Metabolic modeling to understand and redesign microbial systems

Ruben G. A. van Heck

Propositions

1. Metabolic models are incredible.
   (this thesis)

2. The simplification of metabolism into pathways is both convenient and misleading.
   (this thesis)

3. The inexplicable tendency to not share computational code hampers scientific development and results in a tremendous waste of resources.

4. Publishing is a form of dissemination, it should never be a goal in and of itself.

5. Science is too focused on 'showing that' rather than 'determining whether'.

6. Self-improvement comes through constant and thorough reflection on your performance; your results, whether good or bad, are not informative for self-improvement.

7. The meaning of 'highly educated' must not suffer from the desire to have more highly educated people.

Propositions belonging to the thesis entitled:

'Metabolic modeling to understand and redesign microbial systems'

Ruben G. A. van Heck

Wageningen, 6 July 2017

Metabolic modeling to understand and redesign microbial systems

Ruben G. A. van Heck

Thesis committee

Promotor
Prof. Dr Vitor A. P. Martins dos Santos
Professor of Systems and Synthetic Biology
Wageningen University & Research

Co-promotor
Dr María Suárez Diez
Assistant professor, Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr Jaap Molenaar, Wageningen University & Research
Prof. Dr Bas Teusink, VU Amsterdam
Prof. Dr Ralf Takors, University of Stuttgart, DE
Prof. Dr Markus Herrgard, Technical University of Denmark, Lyngby, DK

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Metabolic modeling to understand and redesign microbial systems

Ruben G. A. van Heck

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Thursday 6 July 2017
at 11 a.m. in the Aula.

Ruben G. A. van Heck
Metabolic modeling to understand and redesign microbial systems,
239 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with summary in English

ISBN: 978-94-6343-455-3
DOI: 10.18174/416473

"I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail."

Abraham Maslow

Contents

Chapter 1

Introduction: Metabolic modeling

It is the year 2009 and I am choosing my Bachelor programme for after the summer break. The BSc 'molecular life sciences' is the first I scratch off of my preselection and the search goes on. My search becomes focused after I stumble on a press article about the 2008 TU Delft student team of the international Genetically Engineered Machine (iGEM) competition. The article describes how micro-organisms can be modified to carry out all kinds of useful tasks for our society, for example, giving bright colors to dog feces so people do not accidentally step in. A silly example, to be sure. Nonetheless, the underlying concept of 'synthetic biology' — self-replicating micro-machines that carry out specific tasks — is fascinating, appears widely applicable, and ties in with societal goals such as reducing oil and pesticide usage, or performing bio-remediation. I decide to specifically look for BSc programmes that enable a future move into synthetic biology. After the summer, I start the BSc Molecular Life Sciences at Wageningen University "To explore the potential of nature to improve the quality of life".

In 2011, the second year of my BSc, I decided to join the Wageningen iGEM team; one of the best decisions I've ever made. This was a team full of motivated — and motivating — students who shared my interest in synthetic biology and together we created an educationally invaluable experience. I learned that the key aspect distinguishing synthetic biology from earlier biological disciplines is the 'engineering mindset'; the *ethos* to not use trial-and-error, but rational design. The rational redesign of biological system requires an extremely thorough and comprehensive understanding of the original system as there are so many relevant factors. This explains the 'marriage' between the fields of synthetic biology and systems biology. Systems biology focuses on combining all available sources of biological information in mathematical models in order to understand as much as possible of the biological system.

Although the mathematical modeling in systems biology is still mostly descriptive, the models could be used to redesign biological systems for specific purposes. Therefore, I enrolled in the MSc Bioinformatics with the Systems Biology specialization.

I start my PhD in 2013 on a project entitled "Modeling, refactoring and reprogramming the *Pseudomonas putida* chassis for biocatalysis *à la carte*". This project aligns perfectly with my developed skills and interests: The application of systems biology methods to study and model microbial metabolism with the ultimate goal of redesigning it for the production of compounds of biotechnological interest. Although I recognize that the identity of the produced compound is a crucial factor for any industrial exploitation, my own interests lie in the fundamental ability to rationally and functionally redesign biological systems using computational models. I strongly suspect that in the next few decades such efforts will prove a crucial centerpiece of biotechnology and synthetic biology applications for many organisms. One organism that is currently at the core of these ongoing developments is *P. putida*.

P. putida is a gram-negative soil bacterium renowned for its versatile metabolism [84, 195, 285, 424]. In particular, its ability to degrade aromatic compounds such as toluene and benzene [424] sparked initial interest in the potential of this 'oil-eating' organism. In fact, the first patent ever granted on a genetically modified organism regards the use of a modified *P. putida* strain for the bioremediation of oil spills [210]. *P. putida's* rise to fame did not end there, however, as it is currently regarded as a top synthetic biology chassis [264, 296], as well as industrial workhorse [264, 296, 338], and it stars in two Horizon2020 European Union funded projects (EmPowerPutida, project number 635536; and P4SB, project number 633962). The physiological and metabolic features underlying these interests are many: It grows fast, has a low nutrient requirement, promotes plant growth, thrives in a broad pH range, is tolerant to toxins and organic solvents, and has a high reductive power [72, 293, 296]. In addition, *P. putida* is genetically accessible, and a Generally Recognised As Safe (GRAS) status has been granted to strain KT2440. This strain has been the most thoroughly studied *P. putida* to date [296] and four genome-scale metabolic models extensively describe its metabolism [298, 311, 341, 395].

Metabolism is the interplay of all biochemical processes taking place within an organism. An ideal and comprehensive model of metabolism would accurately describe the dynamics of all metabolite concentrations and reaction rates over time. However, at any point in time, the rate of each reaction is determined through a complex interplay of many variables including (i) temperature, (ii) pH, (iii) concentration of substrates, products, and enzymes, (iv) enzyme properties, and (v) formation energies of substrates and products. In addition, metabolism is subject to several layers of regulation including (i) translational regulation, (ii) transcriptional regulation, (iii) active enzyme

degradation, and (iv) allosteric control. The aforementioned ideal model of metabolism would incorporate all these heterogeneous processes and types of data. Unfortunately, most of the required data is currently unavailable, and there is no mathematical framework in place to model all these processes simultaneously. Fortunately, a solution to both the lack of data and lack of suitable mathematical analysis methods is provided by Genome-Scale constraint-based Metabolic models (GSMs).

