journal, 100(3), 544–553.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., & Palsson, B. Ø. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols, 6(9), 1290–1307.

Scheper, T., Klinkenberg, D., Pennartz, C., & van Pelt, J. (1999). A mathematical model for the intracellular circadian rhythm generator. The Journal of neuroscience : the official journal of the Society for Neuroscience, 19(1), 40–47.

Schilling, C. H. & Palsson, B. Ø. (2000). Assessment of the Metabolic Capabilities of Haemophilus influenzae Rd through a Genome-scale Pathway Analysis. Journal of Theoretical Biology, 203(3), 249–283.

Schmidt, S., Sunyaev, S., Bork, P., & Dandekar, T. (2003). Metabolites: A helping hand for pathway evolution? Trends in Biochemical Sciences, 28(6), 336–341.

Schultes, E. A. & Bartel, D. P. (2000). One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science, 289(5478), 448–452.

Schuster, S., Hilgetag, C., Woods, J. H., & Fell, D. A. (1996). Elementary modes of functioning in biochemical networks. In Computation

in cellular and molecular biological systems (pp. 151–165). World Scientific.

Suthers, P. F., Dasika, M. S., Kumar, V. S., Denisov, G., Glass, J. I., & Maranas, C. D. (2009). A Genome-Scale Metabolic Reconstruction of Mycoplasma genitalium, iPS189. PLoS Computational Biology, 5(2), e1000285.

Thiele, I. & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, 5(1), 93–121.

Van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. Trends in Genetics, 19(9), 479–484.

Wagner, A. (2008). Neutralism and selectionism: a network-based reconciliation. Nat. Rev. Genet. 9(12), 965–974.

Wagner, G. P. & Altenberg, L. (1996). Perspective: complex adaptations and the evolution of evolvability. Evolution, 50(3), 967–976.

Waldrop, M. M. (2016). The chips are down for Moore's law. Nature News, 530(7589), 144.

Walter, K. U., Vamvaca, K., & Hilvert, D. (2005). An active enzyme constructed from a 9-amino acid alphabet. Journal of Biological Chemistry, 280(45), 37742–37746.

Watson, M. R. (1984). Metabolic maps for the Apple II. Biochemical Society Transactions, 12, 1093–1094.

Watts, E., Hossfeld, U., Tolstikova, I., & Levit, G. (2016). Beyond borders: on the influence of the creationist movement on the educational landscape in the USA and Russia. Theory in Biosciences, 1–18.

Wiechert, W. & Noack, S. (2011). Mechanistic pathway modeling for industrial biotechnology: Challenging but worthwhile.

Wolf, Y. I. & Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. BioEssays, 35(9), 829–837.

Yuan, H., Cheung, C. Y. M., Hilbers, P., & Riel, N. V. (2016). Flux balance analysis of plant metabolism: the effect of biomass composition and model structure on model predictions. Frontiers in Plant Science, 7(April), 1–13.

Zanghellini, J., Ruckerbauer, D. E., Hanscho, M., & Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell: Properties, calculation and applications. Biotechnology Journal, 8(9), 1009–1016.

6 Manuscripts

This section includes four manuscripts with C.J.F. being first author, shared first author, or co-author. A copy of each manuscript is followed by a paragraph indicating the contributions to that manuscript. This is followed by a section giving some outlook on the development in this field.

6.1 *Manuscript 1*: Sybil – Efficient Constraint-Based Modelling in R

This manuscript is published as:
Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J., & Lercher, M. J. (2013). sybil - Efficient constraint-based modelling in R. BMC systems biology, 7(1), 125
Digital Object Identifier (DOI): 10.1186/1752-0509-7-125
URL: https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-125

6.1.1 Contributions

C.J.F. developed the sybil add-on package RSeed. C.J.F tested and applied sybil and made suggestions for improvements and additions. Currently, C.J.F. is maintaining the package on the Comprehensive R Archive Network (CRAN).

6.1.2 Outlook

Currently, sybil is used in many labs as software for constraint-based modelling. Additional packages were developed by other authors: *e.g.*, Desouki et al. (2015) implemented a method to remove futile cycles from the flux distribution, and Hartleb et al. (2016) built the GlobalFit (Section 5.1.4) algorithm within the *sybil* package. Further packages

extending sybil are sybilccFBA[1], sybilEFBA[2], and sybilDynFBA[3]. Additionally, the package BacArena[4] is in development. It can be used to simulate bacterial growth of different organisms and their interaction in time and space and uses sybil as its engine to perform the constraint-based calculations. All further presented manuscripts in this thesis are based on the package *sybil*. Also a manuscript about the integration of kinetic reaction parameters and non-metabolic processes into *sybil* is in preparation.

6.1.3 References

Desouki, A. A., Jarre, F., Gelius-Dietrich, G., & Lercher, M. J. (2015). CycleFreeFlux: efficient removal of thermodynamically infeasible loops from flux distributions. Bioinformatics (Oxford, England), 31(13), 2159–65.

Hartleb, D., Jarre, F., & Lercher, M. J. (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(8), e1005036.

Szappanos, B., Fritzemeier, J., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R. A., Lercher, M. J., Pál, C., & Papp, B. (2016). Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nature Communications, 7(11607).

[1] https://CRAN.R-project.org/package=sybilccFBA
[2] https://CRAN.R-project.org/package=sybilEFBA
[3] https://CRAN.R-project.org/package=sybilDynFBA
[4] https://github.com/euba/BacArena

6.2 *Manuscript 2*: Adaptive Evolution of Complex Innovations Through Stepwise Metabolic Niche Expansion

This manuscript is published as:
Szappanos, B., Fritzemeier, J., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R. A., Lercher, M. J., Pál, C., & Papp, B. (2016). Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nature Communications, 7(11607)
Digital Object Identifier (DOI): 10.1038/ncomms11607
URL: https://www.nature.com/articles/ncomms11607

6.2.1 Contributions

C.J.F. merged the *E.coli* metabolic model with the universal reaction set and conducted the analysis of growth promoting reactions sets in new environments. Growth media were defined by B.Z. and C.J.F. A draft manuscript was prepared by C.J.F. for the parts regarding his work. The manuscript was finalized jointly by all authors including C.J.F.

6.2.2 Outlook

Manuscript 2 presents a new evolutionary model that explains how changing environments accelerate the evolution of complex innovations. This was done by showing the ability of an *E. coli* strain to adapt to new environments. Surprisingly, in most cases *E. coli* needed three or less additional reactions to become viable in new environments that did not support growth beforehand. This raised the question if adaptability within and across species, possibly depending on genome size. This question is answered by *Manuscript 4*.

The work of *Manuscript 2* only considers adaptations from exactly one starting point to another new environment. Unclear is yet how adaptations happen from this point on and how different evolutionary trajectories

converge at one point. Thus, an interesting investigation would be to use a specific organism model as starting point and let this model adapt to one environment after another by adding and removing reactions. Repeating such a run with, first, a varying order of environments and, second, alternative parameters (*e.g.*, speed of changing environments, time for removal of unused reactions), would yield distinct genotypes. By choosing appropriate parameters to realistic evolutionary trajectories can be simulated and validated. Although the methodology is already founded in *Manuscript 2*, this analysis might be challenging to conduct because it is expected to be computational expensive. A theory contradicting the findings in *Manuscript 2* is the "White-Knight Hypothesis" (Wagner, 2016). The conceptual model posits that non-adaptive traits are important for the evolution of versatile organisms in nutritional-sparse environments. Although Wagner (2016) suggests wet lab experiments for validation of his conceptual model, simulations on the basis of *Manuscript 2* might be a reasonable alternative.

6.2.3 References

Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B., & Lercher, M. J. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Computational Biology, 13(4), e1005494.

Wagner, A. (2016). The White-Knight Hypothesis, or Does the Environment Limit Innovations? Trends in Ecology & Evolution, 20(2), 1–10.

6.3 *Manuscript 3*: Erroneous Energy-Generating Cycles in Published Genome-Scale Metabolic Networks: Identification and Removal

This manuscript is published as:
Fritzemeier, C. J., Hartleb, D., Szappanos, B., Papp, B., & Lercher, M. J.

(2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Computational Biology, 13(4), e1005494
Digital Object Identifier (DOI): 10.1371/journal.pcbi.1005494
URL: http://dx.plos.org/10.1371/journal.pcbi.1005494

6.3.1 Contributions

C.J.F. discovered the problem of energy generating cycles in published metabolic networks and performed all data collection. C.J.F. developed the method to detect EGCs. Together with D.H., C.J.F. designed the method to remove energy-generating cycles from metabolic network reconstructions. C.J.F. implemented and ran all computations, drafted the manuscript, and created the figures. The manuscript was finalized by C.J.F., D.H., and M.J.L.

6.3.2 Outlook

In *Manuscript 3* the authors point out a severe flaw in current protocols (Thiele & Palsson, 2010) and pipelines (Henry et al., 2010) for manual and automatic metabolic network reconstructions. The presented method for the detection of energy-generating cycles should become part of those pipelines and protocols to further improve the quality of metabolic models. As already pointed out in the paper, "the seed", the database where the most EGCs were detected, is filled with models from the ModelSEED-pipeline (Henry et al., 2010). An integration of the newly presented method into this pipeline could eliminate EGCs on the spot. Additionally, *Manuscript 3* raises the question whether FBA is the right method to handle simulations of metabolism in different environments correctly. It should be considered to publish, along with a GENRE, environment specific reaction directions that exclude EGCs.

6.3.3 References

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., & Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature biotechnology, 28(9), 977–82.

Thiele, I. & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, 5(1), 93–121.

6.4 *Manuscript 4*: Differences in the Adaptability of Generalist and Specialist Bacteria: the Influence of Metabolic Network Size and Structure

Differences in the adaptability of generalist and specialist bacteria: the influence of metabolic network size and structure

Claus Jonathan Fritzemeier[1*], Felix Lieder[2], Balázs Szappanos[3], Balázs Papp[3], Csaba Pál[3], Martin J. Lercher[1]

[1] Institute for Computer Science and Department of Biology, Heinrich Heine University, Universitätsstraße 1, D-40225 Düsseldorf, Germany

[2] Institute for Mathematics, Heinrich Heine University, Universitätsstraße 1, D-40225 Düsseldorf, Germany

[3] Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Temesvári krt. 62, Szeged H-6726, Hungary

* To whom correspondence should be addressed

Abstract

Bacteria show an astounding ability to adapt to new environments through horizontal gene transfer. Anecdotal evidence suggests that some bacteria are more adaptable than others; *e.g.*, strains of intestinal *E. coli* frequently spin off pathogenic strains adapted to other human tissues, while gastric *Helicobacter pylori* do not. However, it is unclear what determines the ability of individual strains or species to adapt to new environments. Here, we use pan-genome-scale modeling to explore the ability of 71 different unicellular organisms to adapt to each of 5000+ diverse nutritional environments. In agreement with previous simulations of abstract metabolic networks, we find that the number of viable environments scales approximately quadratically with the number of metabolic genes. While the smallest metabolic systems analyzed – those of the endosymbionts *Buchnera aphidicola* and *Helicobacter pylori* – require on average over 50 additional metabolic reactions to adapt to new environments, different strains of the generalist *E. coli* require on average less than 5 new reactions. When adapting to a new environment, organisms may "accidentally" acquire the ability to also grow in additional, non-selected environments. This effect of collateral adaptation is much stronger for generalist than for specialist organisms. Conversely, specialists are more likely than generalists to re-use (or exapt) previously acquired reactions for later adaptations, likely because of their more linear network structure. Thus, the ease with which microbes adapt to nutritional environments depends not only on how metabolically versatile they already are, but also on their metabolic network structure.