Genome-scale metabolic models

The use of GSMs relies on two assumptions that substantially simplify metabolism. The foremost assumption is that metabolism is in a steady state, which can experimentally be achieved in a chemostat [182], a turbidostat, and during the exponential growth phase [137]. Conceptually, this represents that over a long period of time there can not be a net increase in concentration of intracellular metabolites due to limited space, nor can there be a net decrease in concentration of a metabolite due to limited starting material. Mathematically, the steady state assumption removes the need for detailed multi-parameter differential equations for each reaction and replaces them by relatively simple parameterless constraints for each metabolite. The second major assumption is that evolution has selected for organisms that optimally use their metabolic capabilities for a particular metabolic objective. This optimality assumption implies that the regulatory mechanisms in a cell steer towards optimal metabolic activities. Therefore, there is no need to explicitly model regulatory mechanisms as they are implicitly modeled through determining optimal metabolic behavior.

These simplifications of metabolism have enabled a large range of succesfull GSM applications, for example: phenotype predictions [39, 284, 341], drug-target identification [39, 189, 337], metabolic engineering [341, 461], genome annotation [149, 312], analysis of omics data [69, 87, 473], and organism comparison [17, 27, 284, 311]. Hence, GSMs have been generated for hundreds of different organisms during the last few decades, and multiple GSMs have been generated for organisms of particular societal, industrial, or scientific interest (see figure 1.1).

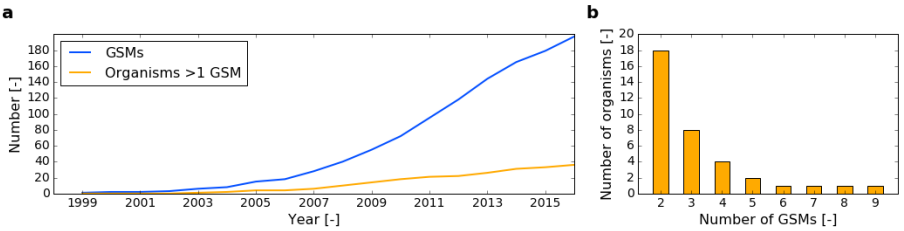


FIGURE 1.1: **Published manually curated GSMs. (a)** Cumulative numbers of published manually curated GSMs (blue) and species for which multiple manually curated GSMs were published (orange). **(b)** Histogram of species for which multiple manually curated GSMs have been published.

GSM generation starts with the construction of a draft model based on the genome of the organism of interest [418]. The functional genome annotation provides an overview of encoded enzymes and, thereby, an overview of biochemical reactions the organism can perform; typically in the form of EC numbers. EC numbers can be used to retrieve their corresponding reactions from a reaction database such as KEGG [200] or Metacyc [66]. These enzymatic reactions are further supplemented by spontaneous non-enzymatic reactions and artificial reactions that are required for the mathematical analysis. These artificial reactions include: (i) exchange reactions that represent changes to the growth medium, (ii) maintenance reactions that represent energy expenditure for unmodelled and unknown processes, and (iii) biomass reactions that represent growth as the production of a specific biomass composition. The biomass composition in GSMs is a rough breakdown of the dry weight biomass in individual metabolites [31, 129]. Although the biomass composition is ideally based on species-specific measurements, it is also often inferred from similar organisms [454]. This list of enzymatic, spontaneous and artificial reactions forms the draft GSM.

In principle, draft GSMs could be able to predict growth phenotypes: they contain an overview of the biochemical capabilities of the organism as well as the biomass reaction. Typically, however, not all biomass components can be produced in the draft GSM due to missing reactions, commonly referred to as 'gaps', even for well-characterised organisms [131, 234]. Reactions may be missing due to incorrect, non-specific or missing annotations [399], or even due to missing reactions in reaction databases [414]. In addition, GSM reactions are required to be either bidirectional (reversible) or unidirectional (irreversible), but the determination of reaction directionalities remains uncertain [138, 149]. In other words, a reaction could be present in the draft GSM but with an incorrect directionality, thereby preventing its proper functioning. These problems can - mostly - be addressed using gap-filling algorithms [234,

472]. Gap-filling algorithms add and modify a minimal number of reactions in a GSM in order to enable biomass production (or any other desired phenotype). However, any modification made by such an algorithm will require manual inspection. In fact, a draft GSM may require years of manual curation before being finished [418]. This mostly manual process is extensively covered in a 96-step protocol [418].

The aforementioned separate steps in GSM construction have recently been combined into fully automatic workflows [3, 176]. For example, the Model Seed [176] only requires a genome sequence as input to generate a draft GSM. This draft GSM can then also directly be subjected to gap-filling to obtain a GSM that can simulate biomass production in a predefined growth medium. Although automatic GSM generation provides an immense acceleration in GSM construction, manual curation of the generated GSM is still recommended — if not required — to warrant sufficient quality [3, 173, 176].

The quality of a new GSM is typically benchmarked *versus* previously published GSMs, especially if a GSM for the same organism already exists. However, the quality of a GSM is a somewhat ambiguous - perhaps even controversial - characteristic [309]. They should be comprehensive, should represent our biological knowledge, and should correctly predict growth phenotypes [320, 363]. Comprehensiveness is easily gauged via the sheer number of reactions, metabolites and genes included in a GSM; 'more is better' [320]. The representation of biological knowledge, on the other hand, is practically inassessible due to the requirement of substantial manual curation. The assessment of this criterion is thereby limited to, for example, mentioning the fraction of reactions with gene association [363], or the number of literature references that were used during GSM construction [10]. Growth phenotype predictions are qualitative predictions of whether or not the organism - or a mutant strain - can grow in a large number of defined media varying in carbon, nitrogen, phoshpor, and sulfur sources [235, 320]. This last criterion, growth phenotype predictions, is the most commonly used criterion when comparing the quality of different GSMs [309]. The three aformenetioned criteria are, however, not independent of one another. In fact, a recent assessment of updates of the *Sacchromyces cerevisiea* GSM demonstrated that typically the updates that increased comprehensiveness reduced accuracy of growth phenotype predictions and *vice versa* [320]. The evaluation of GSMs is further discussed in **chapter 8**.