Introduction

Unicellular organisms show an astounding ability to adapt to new environments (1). Different phylogenetic lineages differ widely in the frequency with which they give rise to new strains or even new species, but it is currently unclear what determines these differences. Speciation may occur through reproductive isolation, *i.e.,* through strongly reduced genomic recombination. Conversely, speciation (or the generation of distinct strains) may be adaptive, with the new lineage specializing to a different life style or environment. Among bacteria, such specialization is typically accompanied by the the loss of now superfluous genes from the genome and/or the acquisition of additional genes via horizontal gene transfer (HGT, also called lateral gene transfer) (2, 3). Examples in point are the loss of a majority of metabolic genes in the endosymbitic *Buchnera* species (4) and the acquisition of additional metabolic pathways by pathogenic *E. coli* strains to survive in the human urinary tract (5). As a consequence of these evolutionary dynamics, bacterial pan-genomes can be partitioned into core genes (found in almost all strains), shell genes (found in several strains), and cloud genes (which are strain-specific) (2).

Bacterial strains often differ widely in their metabolic capabilities; *e.g.*, a study comparing strains of *E. coli* found individual strains be able to grow in between 437 and 624 of the tested environments (6). Based on such differences, lineages can be categorized as metabolic generalists or specialists. A prolonged reduction in environmental complexity – as that experienced by a generalist bacterium becoming a permanent endosymbiont – causes a corresponding reduction in metabolic complexity, which can be predicted quantitatively from genome-scale metabolic modeling (4). That bacterial evolution appears to organize itself into short bursts of innovation followed by long phases of genome reduction (7) indicates that the inverse process – a specialist evolving into a generalist – is comparatively rare.

In previous work (8), we utilized metabolic simulations to show that the standard lab strain *E. coli* K-12 can adapt to most previously unviable nutritional environments by acquiring at most three additional enzymes and/or transporters via HGT. In many cases, different new environments required the acquisition of overlapping gene sets; we found that complex metabolic innovations can evolve via the successive acquisition of single biochemical reactions that each confers a benefit to utilize specific nutrients. This demonstrates an important role of exaptations in metabolic evolution, where stepwise metabolic niche expansion can accelerate adaptation substantially (8). These findings also demonstrated that complex innovations can evolve without the need to resort to neutral explorations of phenotype space, as had been suggested earlier (9). Such non-adaptive evolution is not only expected to be extremely slow; there is also no direct empirical support for this scenario in bacteria (8).Moreover, support for the hypothesis is based on simulations that are biologically highly unrealistic, as *E. coli* is evolved *in silico* over prolonged periods of time in a single minimal environment without a reduction in genome size: unneeded reactions that are lost from the genome are always replaced by new, frequently also unselected reactions via HGT (9). Thus, it appears that bacterial evolution can only be understood from a consideration of adaptive processes (8).

Is the ease of metabolic adaptation and the potential for exaptations observed previously (8) unique to *E. coli*, is it typical of metabolic generalists, or is it representative of bacterial adaptation in general? What determines a lineage's frequency with which it generates metabolic innovations, *i.e.*, the ability to grow in a previously unviable environment? Abstract mathematical models of metabolic network expansion suggest that metabolic network size plays a crucial role for the answers to these questions. Based on an analysis of branching pathways, Maslov *et al.* showed that the number of distinct carbon

sources that can be converted to a central biomass component grows roughly quadratically with network size (10), an observation in agreement with the quadratic scaling between the numbers of transcription factors and enzymes (11, 12). Maslov *et al.* illustrate this behavior through the analogy of metabolic tools in a genomic toolbox. Many tools (enzymes or transporters) can be used in different contexts (environments); thus, an organism with a well-filled toolbox typically requires a much smaller number of additional tools to perform a given task than a competitor with an emptier toolbox (10). Accordingly, we expect that metabolic generalists, which possess larger and more versatile metabolic systems, are better "pre-adapted" to new, previously unselected environments – either because they are already viable without the necessity to acquire any additional reactions, or because they require a smaller number of additional genes compared to competing specialists. However, a study examining how metabolic network size and structure affect the adaptability of real metabolic systems is lacking.

Below, we use careful metabolic simulations on a pan-genome-scale to show that the ease with which microbes adapt to new environments varies widely between species, with metabolic specialists typically requiring an order of magnitude more gene acquisitions than generalists adapting to the same environment. The increased adaptability of generalists is emphasized by their much higher potential for collateral adaptations, *i.e.*, the ability to grow in additional, non-selected environments due to ecologically unrelated previous adaptations. Specialist species, on the other hand, have largely lost their adaptive potential. If they do adapt, however, they show a stronger tendency of exaptation, *i.e.*, they are more likely to re-use previously acquired enzymes and transporters for later adaptations, likely because of their more linear metabolic network structures.

Results and Discussion

Construction of a pan-genome scale metabolic supermodel from organism-specific models

To allow coherent simulations of metabolic network expansion through HGT, we first created a pan-genome-scale metabolic supermodel that contained all examined organism-specific metabolic networks as submodels. The supermodel built from 76 organismal metabolic models contains 7768 unique reactions, 4040 unique metabolites and is parted into thirteen compartments. Fig. 1 shows the sizes of the organism submodels included. In Fig. 2a is the size of compartments shown by metabolite number. Most metabolites are assigned to the compartments extracellular (e), periplasm (p), and cytosol (c). The high number of compartments originates from the three eukaryotic organisms and a cyanobacterium in this model: *Chlamydomonas* (iRC1080), *Saccharomyces cerevisiae* (iMM904, iND750) and *Synechocystis* sp. PCC 6803 (iJN678). For the well-studied *Escherichia coli* str. K-12 substr. MG1655 are five models integrated. Additionally 55 models are from one earlier work and consist of *Escherichia coli* and *Shigella* strains (6). Further details about the used organisms and their GSM properties can be taken from Supplementary Table S1.

Analogous to the partitioning of pan-genomes (2), the reactions in the supermodel can be broken down into ubiquitous "core" reactions, a "shell" of reactions present in many organisms, and a "cloud" of organism-specific reactions only present in one submodel (Fig. 2). 70 out of 97 reactions (72%) of the *E. coli* core network (e_coli_core) are found in more than 60 of the submodels, constituting a substantial fraction of the metabolic core (Fig. 2).

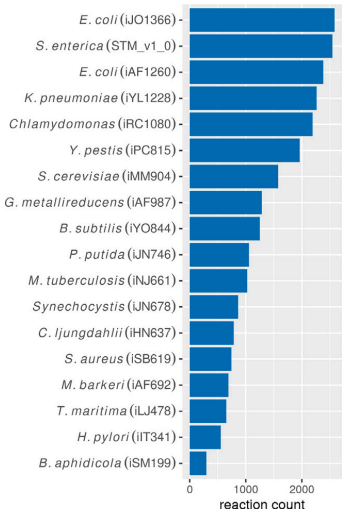


Figure 1. Number of metabolic reactions of the organisms used to build the pan-genome-scale supermodel. The figure shows only one representative strain per species, except for *E. coli*, which is represented by a second model (iAF1260). Supplementary Figure S1 shows all models.

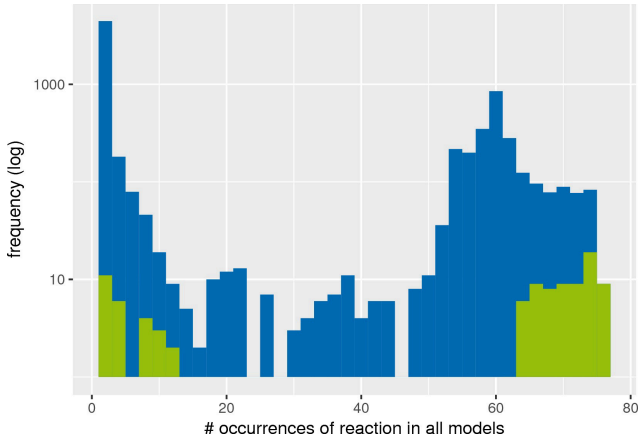


Figure 2. The reactions of the metabolic supermodel can be roughly partitioned according to their occurrence frequencies into ubiquitous “core”, intermediate “shell”, and organism-specific “cloud” reactions (2). The histogram shows reaction counts over all models (blue, log-scale), ignoring reaction directionality. The leftmost reactions each occur in only a single model, while the rightmost reactions are ubiquitous. The frequency of reactions contained in the model of *E. coli* core metabolism across all examined models are shown as a green overlay. Some *E. coli* core reactions occur in less than 13 of the submodels; this is an artifact of the missing compartmentalization of the core model.

Distinguishing metabolic generalists and specialists based on metabolic simulations

We used flux balance analysis (FBA) (13) to estimate the ability of each submodel to grow in a wide set of nutritional environments. To make results comparable, we used the same "general" biomass reaction for all organism-specific submodels, *i.e.*, each metabolic system was required to produce the same metabolic precursors for cellular growth. We examined two sets of nutritional environments: one set that largely contains typical wet lab growth media (14), including those assayed in the Biolog phenotyping system, and another set of random combinations of carbon, nitrogen, sulfur, phosphorus plus trace elements.

As most models cannot grow in any of the random minimal environments, we checked whether all models can grow in a medium that supplies all possible nutrients. Only three models are not viable in this maximally rich condition: the *E. coli* core metabolism (e_coli_core), the hyperthermophilic bacterium *Thermotoga maritima* (iLJ478), and the endosymbiotic bacterium *Buchnera aphidicola* (iSM199). This is because the general biomass objective function contains more amino acids than the original biomass functions of these models.

The minimal random environments provide limited insights into the growth of the submodels in the real world, as these environments are too restricted for most modeled organisms (Fig. 3). Fig. 3 also shows the fraction of viable wet lab environments for each submodel, ranging from 0% to about 36%. The corresponding distribution (Supplementary Fig. S4) is bimodal, naturally dividing these organisms into generalists and specialists, with the dividing line defined here at growth in 18% of assayed media. Note that the models unable to grow in any wet lab environment are those unable to grow in the full medium.

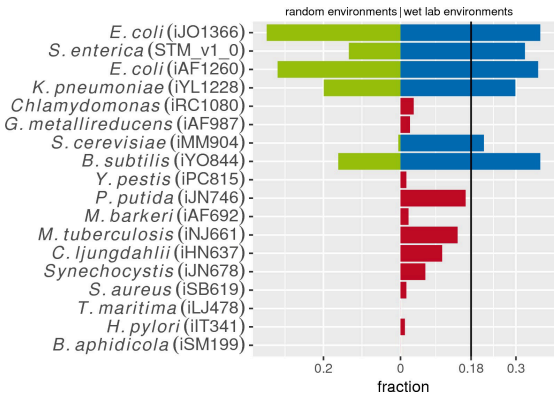


Figure 3. The fraction of viable environments differs widely across submodels, both for random minimal environments (green bars to the left) and for common wet lab environments (blue and red bars to the right). The vertical line indicates the mean fraction of viable wet lab environments; we use it as the threshold for partitioning metabolic systems into generalists (blue) and specialists (red). Models are ordered by top to bottom by decreasing genome size.

Network size predicts adaptability

Every organism is well adapted to a certain set of environments and we now measure the effort needed for an organism to adapt to a new environment in which it was not viable beforehand. This is done by finding the minimal number of reactions that have to be added to an organism-specific submodel in a given environment; below, we refer to this number as "added reactions". The distribution of added reactions varies widely between organisms (Fig. 4a; for a figure including all *E. coli* strains see Fig. Supplementary Fig. S6).