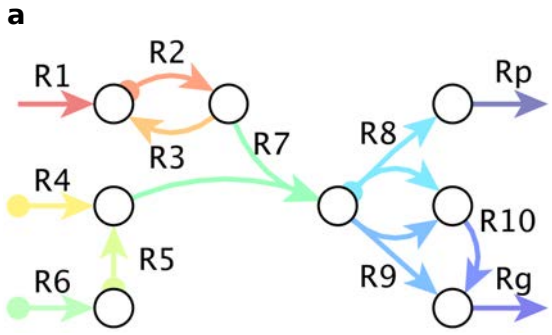
Analysis of metabolic models

The analysis of GSMs is effectively an analysis of an n-dimensional solution space; one dimension for each reaction. Each spot in this solution space corresponds to a flux distribution, which may be feasible or infeasible. A feasible flux distribution is one that does not violate any constraints, and all feasible flux distributions together make up the feasible solution space; a subspace of the total solution space. Figure 1.2 depicts a 12-reaction example model and how its 12-dimensional solution space is affected by the constraints on reaction bounds and steady state, as well as through the objective function.

Lower and upper reaction bounds are predefined for each reaction in a GSM. These bounds represent (i) thermodynamic feasibility of reaction reversibility, (ii) presence of a compound in the medium composition (exchange reactions only), and (iii) known flux limits such as measured substrate uptake rates.

The steady state requirement in GSMs dictate that there can be no net accumulation of intracellular metabolites. A reaction can thus only carry flux if its substrates can be produced by another reaction and its products can be consumed by another reaction. Thereby the possible flux values for reactions are dependent on the flux bounds of other reactions. Together, the flux bounds and the steady state requirement make up the feasible solution space. GSM analysis methods operate within this feasible solution space.

The objective function indicates the optimization goal of the metabolic network and, by extent, the hypothesised goal of the modelled organism. The most commonly used objective function is maximization of growth, which is modelled as the maximization of biomass production. The biomass objective function is conceptually based on evolution: The faster growing organism will outcompete the slower growing organisms, hence maximization of growth is a natural objective. Other objective functions are, for example, the maximization of product yield, or the minimization of ATP expenditure or overall flux [379]. Mathematically, the objective function is essentially a scoring function for each point in the feasible solution space. The points in the feasible solution space that share the highest score thereby form the optimal solution space. GSM analysis methods such as Flux Balance Analysis (FBA) [318], Flux Variability Analaysis (FVA) [261], and OptKnock [59] pinpoint a single spot in the optimal solution space corresponding to their optimization objective.



Reaction	Lower bound	Upper bound
R1	0	5
R2	-10	10
R3	0	10
R4	-10	10
R5	-10	10
R6	-10	10
R7	0	10
R8	-10	10
R9	0	10
R10	0	10
Rg	0	10
Rp	0	10

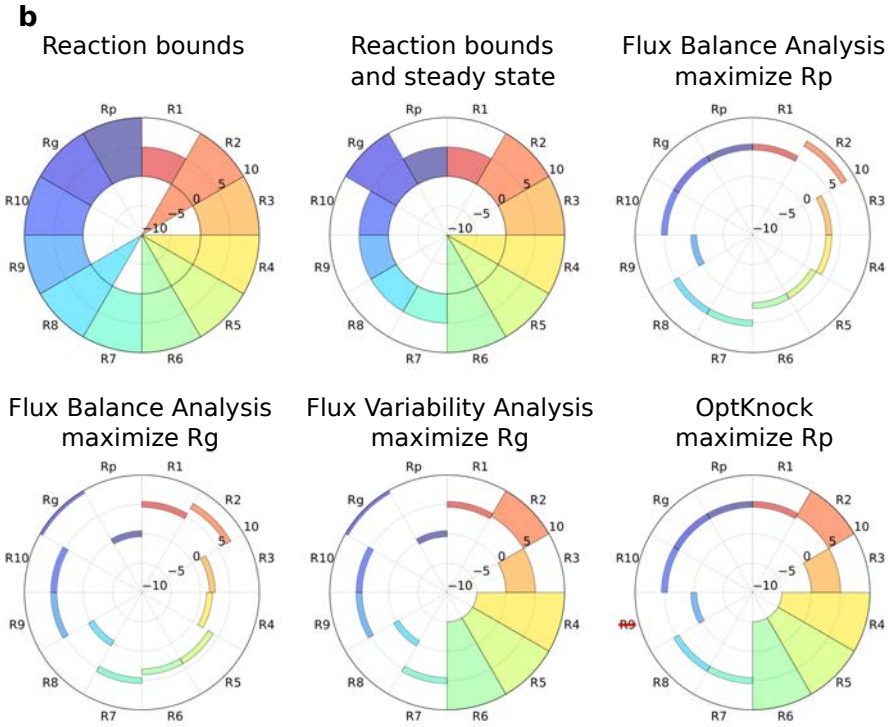


FIGURE 1.2: An example metabolic network and corresponding solution spaces. **(a)** The nodes and edges in the network graph represent metabolites and reactions. Edges with circular bases represent reversible reactions. The reactions denoted as Rg and Rp represent growth and the production of a compound of interest. **(b)** Each wedge represents possible flux values for the corresponding reaction depending on the employed constraints and optimization methods.

FBA, the most common GSM analysis method, determines how an organism can optimally use its metabolic capabilities to maximize biomass production [318]. Specifically, FBA is a linear programming problem that determines a single spot within the feasible solution space that corresponds to the maximally achievable biomass production. The FBA linear programming problem is defined as follows:

$$\begin{array}{ll} \text{Maximize} & \mathbf{cx} \\ \text{Subject to} & S\mathbf{x} = \mathbf{0} \\ & \mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub} \end{array}$$

The vector **c** (r x 1) is a Boolean vector indicating which reaction is the objective reaction. The vector **x** (r x 1) indicates a flux value for each reaction in the GSM. The matrix S (m x r) represents the relation between reactions and metabolites; it indicates how much of each metabolite is consumed or produced for each unit of flux through each reaction. The vector **0** (r x 1) is a null-vector that only contains zeroes. The vectors **lb** (r x 1) and **ub** (r x 1) indicate the lower and upper bounds of possible flux values for each reaction. FBA will identify a vector **x** that leads to the highest value of the scalar product **cx**, which equals the biomass production rate. The product $S\mathbf{x}$ is a vector that indicates the net consumption or accumulation of each metabolite, which is equaled to the null vector **0** so that a steady state is guaranteed. The last line, $\mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub}$ guarantees that the flux of each reaction is between its lower and upper bounds. FBA thus determines a single spot in the subspace of the feasible solution space that corresponds to the maximally attainable biomass production rate; a single spot in the optimal solution space.

FVA identifies the boundaries of the optimal solution space for each reaction in the GSM [261]. These boundaries are of interest as they describe the possible phenotypes of an optimally growing organism. For example, for microbes it is commonly assumed that the fastest growing microbe has an evolutionary advantage and will outcompete other microbes. Therefore, the microbe's metabolism should assume a flux distribution that is optimal. In turn, any flux distribution that is suboptimal is evolutionary disadvantageous and will ultimately not be used. Consequentially, if the optimal solution space requires the absence of flux for a reaction of biotechnological interest, evolutionary pressure will steer the organism away from the biotechnologically desired phenotype.

OptKnock, and other GSM-driven strain engineering methods, will therefore reshape the optimal solution space [59]. Specifically, OptKnock determines one or more reaction deletions that alter the optimal solution space such that a predefined growth rate is maintained, while the maximally possible flux through a reaction of interest is maximized. It is noteworthy that this new optimal solution space is a different subspace of the original feasible

solution space. OptKnock thus designs a knockout strain for which evolutionary pressure drives towards the production of a compound of interest.

Metabolic engineering

OptKnock is but one of many GSM-driven strain design algorithms [262] that vary both in approach to alter the solution space and in their optimization goal [262, 472]. For example, RobustKnock identifies reaction deletions that maximize the minimally possible flux rather than the maximally possible flux of a reaction of interest [411], whereas SimOptStrain maximizes the maximally possible flux of a reaction through both reaction deletions and reaction additions from a reference biochemical reaction database [213]. The addition of reactions from a biochemical reaction database also enables direct GSM-driven strain design for non-native compounds.

Biochemical reaction databases are limited to previously characterized reactions. Theoretically there are many more possible reactions, either uncharacterized or not naturally occuring, which could play important roles for metabolic engineering. Several algorithms aim to identify theoretically possible reactions and use these to design novel pathways [62, 63, 78, 169]. These algorithms infer reaction rules based on reactions in reaction databases, following the assumption that metabolites with similar biochemical properties can undergo similar biochemical conversions. A target metabolite is then selected, and the reaction rules are applied to predict theoretically possible reactions and corresponding metabolites. These metabolites are potential substrates for the production of the target metabolite in a single reaction step. These potential substrates are then again subject to the reaction rules to identify metabolites that are two reaction steps away from the original target. This process is repeated until a metabolite is identified that can already be produced by an organism of interest. This metabolite is then the starting point of a, potentially novel, biosynthesis pathway for the metabolite of interest. Subsequently, GSMs can be expanded with these pathways to evaluate their suitability in the context of the rest of metabolism. Specifically, GSM-driven strain design methods can evaluate whether the pathway is part of an optimal solution space, or if it can be made part of an optimal solution space via reaction deletions or additional reaction additions.

Outline of this thesis

The goals of this thesis are: To increase the understanding of microbial metabolism and to functionally redesign microbial systems using metabolic models. As a case study, I use *Pseudomonas putida*; a bacterium renowned for its versatile metabolism.

The foundation of metabolic modelling using Genome-Scale Metabolic models (GSMs) is a solid genome annotation. Genome annotations are arguably already outdated only days to weeks after they were completed. More than 10 years have passed, however, since the original *P. putida* genome annotation on which its GSMs were built. Therefore, we revisit and update the *P. putida* genome annotation in **chapter 2**. This update includes general structural and functional annotation updates, but also has a strong focus on elucidating the metabolism not priorly described in *P. putida* GSMs. The most comprehensive GSM to date - iJP962 [311] - was used to identify compounds that *P. putida* degrades but for which the degradation pathway remained unknown. These compounds formed the basis of an iterative cycle of targeted manual annotation and GSM expansion to identify suitable degradation pathways for these compounds in *P. putida*.

The *P. putida* GSM used in chapter 2 was only one of the four *P. putida* GSMs that were published at the time. These GSMs were made for the same organism, but by different people, with different expertise, using different parts of the scientific literature. These GSMs thus contain complementary knowledge that could be combined into a single, more comprehensive GSM. This situation of multiple complementary GSMs for one organism is not unique to *P. putida*; in fact, there are multiple GSMs for dozens of organisms (see figure 1.1). However, there is no suitable framework to compare and combine information from multiple GSMs. The study presented in **chapter 3** addresses this problem with the introduction of a computational tool that semi-automatically creates a single consensus GSM from multiple independently generated GSMs of the same species.