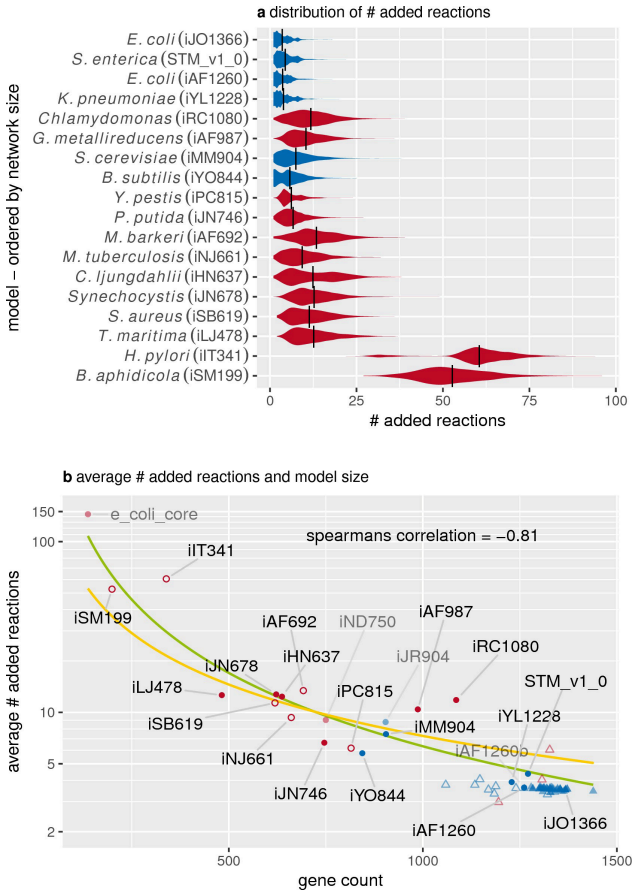


Figure 4. The number of additional reactions required for adaptation decreases with increasing genome size. The colors of circles and points distinguish specialists (red) and generalists (blue). The 55 *E. coli* strains are plotted as triangles. Organisms with a known auxotrophies are shown as open circles. Shaded points are organisms represented by multiple metabolic models. **a)** Distributions of added reactions as violin plots. The width of each "violin" indicates the local density of the distribution, normalized for each model. Supplementary Figure S3 is equivalent, but contains all models of species represented multiple times. **b)** Correlation of the average number of added reactions (log scale) and the gene count for each model. Calculation of Spearman correlation and curve fitting were performed including only iJO1366 as a representative of the 55 *E. coli* strains. A close-up from plot (b) focusing on the 55 *E. coli* strains is shown as Supplementary figure S5.

6.4 *Manuscript 4*: Differences in the Adaptability of Generalist & Specialist Bacteria: the Influence of Metabolic Network Size & Structure

The two smallest and most specialized metabolic networks require the largest number of added reactions to adapt to new environments. The endosymbiont *Buchnera aphidicola* needs at least 27 reactions and on average 52.69 reactions to reach new environments. Similarly, the pathogen *Helicobacter pylori*, which exclusively lives in human stomachs, needs at least 22 and on average 60.52 reactions. Both organisms are highly specialized to a single, relatively stable environment. Accordingly their metabolisms show very little flexibility, reflected in very small genomes (*B. aphidicola:* 199 metabolic genes out of a total of 517 genes (15); *H. pylori:* 341 metabolic genes out of 1590 total genes (16)). At the other end of the spectrum is *E. coli* (Fig. 4a): the standard lab strain K12 (iJO1366) requires on average 3.51 and at most 18 reactions to adapt to any of the tested environments.

Although it is likely that many properties influence the ability of a metabolic system to adapt to new nutritional environments, network size alone explains 65% of the variance across all assayed models (Fig. 4b; Spearman's ρ=-0.81, *P*=9.092e-05).

The dataset contains *E. coli* models of various sizes (1059 to 1439 metabolic genes). These mostly differ only marginally in their adaptability: the average number of added reactions for generalist *E. coli* (including 9 strains with auxotrophies) lies between 3.31 and 4.05, while the average number of added reactions for specialist *E. coli* ranges from 2.98 to 6.04. The outlier requiring the largest number of additional reactions is *E. coli* str. K-12 substr. DH10B (iECDH10B_1368; Supplementary Fig. S5), which is auxotrophic for leucine due to the loss of a complete operon (6). For completeness, we also performed calculations for the incomplete *E. coli* core metabolic model (e_coli_core). As expected, this model requires much larger numbers of reactions (over 100) to achieve the same adaptations as the genome-scale *E. coli* model (iAF1260).

Numbers of added reactions agree with the observed scaling of transcription factor number and toolbox model predictions

Maslov *et al.* used an abstract branching process to estimate the dependence of the number of metabolic pathways on network size for a single type of nutrient (10). They predicted a quadratic scaling between the number of utilizable nutrients, $N_{Nutrients}$ (*i.e.,* the number of independent pathways or, by inference, of transcription factors) and the number of genes, N_{Genes}:

$$N_{Nutrients} = c\, N_{Genes}^2$$

with a constant *c*; this finding agreed with previous observations of a quadratic scaling between transcription factor and metabolic gene numbers (11, 12). The number of added reactions per additional nutrient can be interpreted as the derivative of the inverted relationship, expressed as a function of N_{Genes}:

$$N_{Genes} = \sqrt{\frac{N_{Nutrients}}{c}}$$

$$\Rightarrow \frac{\partial N_{Genes}}{\partial N_{Nutrients}} = \frac{1}{2\sqrt{c}}\, N_{Nutrients}^{-1/2} = \frac{1}{2\, c\, N_{Genes}}$$

By choosing an appropriate value for the constant c, this equation can predict the number of additional genes as a function of the number of metabolic genes; this curve is shown in Fig. 4b for the best fitting *c*=6.86e-05. The mean squared error of the measured data relative to the fitted curve is

MSE=0.16 (to avoid biasing the curve towards *E. coli*, we represented *E. coli* through a single model, iJO1366, when fitting the curve).

Instead of assuming a quadratic scaling between gene and nutrient numbers, we can explore general power law relationships, where the exponent 2 is replaced by an arbitrary number α. The best fitting α is 2.43 (95% CI = [2.03, 2.82]), which reduces the MSE to 0.11 (Fig. 4b). This best fitting α is slightly larger than the α=2 expected from abstract models of metabolic expansion (10). One possible reason for this disagreement is the artificial, simple network structure of the abstract model (10). In addition, our data comes from considering adaptations to new environments (consisting of 4 nutrient types) rather than to single new nutrients; thus, the quadratic scaling may slightly underestimate the number of additional reactions per new environment for small genomes and overestimate this number for larger genomes.

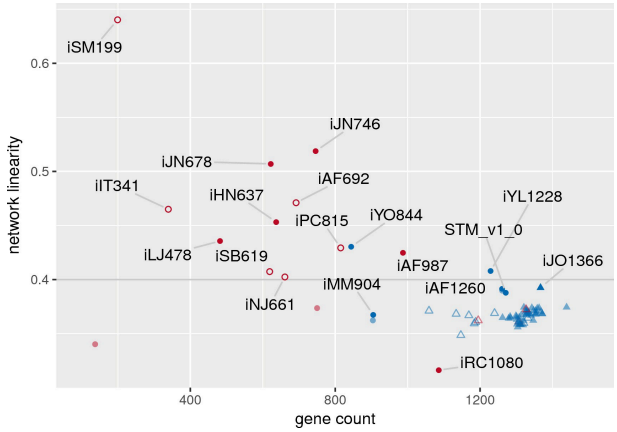


Figure 5. Small, specialist metabolic networks are less branched than large generalist networks. Network linearity is defined as the fraction of metabolites that participate in only two reactions, *i.e.*, metabolites that are intermediates in unbranched pathways. Most models with more than 40% unbranching metabolites are specialists. The colors of circles and points distinguish specialists (red) and generalists (blue). The 55 *E. coli* strains are plotted as triangles. Organisms with a known auxotrophies are shown as open circles. Shaded points are organisms represented by multiple metabolic models. Spearman correlation between network linearity and gene count: ρ = -0.68, using only iJO1366 as representative for the 55 *E. coli*.

Generalists exhibit more collateral adaptations than specialists

One way to quantify an organisms metabolic adaptability is to ask what fraction of previously unviable environments can be reached at no additional cost after a given adaptation (9); in such cases, the added reactions can be seen as exaptations for the additional, non-selected environments. If this "collateral adaptation index" (see methods) is one, the model is viable in all previously unreachable environments, while a value of zero indicates that no other environment is reachable without additional reactions. We generally found higher collateral adaptation values for generalists (median 0.13, Supplementary Figure S7) than for specialists (median 0.08).

Thus, generalists not only require fewer additional reactions than specialists to reach a new environment, these reactions are also more valuable to them, as they simultaneously provide access to a substantial number of other, previously unviable environments. Specialists, in contrast, tend to have a

smaller genome size, need more additional reactions to adapt, and are less likely to profit from collateral adaptations.

The smallest specialist genomes are most likely to re-use gained reactions in later adaptations

Even if the reactions acquired to adapt to environment *A* do not provide immediate access to environment *B*, they may still provide a subset of the reactions required to adapt later to this second environment. We quantify the propensity to profit from adaptations in this way with an "exaptation index" (see Methods). One might hypothesize that while specialists show little collateral adaptation, they may show a high exaptation index, *e.g.*, if added reactions remove an auxotrophy. As expected, we find a higher exaptation index for specialists with small than for specialists with intermediate genome sizes (Supplementary Figure S8).

Specialist metabolic networks are less branched and have lower adaptability

One function of branching points in metabolic networks is to link alternative pathways to central metabolism. Thus, specialist metabolic networks not only have fewer reactions, but may also have fewer branching points, i.e., they may be more linear in their structure. In a linear pathway, primary metabolites participate in two reactions only, with one reaction producing and one consuming the metabolite. To represent a branching point, a primary metabolite has to participate in at least three reactions. We thus define "network linearity" as the fraction of metabolites participating in exactly two reactions. This measurement again reflects the dichotomy of generalists and specialists (Figure 5). Most organisms that have more than 40% of their metabolites in linear pathways are specialists, while those with a lower network linearity index are generalists; this bisection largely coincides with a bisection based on genome size, dividing specialists from generalists at around 800 genes (Figure 5). As expected, the compact but highly branched *E. coli* core metabolism (e_coli_core) exhibits a lower network linearity than any complete *E. coli* model. Interestingly, *Chlamydomonas* (iRC1080) shows the lowest network linearity of all organisms. It is also noteworthy that some specialist *E. coli* cannot be distinguished from generalist *E. coli* in terms of network linearity, as their genomes differ in only a small number of genes and their general metabolic network structure is identical.

Conclusions

Our analysis of potential adaptations of 71 microbes to thousands of different nutritional environments demonstrate quantitatively that small, specialist genomes are much less adaptable than the large genomes of organisms with a well-filled metabolic toolbox that already live a generalist life style; the smallest metabolic systems, those of *Buchnera aphidicola* and *Helicobacter pylori*, are trapped in their endosymbiotic life style, having all but lost their adaptive potential. However, adaptive potential is not only a function of genome (or metabolic network) size; it is also strongly correlated to metabolic network structure, with highly branched systems requiring lower numbers of additional genes to become viable in a new environment.

Exaptation – the utilization of metabolic genes acquired in previous adaptations for adaptive purposes in a new environment – plays an important role in the adaptation of both generalists and specialists, although in different ways. Generalists, but not specialists, show a high degree of collateral adaptation, *i.e.,* previous adaptations often enable growth in environments other than those experienced by the organisms ancestors. Conversely, specialists that acquire new metabolic genes in the adaptation to one environment are more likely to re-use (exapt) these genes in later adaptations to other environ-

ments; thus, stepwise metabolic niche expansion will play an even stronger role in the adaptation of specialists than previously observed for the generalist *E. coli* (8).

Materials and Methods

Supermodel generation

At the beginning of this work were 79 GSM available at the BiGG database (17). Four models of these representing mouse and human were excluded. In contrast the model for *Buchnera aphidicola* str. APS (18) was modified and added. So starting point of the supermodel are 76 genome scale metabolic models (Supplementary Table S1). All models are tested beforehand whether reactions and metabolites with same ids actually represent the same reaction and renamed if necessary. Reactions were compared on basis of coefficients and reversibility, *i.e.* the lower and upper bound, and metabolites were considered different on basis of the chemical formulae, if known. The merging process itself is simply done by pooling reactions and metabolites from all organisms and building a new model from that. In this generally working supermodel mass balance has to be ensured and energy generating cycles (EGC) have to be removed (19). While each individual model passes these quality checks, the reactions in the merged supermodel may be combined in a way that violates thermodynamic laws or the mass balance. Mass balance is considered first, because the EGC removal cannot be done without proper mass balance.

Correction of mass balance

Mass balance of a reaction is generally ensured by contrasting all atoms of the educts and all atoms of the products. But, due to incomplete data the mass balance for many reactions is not known and removing all reactions with uncertain mass balance will render most of the models non-functional. To circumvent this, only mass balance at the exchange reactions was considered. Namely, we ensure the FBA steady state condition on basis of atoms: the number of atoms of the same kind (*e.g.* carbon) entering the model has to be even to the number of (*e.g.* carbon) atoms leaving the model. The only reactions that allow exchange of metabolites and therewith atoms are exchange reactions and biomass reactions. At the same time these are the only reactions in a network that are allowed to be imbalanced. By fixing the exchange of atoms to zero and determining the blocked reactions, *i.e.* reactions that always have a zero flux, potentially imbalanced reactions are identified. This is because the overall mass exchange of the model has to be balanced and a single imbalanced reaction would disturb this balance. Of course some properly balanced reactions will be blocked as well, but these reactions will be blocked anyway, because they share the same elementary modes (20) with the removed imbalanced reactions. Exchange reactions and biomass objective function that contain a metabolite of unknown composition are removed from the model, too.

Removing erroneous energy-generating cycles

Another problem occurring when combining multiple GSMs is the formation of erroneous energy generating cycles (EGC) (8, 19). These cycles can produce energy, *e.g.* ATP, in infinite amounts without consumption of nutrients (19). Even the combination of two networks that are both free of EGCs itself can cause the formation of those. We adapted the method from previous work (8) to remove these EGCs from the multi organism model. First, the set of reactions of only one organism is considered. If this set is free of EGCs, the reactions of the next organism are added to the set. This is repeated for all organism models.

The method calculates the smallest set of reactions to form such a cycle. This problem was solved in previous work with the ARM MILP algorithm, but here the ARM LP was used (see methods, active reaction minimization). One reaction in this cycle is constrained in the used direction, *i.e.* reversible reactions are made irreversible and irreversible reactions are deleted. Then the process starts over until no more cycles can be found. The algorithm is run with every submodel as starting point, because it can notably affect the metabolism and the order of adding models is crucial. This results in a set of EGC free reactions for each organism. The order of adding organisms is determined by the initial number of EGCs; models with less EGCs are added first.

The combination of the blocked reactions from the first step and the reactions that are free of EGCs results in a properly curated supermodel.

Active Reaction Minimization

Mixed integer linear programs (MILP) are frequently used to extend FBA, *e.g.* ROOM (21), or gapfind and gapfill(22). In many of those problems the objective is to minimize the number of active reactions. Thus we call this problem active reaction minimization (ARM). The pan-genomic-scale model in this work is much bigger than the usually used genome-scale models. Current methods of minimizing the number of active reactions under flux balance constraints cannot be applied due to the exponential complexity of those. We use here a method to approximate this calculation with major speedup and minor inaccuracy. Recent work describes how to formulate such a linear approximation in combination with the Gapfill algorithm (23).

Here, we describe an approximation to solve this ARM problem efficiently, by relaxing the following ARM MILP problem into a sequence of ARM LP^k for $k \in \{1, \dots, n\}$ problems. We use the property of the simplex algorithm to find sparse solution vectors.

ARM MILP:

$$\min\left(\sum_{i \in B} b_i\right)$$
$$s.t.:$$
$$S * v = 0$$
$$l_i \leq v_i \leq u_i \; \forall i \in R$$
$$v_i \neq 0 \Rightarrow b_i = 1 \; \forall \, i \in B$$

ARM LP^k:

$$\min\left(\sum_{i \in B} \left| v_i * \frac{1}{\max\left(\varepsilon_i^k, |v_i^{k-1}|\right)} \right|\right)$$
$$s.t.:$$
$$S * v = 0$$
$$l_i \leq v_i \leq u_i \; \forall i \in R$$

$S \in \mathbb{R}^{\lvert M \rvert} \times \mathbb{R}^{\lvert R \rvert}$	stoichiometric matrix
$v \in \mathbb{R}^{\lvert R \rvert}$	Vector of fluxes
$l \in \mathbb{R}^{\lvert R \rvert}$	Vector of lower bounds
$u \in \mathbb{R}^{\lvert R \rvert}$	Vector of upper bounds

$b \in \{0,1\}^{\lvert B \rvert}$	Vector of binary variables
$R \in \{1, \dots, m\}$	Set of m reaction indices
$B \subseteq R$	Set of indices that are objective of the optimization
v_i^k	Flux of the i-th reaction in the k-th optimization (0 if undefined)
k	Optimization step counter
n	Total number of optimization steps
ε_i^k	i-th upper bound of weight factor in optimization round k

Table 1. Definition of variables of the ARM LP.

In this sequence of linear problems, the optimization function of the $(k+1)$-th problem is reweighted with the solution of the k-th problem. The initial values for ε_i^0 are either set to one or some positive random values. After each LP optimization it is recalculated to $\varepsilon_i^{k+1} := \varepsilon_i^k * {}^1/_{10}$ for the $(k+1)$-th optimization.

In order to show the practical application of our linear approximation of active reaction minimization, we show the comparison between the MILP result and the LP approximation. To limit the computation time to a reasonable span, we allowed the solver for the MILP eight parallel threads per problem and a maximum time of two minutes per problem. Thus, some results are suboptimal, but the gap value accounts for the maximal possible difference to the optimal value. For the ARM LP calculations, the linear problem was solved twelve times and the best solution was kept. After every fourth optimization, ε was reinitialized with random values and v is set as undefined.

Supplementary Fig. S9a shows both results in direct comparison for a total of 2830 problems we solved with the *E. coli* model iAF1260 and the standard biomass reaction. The ARM LP performs better for some problems with the non-optimal MILP solutions, *i.e.* with a gap greater zero. This is also the case for the exact solutions. We suspect the solver to have some numerical issues and thus to give a non-optimal solution in four MILP cases. All results were successfully verified with FBA. The differences between the pairwise results are shown in Supplementary Fig. S9b. For over 50% (1587 of 2830) of the problems the ARM LP found a better or equally good optimal value and 75% of the problems were at maximum by two reactions off the MILP ARM solution. The ARM MILP was solved with a time limit of two minutes with eight parallel threads. Notably in four cases the LP ARM found a better solution than the MILP ARM. As the CPLEX code is proprietary, it is impossible to find a reason for this. Without a given time limit the MILP ARM the computation time varies widely from seconds to hours, while the LP ARM is solved in split seconds.

Environmental Distances

The environmental distance is the number of reactions an organism has to obtain in order to survive in an environment that does not support life beforehand. The environment is defined as the set of nutrients available for growth and viability is defined to be a biomass production above the threshold of 0.01. This calculation is dependent of two major factors: the definition of the environments, *i.e.* growth media and the choice of the biomass objective function for a model.

Sets of environments

Each environment consists of a set of nutrients that can either be present or absent. One set is taken from the Seed database (14) and represents wet lab growth media. The second set growth media derived from a minimal growth medium for *E. coli* model iAF1260 and consist of one carbon, nitrogen, sulfur, and phosphorous (CNPS) source accompanied by growth essential trace elements (8). The variety of minimal growth media is achieved by exchanging one of the CNPS-nutrients a time. The third set of growth media is generated by building all combinations of nutrients from each CNPS-category. Because there are nearly $77 * 10^6$ possible combinations, we have randomly chosen 5000 of them.

Biomass objective functions

In each model is at least one biomass reaction defined and for further analysis exactly one of those was selected. The biomass reaction of some models may be blocked in the mass balance step. Those do not have a valid biomass reaction left. In addition we derived from the iAF1260 biomass reaction a general biomass reaction, containing only ribose nucleotides and deoxyribose nucleotides, amino acids, water, and energy. Last, also a biomass reaction is considered that only dissipates ATP and thus indicates if a model is able to produce energy from the nutrients. All combinations of environment types and biomass objective functions can be seen in Supplementary Fig. S3.

Calculation of environmental distance with ARM LP

The calculation of the environmental distance can be split into two parts. First, for each organism we have the organism's metabolic network. Formally this is a subnetwork of the new proposed supermodel. With normal FBA is the biomass production for the subnetwork and the supermodel calculated in each environment. For further analyses are only environments considered that support growth, *i.e.* have a biomass production above a threshold of 0.01, for the supermodel, but do not for the submodels. Secondly, with the LP ARM algorithm is the number of reactions calculated that has to be added to the subnetwork to make it viable in the considered environment. Both steps were done for each organism, environment, and biomass reaction, but results are only shown for the random minimal environments and the general biomass objective function.

Collateral Adaptation Index

The collateral adaptation index is the ratio of environments an adapted organism can reach without additional adaptation and the total number of environments. This definition was similarly described elsewhere (9). Additionally, the collateral adaptation index considers only environments that are distinct from the source environment, *i.e.* none of the carbon, nitrogen, sulfur or phosphate source from the adapted environment is contained in the tested environments. Hence, calculations of the viability for all pairs of environments, for all organism, and biomass reactions are necessary. This results in 2.1 billion FBA calculations. The collateral adaptation index can be seen as a phenotypic measurement for exaptation.

Exaptation Index

Contrary, the exaptation index is defined as a genotypic measurement. Given the organism is adapted to a new environment m_1, it needs the additional reaction set r_1 to survive there. For a second distinct

environment m_2 it needs the set r_2. The fraction of preadapted reactions f_{m_1,m_2} is then $\frac{|r_1 \cap r_2|}{|r_2|}$. Then we define the exaptation index

$$e_{m_j} := mean\left(f_{m_i,m_j}\right) \; \forall \; m_i \in M$$

with M being the environments distinct from m_i. Thus an exaptation index e_{m_j} of 0.5 means the organism already acquired on average half of the needed reactions.

Hardware, Software

All calculations were computed with the constraint based modelling package "sybil" in GNU R and IBM ILOG CPLEX as solver. Calculations were done on a compute cluster with a peak usage of about 600 CPUs. The whole process is implemented as a pipeline reducing human interaction to a minimum. Frequent control points ensure data integrity and correctness of calculations.