The species in the *Pseudomonas* genus exhibit vastly different lifestyles. There are plant growth-promoting species as well as plant pathogens, and soil-inhabiting obligate aerobes as well as lung-inhabiting facultative anaerobes. These vastly different lifestyles beckon the question what it is on a genetic level that defines a *Pseudomonas* species, and whether these defining traits are essential for growth and survival. In **chapter 4** we explore the link between the ubiquity of genetic elements in the *Pseudomonas* genus and their essentiality for growth. As experimental essentiality data is only available for few species in few experimental growth conditions, we employ the available *Pseudomonas* GSMs to predict the (un-)conditional essentiality of genes in an experimentally unattainable number of growth conditions. This data, together

with a high-throughput genetic comparison of 432 *Pseudomonas* strains, provides a clear relationship between gene ubiquity and essentiality in the *Pseudomonas* genus.

Among the distinct lifestyles within the *Pseudomonas* genus, the potential for anaerobic growth is both fundamentally interesting and industrially relevant. Therefore, several previous experimental studies have attempted to convey the ability for anaerobic growth found in other *Pseudomonas* species to *P. putida*, mostly focusing on anaerobic energy generation. So far, these approaches have resulted in increased anaerobic survival, but not in growth. **Chapter 5** describes an *in silico* approach to determine essential oxygen-dependent processes in *P. putida*, and subsequently design an anaerobically growing *P. putida*. This *in silico* approach uses GSMs to describe *P. putida* metabolism, as well as comparative genomics to pinpoint the genetic differences between obligate aerobic and facultative anaerobic *Pseudomonas* species. Together, these *in silico* methods have enabled the theoretical design of anaerobic *P. putida* strains.

Chapter 6 describes the *in silico* design of species-specific synthetic CO_2 fixation pathways. CO_2 fixation is a fundamental component of the biobased economy, but natural CO_2 fixation pathways are rather inefficient. This inefficiency has sparked several prior studies where synthetic CO_2 fixation pathways were designed that are ATP-efficient and have favorable thermodynamics and kinetics. These studies have, however, mostly designed pathways in isolation rather than in the context of the metabolism of any particular microorganism. This lack of contextualization typically introduces a requirement for many non-native enzymes. Therefore, we developed CO2FIX, an *in silico* method that uses species-specific GSMs to design tailored CO_2 fixation pathways that require few non-native reactions. The application of CO2FIX to eight different organisms has resulted in several novel CO_2 fixation pathways that require less modification of native metabolism than previously described synthetic pathways, but are similar in terms of ATP-efficiency and thermodynamic and kinetic favorability.

The previous chapters demonstrate that GSMs for relatively well-understood microbes such as *P. putida* can both further increase the understanding of microbial metabolism and can be used to functionally redesign microbial systems. In **chapters 7 and 8** we explore how GSMs can contribute to these same goals for the much more complex microbial systems that are algae (**chapter 7**) and gut microbial communities (**chapter 8**). Although these systems are starkly different, they face similar issues in terms of cultivation, annotation, lack of species-specific data, and general difficulties in multicompartment and multispecies modeling. These chapters review previous work to address these issues and provide perspectives on how GSMs will continue to contribute to the study of these complex microbial systems.

In **chapter 9** I will discuss how the research presented in this thesis contributes to the objectives previously outlined herein, as well as to the current challenges and opportunities surrounding the use of GSMs.

Chapter 2

The foundation: *Pseudomonas putida* reannotation

Adapted from:

Eugeni Belda*, **Ruben G. A. van Heck***, Maria Jose Lopez-Sanchez, Stephane Cruveiller, Valerie Barbe, Claire Fraser, Hans-Peter Klenk, Jorn Petersen, Anne Morgat, Pablo I. Nikel, David Vallenet, Zoe Rouy, Agnieszka Sekowska, Vitor A. P. Martins dos Santos, Victor de Lorenzo, Antoine Danchin, and Claudine Medigue. "The revisited genome of *Pseudomonas putida* KT2440 enlightens its value as a robust metabolic chassis". In: Environmental Microbiology 18(10) 2016.

*Equal contributions

Abstract

By the time the complete genome sequence of the soil bacterium *Pseudomonas putida* KT2440 was published in 2002 [290] this bacterium was considered a potential agent for environmental bioremediation of industrial waste and a good colonizer of the rhizosphere. However, neither the annotation tools available at that time nor the scarcely available omics data –let alone metabolic modeling and other nowadays common systems biology approaches- allowed us to anticipate the astonishing capacities that are encoded in the genetic complement of this unique microorganism. In this work we have adopted a suite of state-of-the-art genomic analysis tools to revisit the functional and metabolic information encoded in the chromosomal sequence of strain KT2440. We identified 242 new protein-coding genes and reannotated the functions of 1,548 genes, which are linked to almost 4,900 PubMed references. We also predicted catabolic pathways for 92 compounds (carbon, nitrogen, and phosphorus sources) that could not be accommodated by the previously constructed metabolic models. The resulting examination not only accounts for some of the known stress tolerance traits known in *P. putida* but also recognizes the capacity of this bacterium to perform difficult redox reactions, thereby multiplying its value as a platform microorganism for industrial biotechnology.

Introduction

Pseudomonas putida is a soil bacterium generally recognized as safe (GRAS). Belonging to a somewhat fuzzy clade of the *Pseudomonadales* [327], it has been used for decades as a model environmental organism with activity against aromatic pollutants. In 2002, in a successful transatlantic collaboration, scientists at The Institute for Genomic Research (USA) and at four research centers in Germany deciphered and analyzed the genome of strain KT2440 [290]. This strain, which can be used to dispose of organic pollutants in the soil, promotes plant growth and fights plant diseases [359]. Regenhardt *et al.* highlighted the complex and versatile metabolism that gives *P. putida* an important role not only in academic research on soil bacteria but also as an agent for environmental cleanup and other biotechnological uses [359]. Yet, the genome analysis tools available at that time were able to extract only a small portion of the wealth of biological activities encoded in the chromosome of this bacterium.