Acknowledgements

We are grateful for computational support through the Zentrum für Informations- und Medientechnologie (ZIM) at Heinrich Heine University Düsseldorf. We gratefully acknowledge financial support by the German Research Foundation (DFG grants IRTG 1515 supporting CJF; EXC 1028, and CRC 680 to MJL) and through an ENORM graduate school fellowship to CJF.

References

1. Brooks AN, Turkarslan S, Beer KD, Lo FY, Baliga NS. Adaptation of cells to new environments. Wiley Interdiscip Rev Syst Biol Med. 2011;3(5):544-61.

2. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36(21):6688-719.

3. Pal C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 2005;37(12):1372-5.

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.

5. Alteri CJ, Smith SN, Mobley HL. Fitness of Escherichia coli during urinary tract infection requires gluconeogenesis and the TCA cycle. PLoS Pathog. 2009;5(5):e1000448.

6. 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.

7. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. Bioessays. 2013;35(9):829-37.

8. Szappanos B, Fritzemeier J, Csorgo B, Lazar V, Lu X, Fekete G, et al. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun. 2016;7:11607.

9. Barve A, Wagner A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature. 2013;500(7461):203-6.

10. Maslov S, Krishna S, Pang TY, Sneppen K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proc Natl Acad Sci U S A. 2009;106(24):9743-8.

11. Croft LJ, Lercher MJ, Gagen MJ, Mattick JS. Is prokaryotic complexity limited by accelerated growth inregulatory overhead? Genome Biology. 2003;5(1):P2.

12. van Nimwegen E. Scaling laws in the functional content of genomes. Trends Genet. 2003;19(9):479-84.

13. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Biotech. 2010;28(3):245-8.

14. 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.

15. Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H. Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature. 2000;407(6800):81-6.

16. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, et al. The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature. 1997;388(6642):539-47.

17. 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.

18. Macdonald SJ, Lin GG, Russell CW, Thomas GH, Douglas AE. The central role of the host cell in symbiotic nitrogen metabolism. Proc Biol Sci. 2012;279(1740):2965-73.

19. 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;13(4):e1005494.

20. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C. Elementary flux modes in a nutshell: properties, calculation and applications. Biotechnol J. 2013;8(9):1009-16.

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

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

23. Thiele I, Vlassis N, Fleming RM. fastGapFill: efficient gap filling in metabolic networks. Bioinformatics. 2014;30(17):2529-31.

Supporting Information

S1 Table. Organism specific models and their properties. The models in this table are sorted in ascending order by the number of metabolic genes. The column "55 *E.coli*" indicates, whether models from the publication Monk et al. (2013).

S1 Figure. Networks sizes (number of reactions) of all models combined in the supermodel. The 55 *E. coli* models shown in an extra group and are depicted in lighter shade. The taxonomy ID refers to the NCBI taxonomy and the PubMed ID refers to the respective publication of the model.

S2 Figure. Additional properties of the supermodel. **a)** Number of metabolites in the compartments. The compartment with the most reactions is the cytosol (c) followed by the extracellular (e) and periplasm (p). The Remaining compartment originate from the eukaryotic models used. **b)** Barchart of published models per year. Here are only the 18 models of unique organisms counted, which were created by about two publications per year.

S3 Figure. Growth for submodels and supermodel in all three environment types (minimal environments, random minimal environments, and wet lab (seed) environments) and with three types of biomass objective functions (energy production, general biomass, and organism specific).

S4 Figure. Fractions of viable environments for submodels in wet lab environments (seed). The vertical line indicates the threshold to split models into specialists (red) and generalists (blue).

S5 Figure. The variation of added reactions within *E. coli* species is low, although the genome size varies. Specialists and generalists are distinguished by colors red and blue, respectively. Organism with known auxotrophy are shown with hollow points.

S6 Figure. Distributions of added reactions per submodel. Models in the groups on the y-axis are sorted by gene count. Specialists and generalists are distinguished by colors red and blue, respectively. The mean of each distribution is marked with a vertical line. Bar widths are normalized for each model.

S7 Figure. Distributions of the collateral adaptation index per submodel. Models in the groups on the y-axis are sorted by gene count. Specialists and generalists are distinguished by colors red and blue, respectively. The mean of each distribution is marked with a vertical line. Bar widths are normalized for each model.

S8 Figure. Distributions of the exaptation index per submodel. Models in the groups on the y-axis are sorted by gene count. Specialists and generalists are distinguished by colors red and blue, respectively. The mean of each distribution is marked with a vertical line. Bar widths are normalized for each model.

S9 Figure. Practical application of ARM LP shows equally good performance as ARM MILP. **a)** Result (objective value) comparison of ARM MILP and ARM LP. Dot colour indicates the gap size (smaller is better). In the left panel, ARM MILP solutions are suboptimal due to the limited computation time. Results shown in the right panel could be solved exactly within the time limit. The blue lines indicate equal objective values. **b)** Distribution of the difference between ARM LP and ARM MILP results.

Table S1. Organism specific models and their properties

Model ID	Organism	Gene count	Metabolite count	Reaction count	55 *E.coli*	Taxonomy ID	PubMed ID
e_coli_core	Escherichia coli str. K-12 substr. MG1655	137	72	95		511145	26443778
iSM199	Buchnera aphidicola str. APS	199	298	297		107806	22513857
iIT341	Helicobacter pylori 26695	339	485	554		85962	16077130
iLJ478	Thermotoga maritima MSB8	482	570	652		243274	19762644
iSB619	Staphylococcus aureus subsp. aureus N315	619	655	743		158879	15752426
iJN678	Synechocystis sp. PCC 6803	622	795	863		1148	22308420
iHN637	Clostridium ljungdahlii DSM 13528	637	698	785		748727	24274140
iNJ661	Mycobacterium tuberculosis H37Rv	661	826	1025		83332	17555602
iAF692	Methanosarcina barkeri str. Fusaro	692	628	690		269797	16738551
iJN746	Pseudomonas putida KT2440	746	909	1056		160488	18793442
iND750	Saccharomyces cerevisiae S288c	750	1059	1266		559292	15197165
iPC815	Yersinia pestis CO92	815	1552	1961		214092	21995956
iYO844	Bacillus subtilis subsp. subtilis str. 168	844	991	1250		224308	17573341
iJR904	Escherichia coli str. K-12 substr. MG1655	904	761	1075		511145	12952533
iMM904	Saccharomyces cerevisiae S288c	905	1226	1577		559292	19321003
iAF987	Geobacter metallireducens GS-15	987	1109	1285		269799	24762737
iSDY_1059	Shigella dysenteriae Sd197	1059	1890	2540	X	300267	24277855
iRC1080	Chlamydomonas	1086	1706	2191		3052	21811229
iSBO_1134	Shigella boydii Sb227	1134	1910	2592	X	300268	24277855
iSbBS512_1146	Shigella boydii CDC 3083-94	1147	1912	2592	X	344609	24277855
iSFxv_1172	Shigella flexneri 2002017	1169	1918	2639	X	591020	24277855
iSFV_1184	Shigella flexneri 5 str. 8401	1184	1917	2622	X	373384	24277855
iS_1188	Shigella flexneri 2a str. 2457T	1188	1914	2620	X	198215	24277855
iSF_1195	Shigella flexneri 2a str. 301	1195	1917	2631	X	198214	24277855
iYL1228	Klebsiella pneumoniae subsp. pneumoniae MGH 78578	1229	1658	2262		272620	21296962
iSSON_1240	Shigella sonnei Ss046	1240	1938	2694	X	300269	24277855
iAF1260	Escherichia coli str. K-12 substr. MG1655	1261	1668	2382		511145	17593909
iAF1260b	Escherichia coli str. K-12 substr. MG1655	1261	1668	2388		511145	19840862
iECH74115_1262	Escherichia coli O157:H7 str. EC4115	1262	1918	2695	X	444450	24277855
STM_v1_0	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	1271	1802	2545		99287	21244678
iECED1_1282	Escherichia coli ED1a	1279	1929	2707	X	585397	24277855
iG2583_1286	Escherichia coli O55:H7 str. CB9615	1283	1919	2705	X	701177	24277855
iE2348C_1286	Escherichia coli O127:H6 str. E2348/69	1284	1919	2704	X	574521	24277855
iECSP_1301	Escherichia coli O157:H7 str. TW14359	1299	1920	2713	X	544404	24277855
iECNA114_1301	Escherichia coli NA114	1301	1927	2719	X	1033813	24277855
iECs_1301	Escherichia coli O157:H7 str. Sakai	1301	1923	2721	X	386585	24277855
iLF82_1304	Escherichia coli LF82	1302	1940	2727	X	591946	24277855
iECOK1_1307	Escherichia coli IHE3034	1304	1943	2730	X	714962	24277855
iECS88_1305	Escherichia coli S88	1305	1944	2730	X	585035	24277855
ic_1306	Escherichia coli CFT073	1307	1938	2727	X	199310	24277855
iZ_1308	Escherichia coli O157:H7 str. EDL933	1308	1923	2722	X	155864	24277855
iECP_1309	Escherichia coli 536	1309	1943	2740	X	362663	24277855
iUTI89_1310	Escherichia coli UTI89	1310	1942	2726	X	364106	24277855
iNRG857_1313	Escherichia coli O83:H1 str. NRG 857C	1311	1945	2736	X	685038	24277855

Continued on next page

Table S1. Organism specific models and their properties

Model ID	Organism	Gene count	Metabolite count	Reaction count	55 *E.coli*	Taxonomy ID	PubMed ID
iAPECO1_1312	Escherichia coli APEC O1	1313	1944	2736	X	405955	24277855
iEC042_1314	Escherichia coli 042	1314	1926	2715	X	216592	24277855
iUMN146_1321	Escherichia coli UM146	1319	1944	2736	X	869729	24277855
iECABU_c1320	Escherichia coli ABU 83972	1320	1944	2732	X	655817	24277855
iEcHS_1320	Escherichia coli HS	1321	1965	2754	X	331112	24277855
iECIAI39_1322	Escherichia coli IAI39	1321	1957	2722	X	585057	24277855
iECO103_1326	Escherichia coli O103:H2 str. 12009	1327	1958	2759	X	585395	24277855
iECSF_1327	Escherichia coli SE15	1327	1951	2743	X	431946	24277855
iECDH10B_1368	Escherichia coli str. K-12 substr. DH10B	1327	1947	2743	X	316385	24277855
iBWG_1329	Escherichia coli BW2952	1328	1949	2742	X	595496	24277855
iECO111_1330	Escherichia coli O111:H- str. 11128	1328	1959	2761	X	585396	24277855
iECB_1328	Escherichia coli B str. REL606	1329	1953	2749	X	413997	24277855
iEC55989_1330	Escherichia coli 55989	1330	1953	2757	X	585055	24277855
iECUMN_1333	Escherichia coli UMN026	1332	1935	2741	X	585056	24277855
iECD_1391	Escherichia coli BL21(DE3)	1333	1945	2742	X	469008	24277855
iETEC_1333	Escherichia coli ETEC H10407	1333	1964	2757	X	316401	24277855
iB21_1397	Escherichia coli BL21(DE3)	1337	1945	2742	X	469008	24277855
iEcE24377_1341	Escherichia coli E24377A	1341	1974	2764	X	331111	24277855
iECIAI1_1343	Escherichia coli IAI1	1343	1970	2766	X	585034	24277855
iEcSMS35_1347	Escherichia coli SMS-3-5	1347	1949	2747	X	439855	24277855
iECSE_1348	Escherichia coli SE11	1348	1957	2769	X	409438	24277855
iUMNK88_1353	Escherichia coli UMNK88	1353	1971	2778	X	696406	24277855
iECBD_1354	Escherichia coli BL21-Gold(DE3)pLysS AG	1354	1954	2749	X	866768	24277855
iEKO11_1354	Escherichia coli KO11FL	1354	1974	2779	X	595495	24277855
iECO26_1355	Escherichia coli O26:H11 str. 11368	1355	1965	2781	X	573235	24277855
iY75_1357	Escherichia coli str. K-12 substr. W3110	1358	1953	2760	X	316407	24277855
iEcDH1_1363	Escherichia coli DH1	1363	1949	2751	X	536056	24277855
iJO1366	Escherichia coli str. K-12 substr. MG1655	1367	1805	2583	X	511145	21988831
iEcolC_1368	Escherichia coli ATCC 8739	1368	1971	2769	X	481805	24277855
iECW_1372	Escherichia coli W	1372	1975	2783	X	566546	24277855
iWFL_1372	Escherichia coli W	1372	1975	2783	X	566546	24277855
iECDH1ME8569_1439	Escherichia coli DH1	1439	1950	2756	X	536056	24277855