In this work we set out to revisit the metabolic and physiological setup of this organism by re-analysing the content of its genome using several approaches. We first re-sequenced the *P. putida* KT2440 wild-type strain in parallel with that of a streamlined derivative as a control for possible evolution in laboratory settings [245] and compared it to the original published sequence [290]. Combined with transcriptomics data analysis [145, 214], a complete structural reannotation of the KT2440 genome sequence led us to eliminate original erroneously predicted protein-coding genes, to correct disrupted genes and to identify potential new genes, some of which encode enzymatic activities. In a second step, we functionally reannotated these genes based on recent progress in our knowledge of metabolic pathways [386, 453]. Thirdly, we used this reannotation to reconcile *in silico* predictions from Genome-Scale constraint-based Metabolic models (GSMs) [298, 311, 341, 395] with metabolic phenotype data obtained with BIOLOG plates[43]. The updated annotation was then used to extend the GSM iJP962 [311] with newly curated Gene-Protein-Reaction associations. Finally, this extended GSM was evaluated for its ability to correctly predict positive/negative phenotypes of wild-type and mutant strains.

During the curation process, we surveyed metabolic pathways involved in coping with stressful environments and explored in some details the general context of aromatic compounds degradation. In biochemical terms, the synthesis of aromatic molecules is costly as it requires much energy and reducing power [4]. In genetic terms, the synthesis and degradation of aromatic compounds are costly too, because of the fairly large number of genes involved in these processes. In physico-chemical terms, the degradation of aromatics is problematic due to the fact that, because of their electronic set up, they tend to cross membranes when uncharged, often disrupting the lipid

bilayer of the membrane and leaking in and out of the compartments where they should be confined [373]. Furthermore, because of this property, they frequently behave as proton-carriers that shunt the vectorial proton transport that would be used to build up ATP otherwise (*i.e.,* chemical uncoupling). As a consequence, catabolic processes must be compartmentalized in a way that matches proton availability with the propensity of a protonated molecule to pass through the membrane [99, 205]. This requires an efficient management of transport processes and control of the electrochemical potential of the cell as well as osmolarity. For this reason, we explored the metabolic capacity of *P. putida*, as indicated by the presence of relevant genes, in the context of control of osmolarity, control of proton availability and aromatic compounds degradation. Taken together, the novelties and metabolic updates presented in this work should contribute to the implementation of biocatalysis strategies using *P. putida* as a chassis for Synthetic Biology constructs.

The updated *P. putida* KT2440 genome sequence is deposited at the International Nucleotide Sequence Data Collaboration (identical accession number: AE015451, version 2). The reannotated data can also be explored and downloaded using the MicroScope platform. The curated genome-scale metabolic network is available at the MicroCyc repository and can be downloaded using the "Download Data" functionality of the "Search/Export" menu of the MicroScope platform. Finally, the updated metabolic model is available in the Supporting Information (SBML file formats).

Results and Discussion

New features of the genome of strain KT2440

P. putida genome sequence and its structural reannotation

The revised *P. putida* genome has 10 additional nucleotides compared to the earlier version (6,181,873 bp instead of 6,181,863 bp). We found 140 uncovered regions in the re-sequenced genome (the largest being 5-kb long), encompassing regions annotated as rRNAs, tRNAs, transposons, group II intron-encoding sequences. Since all those genetic elements are repeats, this clearly indicated that the reference genome has been fully covered. The *P. putida* genome displays a GC content of 61.5%. The consensus sequence correction (see Experimental Procedures) shows that the original sequence was of outstanding quality [290]. Indeed, among the 83 detected variations, 46 accounted for Single Nucleotide Polymorphisms (SNPs), 23 for short insertions and 14 for small deletions. It is known that strains kept in laboratories tend to evolve [26]. In order to substantiate the validity of our re-sequencing of the

genome we compared the regions of variation with the sequence of a streamlined mutant [245]. 96% of the variations were present in both sequences, showing that they were present at an early stage. A significant part of the events (54) were found to affect 20 CoDing Sequences (CDSs; see table S1). In most cases, insertion/deletion (InDel) events restored the reading frame (*i.e.,* either the new CDS is longer than the published one or two CDSs are fused in one gene, see table S1). Only *PP0253* (encoding phosphoenolpyruvate carboxykinase) and *PP5662* (encoding two fragments of a conserved gene of unknown function) remain pseudogenes (see below). Curiously, *PP5662* and *PP4302* (encoding an urea transporter) gather most of the detected SNPs (73.3%, either transitions or transversions). The reannotated genome sequence (see Experimental Procedures) comprises 5,592 CDSs plus 56 fragments of CDS (versus 5,350 CDS stored in the last release of the NCBI GenBank file, NC_002947, or in the Pseudomonas.com database [451]), 22 rRNA genes, and 75 tRNA genes (versus 74). The non-coding regions account for 11.5% of the *P. putida* genome and contain 7.5% of repeated sequences. Only nine non-coding regions of more than 1 kb have been identified (table S2). Among the annotated CDSs (complete genes and pseudogenes), we identified (i) common gene annotations between the original data and the AMIGene predictions: 5,301 genes (94.8%), the original start codon positions of which were automatically kept, (ii) gene annotation unique to the original GenBank file: 116 genes, and (iii) gene annotation unique to the AMIGene prediction: 607 potential new CDSs. Following the manual curation process described in the Experimental Procedures section, 311 CDSs unique to the present version and 36 CDSs unique to the original annotation were kept. Moreover, 102 original CDSs were considered false positive predictions and removed from the final set of genes (table S3). All of them would encode proteins of unknown function and 38 (37.2%) were found at a position where a new gene has been annotated, generally on the complementary strand [see [272] for a similar rationale used for annotation of *Helicobacter pylori* genes] (tables S3 and S4). As shown in Supplementary table S4, the validity of most of the 311 newly annotated genes is supported by transcription expression profiling [145, 214] and/or by sequence similarity with authentic genes: for example, *PP5706* encodes a protein involved in the Sec translocation complex (SecG subunit), and *PP5602* encodes the αsubunit of the quinohaemoprotein amine dehydrogenase (the *peaA* gene within the *peaACB* operon) which is known to be involved in the conversion of 2-phenylethylamine and 2-phenylethanol into phenylacetic acid in *P. putida* U [14]. Indeed, 118 newly annotated genes among the 143 novel CDSs listed in table S2 of the publication by Frank *et al.* [145], show up in table S4 (the 25 missing ones correspond to predicted genes that were considered as false positives by our curation process). Forty-five new genes (14.5%) were assigned a gene product type and a biological process (table S4) whereas the

remaining genes (266) correspond to functions that remain to be identified.

Remaining pseudogenes

The re-sequencing process followed by expert curation of gene fragments and fusion/fission events using the MicroScope platform [432], identified a total of 71 CDSs as partial genes (14), pseudogenes (54 fragments of CDSs corresponding to 28 pseudogenes, and 1 CDS, *PP3752*, which contains an internal stop codon) and one programmed frameshift (2 CDSs corresponding to the peptide chain release factor 2 gene, *prfB*). Partial genes were essentially grouped into classes of genes either encoding proteins containing Rhs repeat domains, transcriptional regulators (LysR family) or transposases (table S5). Two of the 27 pseudogenes are of particular interest:

- The gene *PP0253* is split into two fragments that have 100% amino acid identity with fragments of the *pckA* gene encoding phosphoenolpyruvate carboxykinase (ATP dependent) in *P. putida* F1 (UniProt entry A5VX32). This enzyme is involved in gluconeogenesis, where it catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP). The present UniProt functional annotation is supported by sequence similarity using the UniRule annotation procedure [88]. Indeed, similarity with an experimentally validated phosphoenolpyruvate carboxykinase is found with the *Staphylococcus aureus* PckA protein (Q2G1W2, 45.4% amino-acid identity) [382]. The underlying reason for this loss of function in strain KT2440 is unknown, but we note that this enzyme is a key enzyme required for gluconeogenesis, under conditions where *P. putida* strains display a tight regulation of the balance between fluxes going from glucose to pyruvate and from succinate to pyruvate [236]. In *E. coli* O157:H7, PckA is important for maintaining the pathogenic bacteria in competition with the bulk of the microbiome [38]; inactivation of the gene may contribute to the GRAS phenotype of strain KT2440. Additionally, the enzyme is allosterically regulated by Ca2+ in other γ-proteobacteria [402], and this feature might point at a particular role of the inactivation of this gene in the *P. putida* KT2440 niche.

- The two fragments of gene *PP1919* encode a protein similar to *E. coli* K-12 thymidylate kinase (Tmk protein; >50% identity), a key enzyme for DNA synthesis. This protein catalyzes the phosphorylation of deoxythymidine monophosphate (dTMP) to deoxythymidine diphosphate (dTDP) in the presence of ATP and Mg2+. Tmk is essential for DNA synthesis and cell growth in *E. coli* [361] and it would be expected to be essential in *P. putida* as well. Interestingly, in strain KT2440, but not in

other sequenced *P. putida* strains, the *tmk* gene has been disrupted by the integration of a large genomic island of about 65 kb (the 3'-end of the first part of *tmk* is found at position 2,162,696 bp, while the 5'-end of the second part is found at position 2,227,487 bp). This region is obviously of phage origin (it contains genes for phage integrases, a transcriptional regulator of the Cro/cI family as well as site-specific recombinases), and harbors several clusters of metabolic genes (monooxygenases, dehydrogenases, etc.) together with a cluster of genes involved in arsenic resistance (*PP1927-PP1930*). Remarkably, *PP1964*, the prophage gene located next to the truncated *tmk* gene, is likely to encode a deoxyribonucleotide monophosphate kinase [278], that could substitute for the missing essential *tmk* gene. Alternatively, the two halves of the *tmk* gene could be expressed separately and the resulting polypeptides reconstruct the enzyme activity through protein trans-complementation, a possibility currently under investigation.

Functional reannotation of protein-coding genes

The outcome of the automatic functional annotation procedure was followed by manual curation of *P. putida* genes previously recorded as encoding unknown functions, while showing significant similarity with one of the protein and domain resources used in the platform (see Experimental Procedures). Among those, 197 CDSs were reviewed (table S6). Most of these proteins were labeled as (putative) enzymes (56%), (putative) transporters (20%) or (putative) regulators (9%). We further annotated 61 genes encoding proteins highly similar to proteins with functions experimentally demonstrated either in *Pseudomonas* species/genus or in other organisms. This is the case for genes involved in the catabolism of carnitine (*PP0301-PP0305*; [28, 439]), in phenylethylamine degradation (*PP3459* and *PP3460*; [14]), in gallate degradation (*PP2513*, *PP2514* and *PP2515*; [300]), and in urate degradation (*PP4287*; [351]). In order to provide accurate annotations, the global curation process was directed by the results of the growth phenotype data obtained in this work as well as extracted from experimentally based literature (see next section).

TABLE 2.1: Summary of the main *P. putida* KT2440 features annotation update in comparison with the original one.

		New annotations	Original annotations
CDS	Total number	5592	5350
	Unknown functions / hypothetical proteins	1151*[1]	1505
	Pseudogenes	28*[2]	9*[3]
	Partial genes	14	61*[4]
	Additional genes	311	
	False positive genes in original annotations		102
rRNA genes	Total number	22	22
tRNA genes	Total number	75	74
EC numbers	CDS associated with an EC number	1250	463
	Total unique EC numbers	902	360
	Complete EC numbers	811	360
	Partial EC numbers	91	0
GPR associations	Number of CDSs associated to reactions	1485	0
	Number of reactions	1898*[5]	0
	Total number of GPR associations	3185	0
PMID annotations	Genes with associated PMID references	1371	18
	Number of different PMID references	4837	1

*[1] 1040 conserved proteins of unknown function + 111 proteins of unknown function. *[2] 28 pseudogenes made of 54 fragments of CDSs corresponding to 27 pseudogenes, and 1 CDS, *PP3752*, which contains an internal stop codon. *[3] 9 genes without/product annotation; note = "This region contains a pseudogene, one or more premature stops, and is not the result of a sequencing artifact"; following the sequencing and the manual curation processes these 9 pseudogenes have been reannotated as functional. *[4] 61 genes without/product annotation; note = "This region contains an authentic frame shift and is not the result of a sequencing artifact." The sequence of 10 of these partial genes has been corrected after the re-sequencing process. *[5] 1406 MetaCyc reactions [66] 492 Rhea reactions [286].