Figure S1. Networks sizes

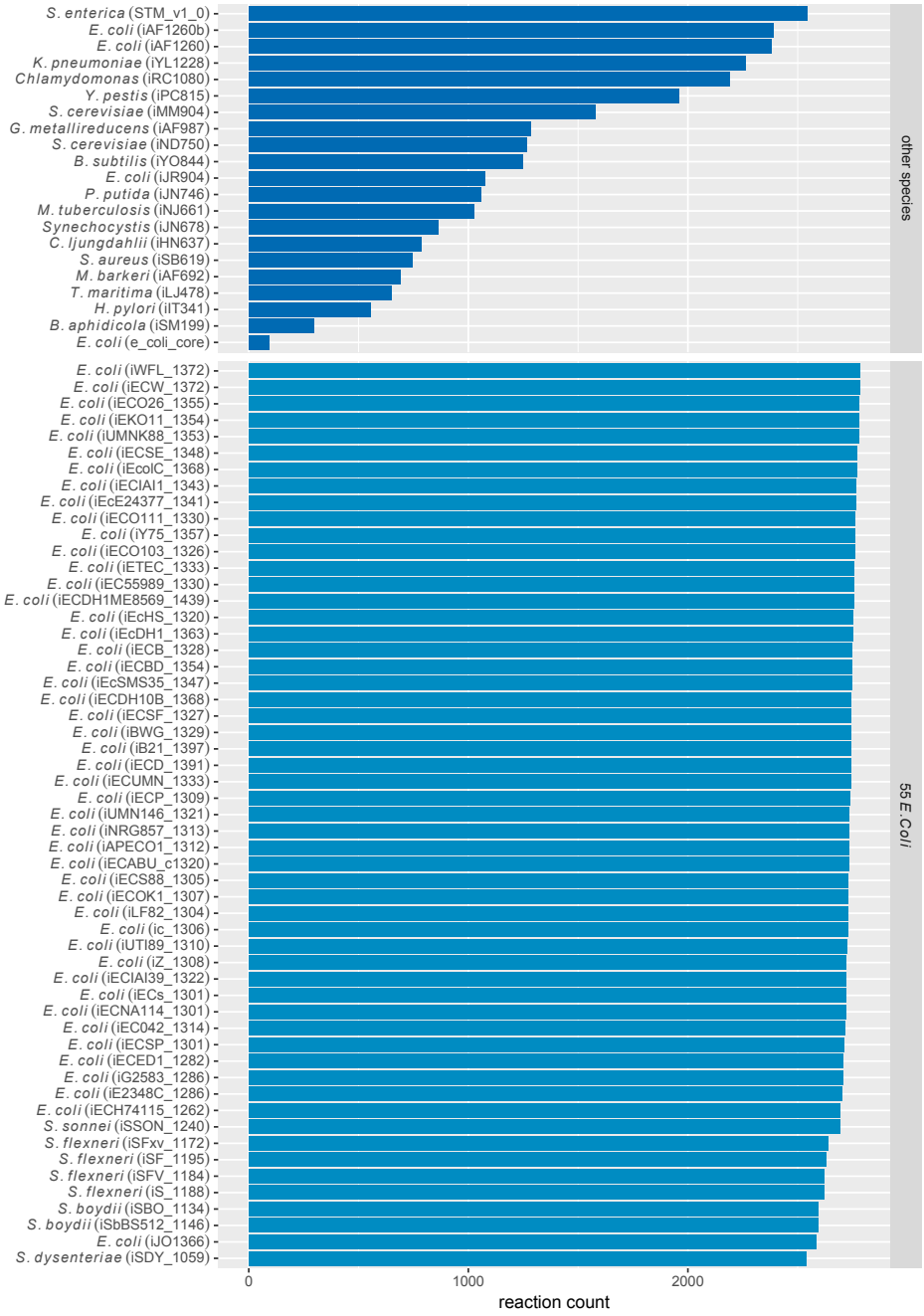


Figure S2a. Number of metabolites in compartments

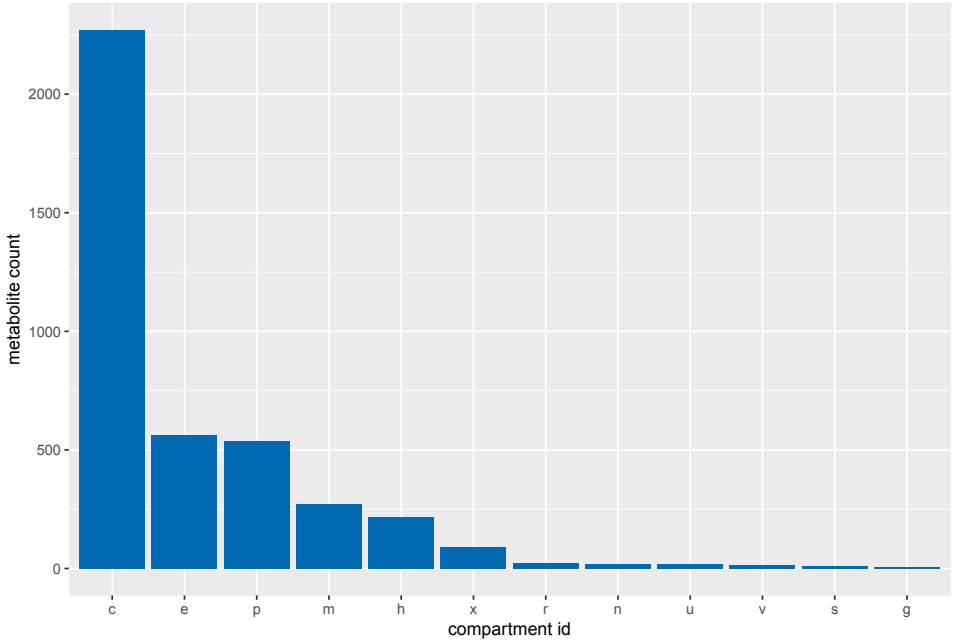


Figure S2b. published models per year

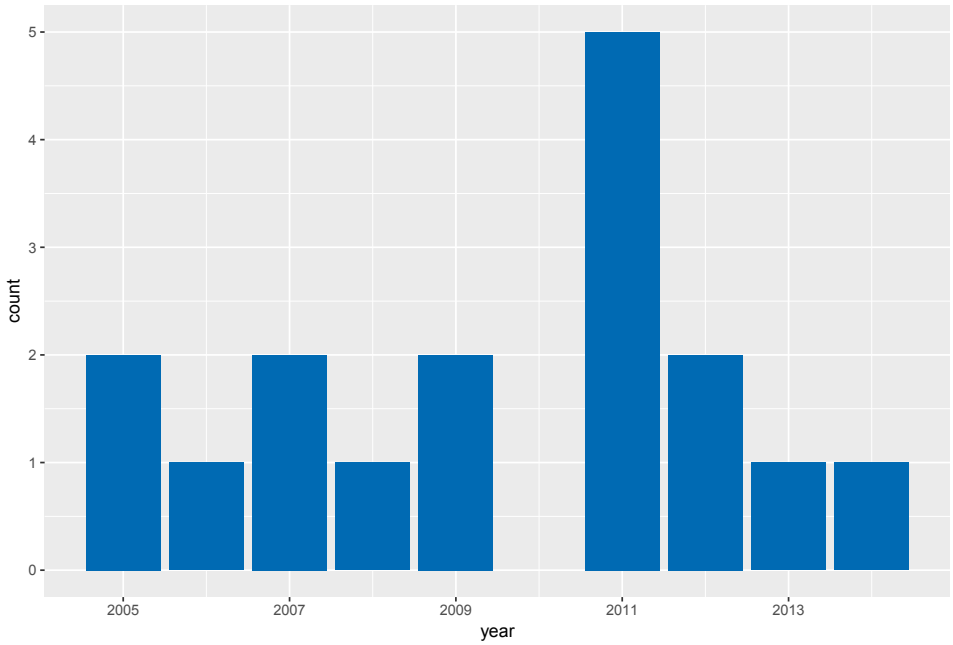


Figure S3. Growth for submodels and supermodel

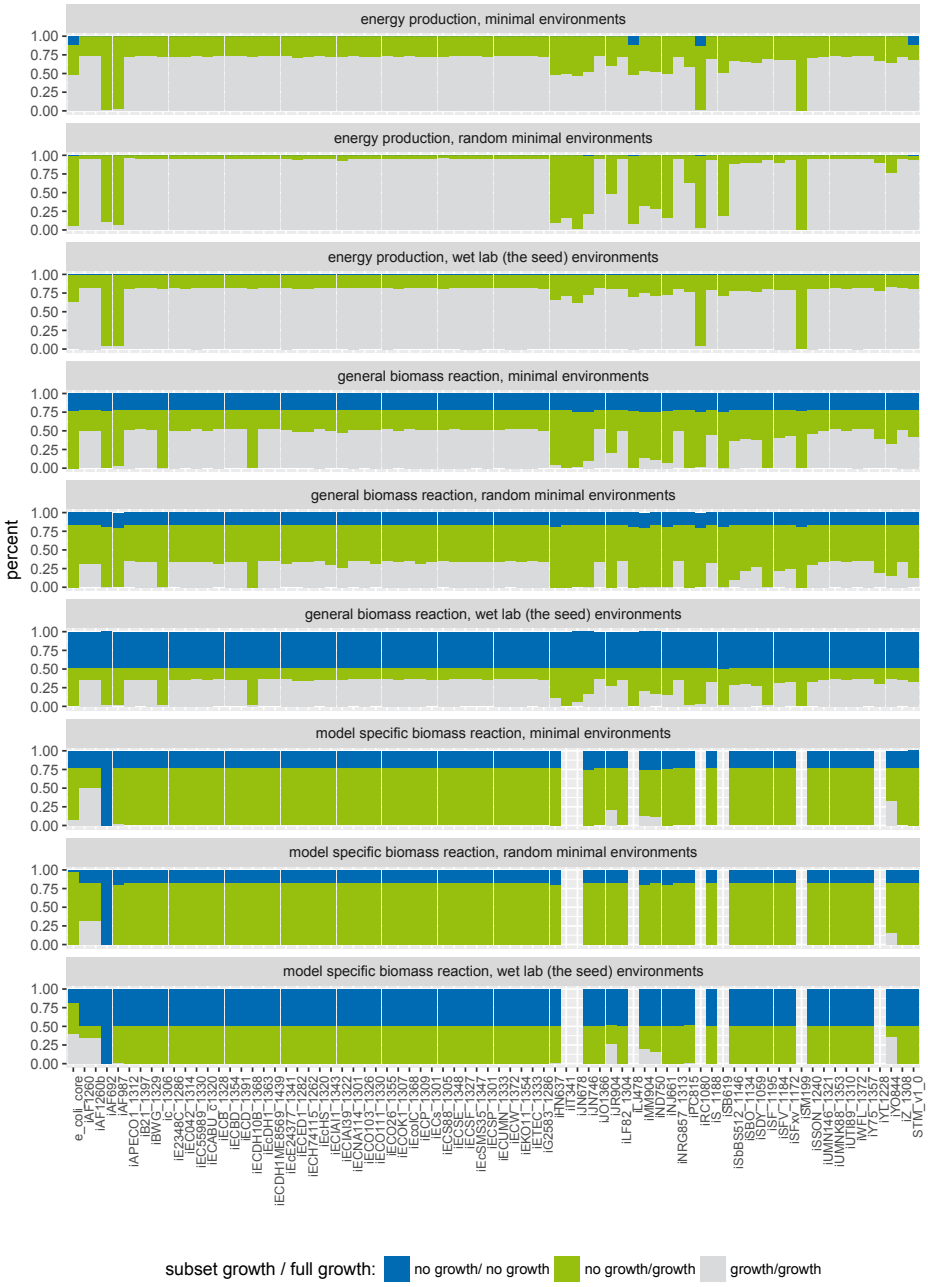


Figure S4. growth in wet lab media

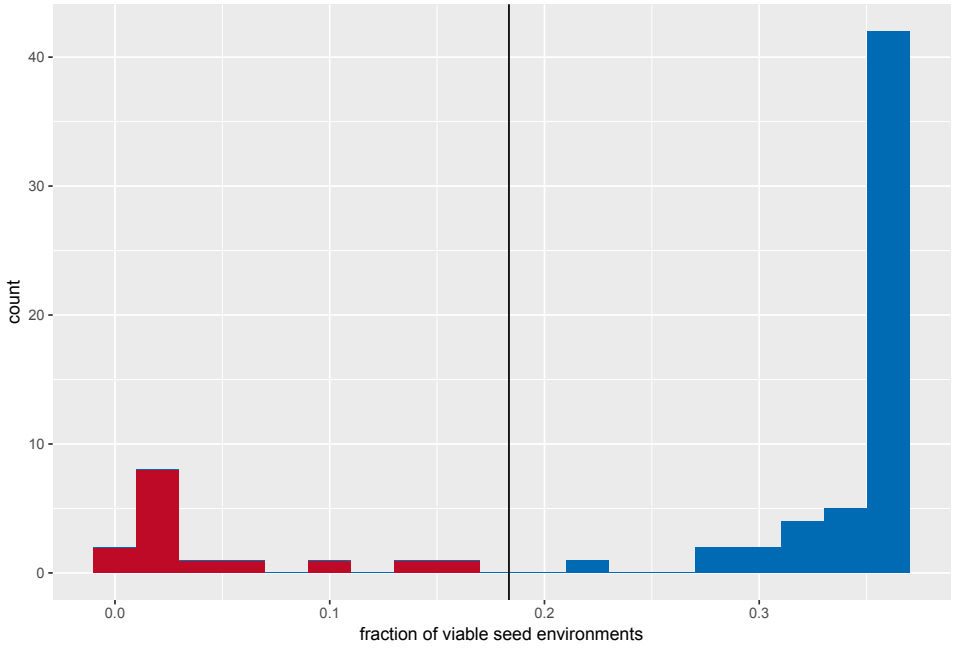


Figure S5. average # added reactions and model size for 55 *E.coli* only

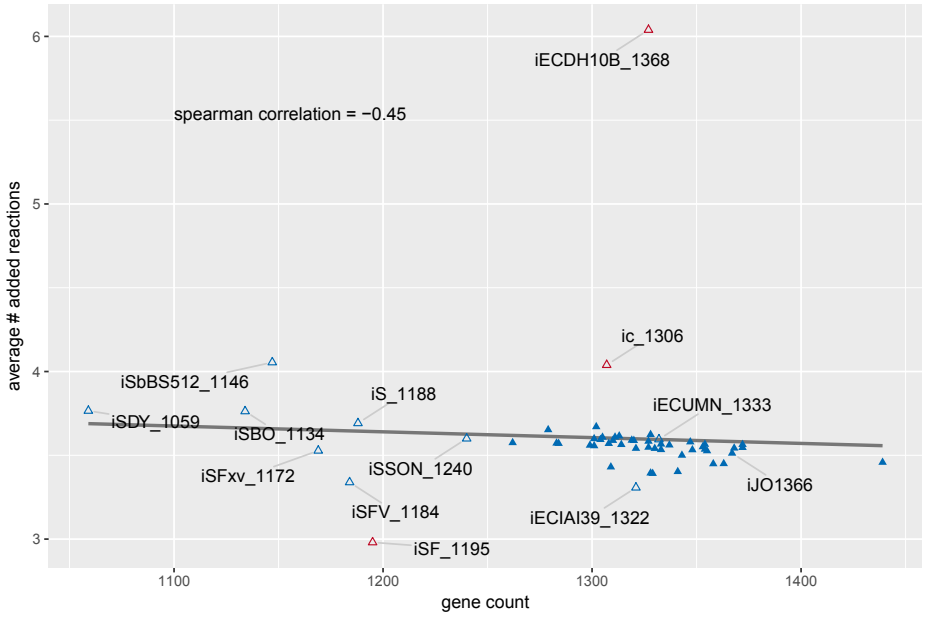


Figure S6. Distributions of added reactions per submodel

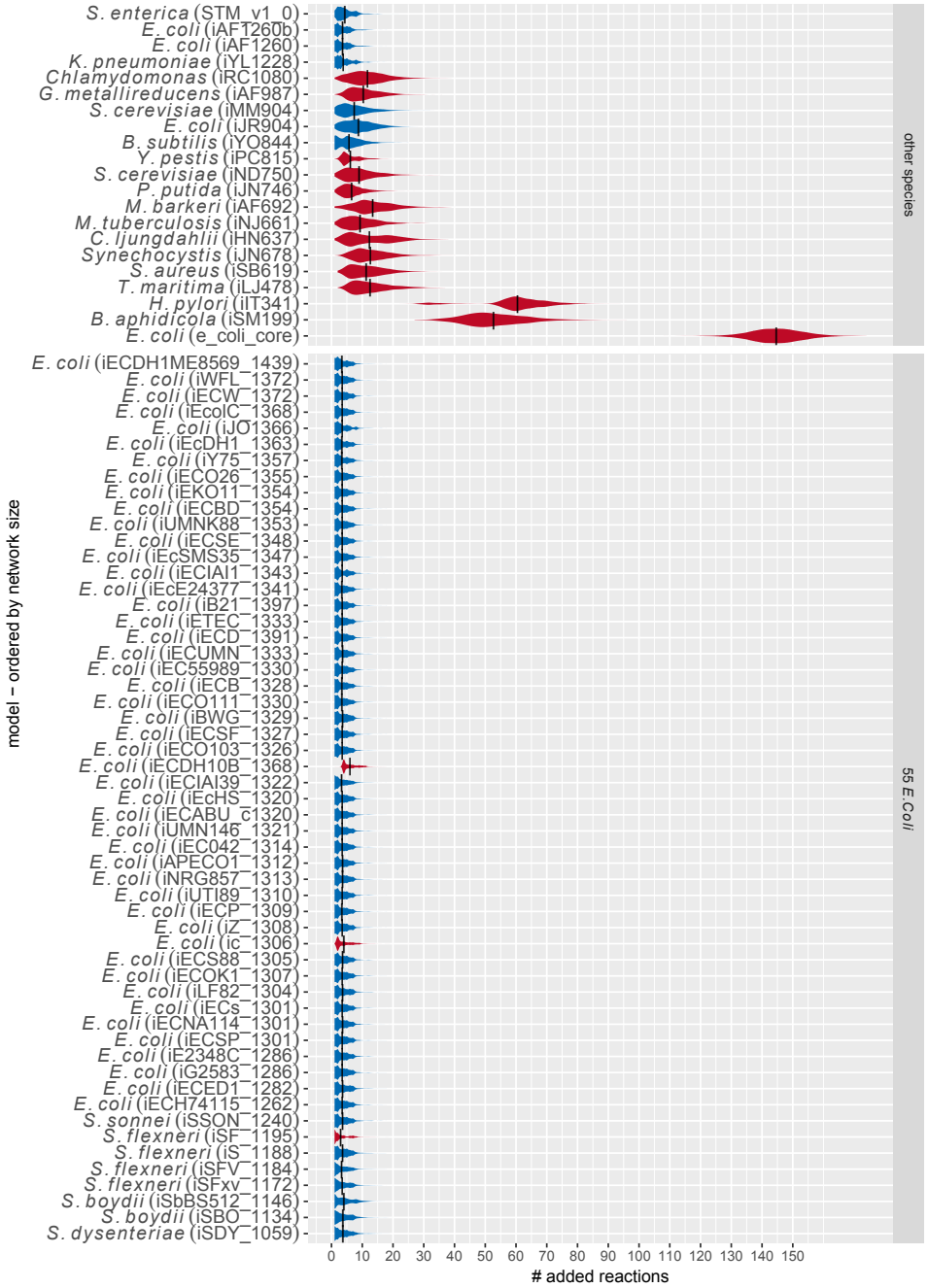


Figure S7. Distributions of the collateral adaptation index per submodel

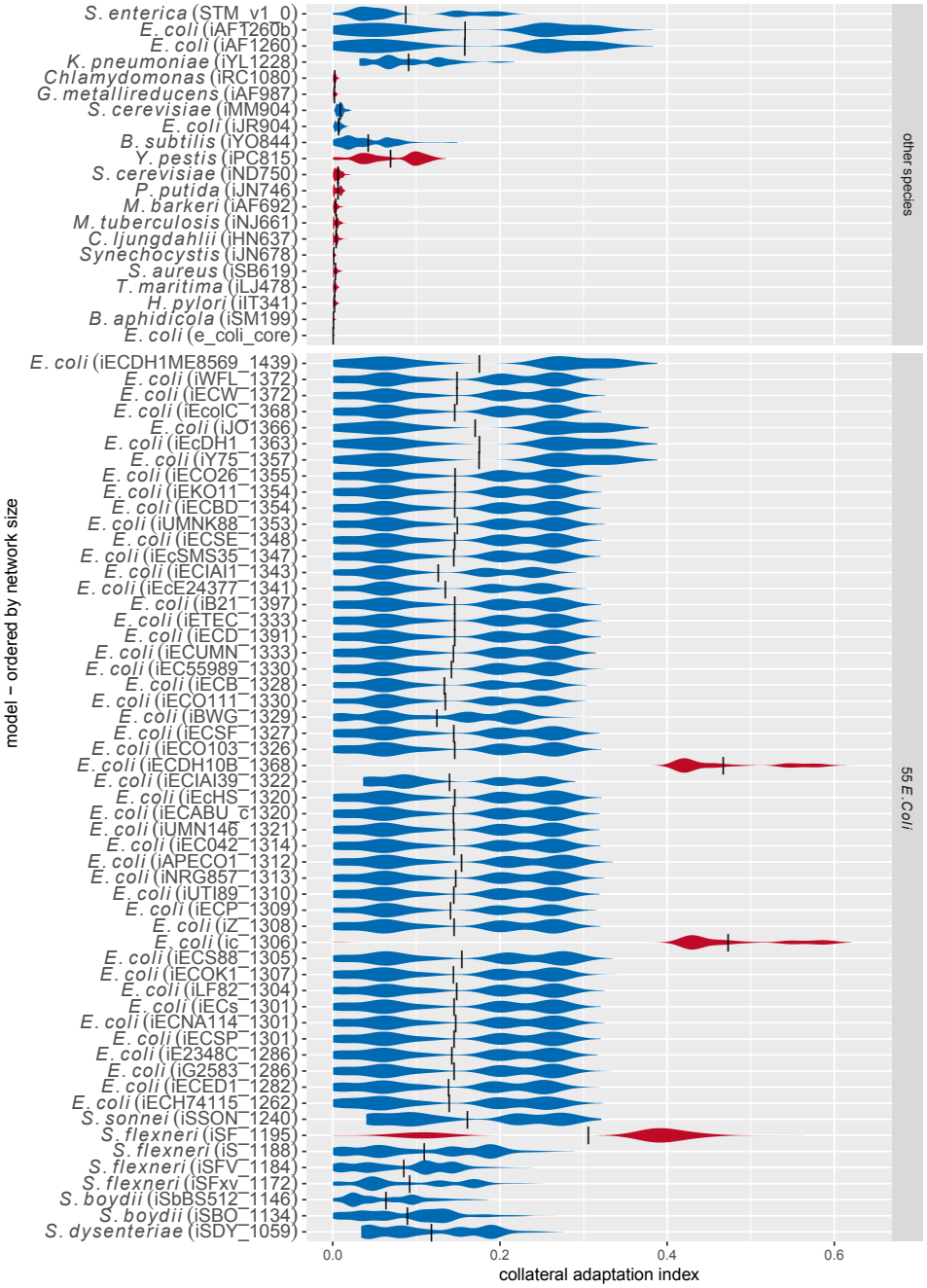


Figure S8. Distributions of the exaptation index per submodel

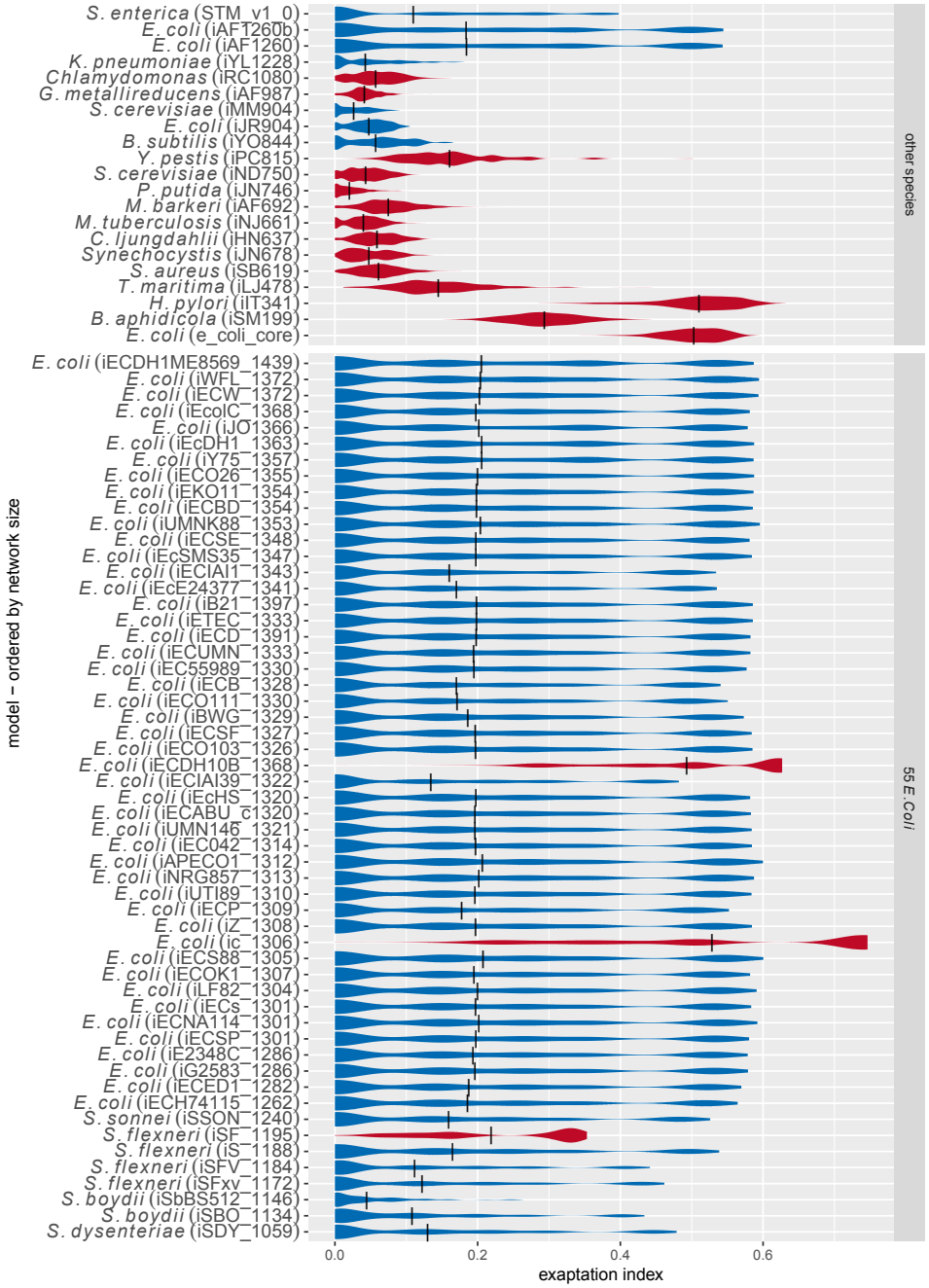


Figure S9a.

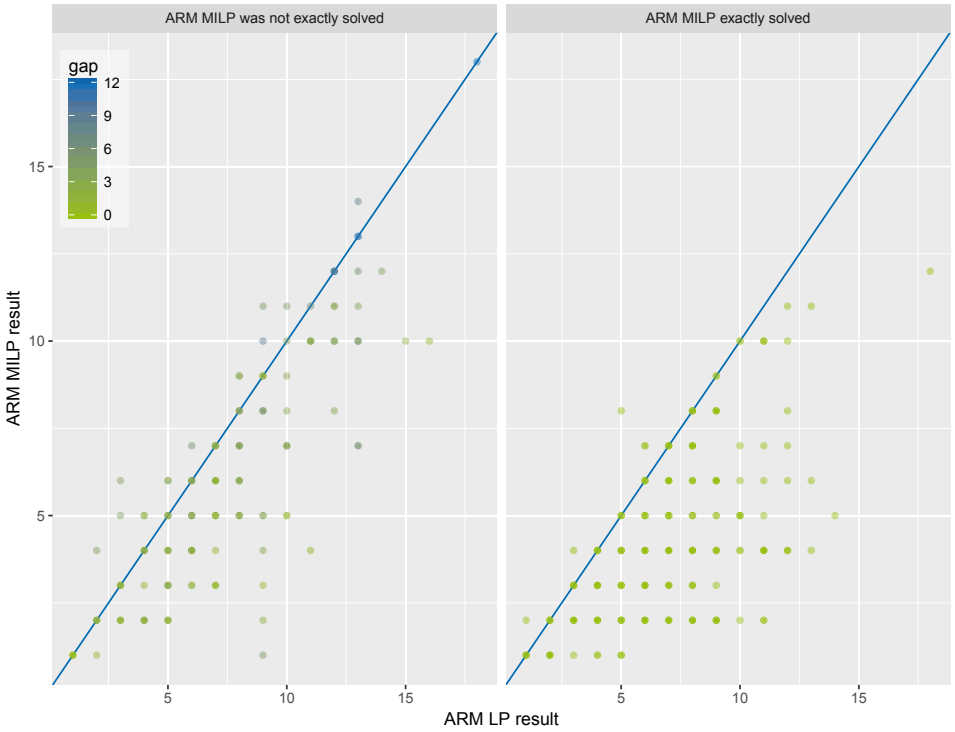
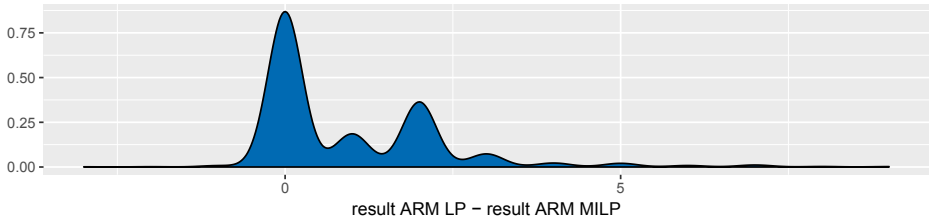


Figure S9b.



6.4.1 Contributions

C.J.F. constructed the pipeline that merges the previously collected metabolic network reconstructions into one pan-genome model. Together with F.L., C.J.F. developed the accelerated algorithm to solve the active reaction minimization problem. With this algorithm, C.J.F. performed the calculations to measure adaptation. C.J.F. analysed the data. The manuscript was drafted and figures were created by C.J.F. The manuscript was finalized by C.J.F. and M.J.L.

6.4.2 Outlook

In this study, the pan-genome-scale metabolic model was created to investigate the adaptability of 71 distinct organisms that vary widely in genome size and metabolic network structure. But this model opens a variety of novel possibilities. In future work, this model can be used to compare other metabolic properties of these organisms, *e.g.*, the ability to utilize certain nutrients and compare the necessary reactions. Further, it opens a new way of reconstructing GENREs. The model consists of a universal set of metabolic reactions. By choosing the right subset of reactions, the metabolism of arbitrary organisms can be modelled. A program like GlobalFit (Hartleb et al., 2016) is able to choose such a subset of reactions from the pan-genome-scale model. All reactions in the pan-genome-scale model get assigned weights according to their presence in the organism's genome. Given information about growth-promoting and non-growth-promoting environments GlobalFit finds a set of reactions that on one hand fits best to the organism of interest and on the other hand directly matches the growth data. The reconstruction process is then a top-down approach instead of the usual bottom-up approach.

6.4.3 References

Hartleb, D., Jarre, F., & Lercher, M. J. (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(8), e1005036.

7 Acknowledgements

Martin Lercher	for teaching, support, and giving me the opportunity to write this thesis.
William Martin	for support and especially for initiating the E-Norm graduate school.
Csaba Pál Balázs Papp, and Balázs Szappanos	for friendly collaboration and fruitful discussions.
Daniel Hartleb	for friendly collaboration.
Felix Lieder	for mathematical support.
David Heckmann	for support and discussion.
Gabriel Gelius-Dietrich	for teaching me, support, and discussion.
Rajen Brass Benjamin Braasch, and Marc-Andre Daxer	for support, discussion, and coffee breaks.
The workgroup	support, discussion, and punctual lunch breaks.
E-Norm members	for company in workshops, interesting seminars, and funny retreats.
Kathrin Kockel	for love and support.
My friends.	
My family.	

I am grateful for funding through the interdisciplinary Graduate School: Evolutionary Networks: Organisms, Reactions, Molecules (E-Norm).
Also I acknowledge the computational support through the Zentrum für Informations- und Medientechnologie (ZIM) at Heinrich-Heine-University Düsseldorf, especially the support through the HPC-Team.