Overall, the function of 1,548 genes has been manually reannotated and linked to updated literature references (4,837 PubMed references in the current annotation release). To provide a comprehensive reconstruction of the global metabolic map of *P. putida*, the utmost care was taken in the curation of associations between genes encoding enzymes and the biochemical reactions they catalyze. A total of 1,485 CDSs has been associated to 1,898 chemical reactions (1,406 reactions from MetaCyc [66] and 492 from Rhea [286]) comprising a total of 3,185 gene-reaction associations. In these associations, the role of 229 genes, displaying a high degree of similarity with their counterparts, was automatically annotated via transfer of the related *E. coli* K-12 reactions (see Experimental Procedures). In the current update of the *P. putida* KT2440 genome annotation, about 21% of the protein-coding genes still remain of unknown function. A summary of the main *P. putida* KT2440 genome annotation

updates in comparison with the original annotation can be found in table 2.1.

An updated view of strain KT2440 metabolic capabilities through genome-scale modeling and phenotyping data

The updated genome annotation and corresponding functions were subsequently reviewed by computer simulations, assessing their contribution to the GSM iJP962 [311], which progresses towards a comprehensive model of the current knowledge of *P. putida* metabolism. First, we pinpointed knowledge gaps in the original GSM by comparing its *in silico* growth predictions to the output of BIOLOG experiments on carbon, nitrogen and phosphorus sources. This comparison identified 108 compounds, the *in silico* growth prediction of which did not match the BIOLOG outcome. Furthermore, we added an extra set of 12 aromatic compounds that were not included in the BIOLOG assay but were known to serve as carbon source to *P. putida* [216]. Eventually, the knowledge gap set comprised a total of 120 compounds, among which 43 carbon sources, 43 nitrogen sources, 31 phosphorus sources, and 3 compounds that are both carbon and nitrogen sources (uridine, glycyl-glutamate, alanine-glycine) (table 2.2).

Initial expansion of the iJP962 model with the automatically reconstructed metabolic network yielded a disappointing total of only 3 (all nitrogen sources) out of 120 compounds, the knowledge gap of which could be closed (i.e., a complete degradation route with reactions connecting the query compound to the central metabolism was present). This observation suggested that the GSM and the automatic genome reannotation were missing catabolic pathways for the remaining 117 compounds. This prompted us to include the full set of 120 compounds as a starting point for a manual metabolic pathway curation process (further described in Experimental Procedures). The outcome of this effort allowed us to identify catabolic pathways for 92 of these compounds (32/43 carbon, 28/43 nitrogen, 29/31 phosphorus, and 3/3 carbon and nitrogen sources; see table 2.1). Some of those metabolic routes, absent from public metabolic pathways databases, are associated to extended substrate specificity of enzymatic activities experimentally described in other organisms. This aspect is illustrated for some compounds of general interest: L-2-hydroxybutyrate degradation, degradation of D-amino acids as nitrogen sources, and dipeptides degradation (see Supplemental Results and table S7 in the Supporting Information). Detailed information about the update and novelties in *P. putida* metabolic competence revealed by the present work can be found in the Supplemental Results file. In the following sections we provide an overview of novel features that may have direct relevance to control and expression of biocatalytic activities, mainly control of osmolarity, management of proton availability and transformations of aromatic compounds.

TABLE 2.2: Results of the integration of the updated catabolic pathways into the metabolic model iJP962.

	iJP962			iJP962			iJP962	
Carbon sources	Pre	Cur	Nitrogen sources	Pre	Cur	Phosphorus sources	Pre	Cur
L-Alanyl-Glycine			L-Histidine			D-Glucosamine-6-P		
Glycyl-L-Proline			Uracil			Trimetaphosphate		
Glycyl-L-Glutamic Acid			Xanthine			b-Glycerol Phosphate		
β-Hydroxy-Butyric Acid			Allantoin			D-Glucose-6-Phosphate		
γ-Hydroxy-Butyric Acid			Gly-Met			Cyclic 3′,5′-CMP		
CHEBI:73677			Met-Ala			Phosphocreatine		
a-D-Glucose			L-Cysteine			Phosphoryl Choline		
Butyric Acid			Thymine			Phosphoethanolamine		
Dihydroxyacetone			Ala-Asp D,L-a-Glycerol			Phosphate		
L-Pyroglutamic Acid			Ala-Gln			Hypophosphite		
Uridine			Ala-Glu			2′-AMP		
4-Hydroxy-L-Proline			L-Alanyl-Glycine			3′-AMP		
a-Hydroxy-Butyric Acid			Ala-His			Cyclic 2′,3′-AMP		
D-Galacturonic Acid			Ala-Leu			2′-GMP		
D-Glucuronic Acid			Ala-Thr			3′-GMP		
Quinic Acid			Gly-Asn			Cyclic 2′,3′-GMP		
b-Phenylethylamine			Gly-Gln			2′-CMP		
Bromo-Succinic Acid			Glycyl-L-Glutamic Acid			O-Phospho-D-Serine		
D,L-Carnitine			L-Pyroglutamic acid			2′-UMP		
D-Ribose			Cytidine			3′-UMP		
D-Ribono-1,4-Lactone			Uridine			Cyclic 2′,3′-UMP		
L-Alaninamide			Inosine			O-Phospho-L-Tyrosine		
Methyl Pyruvate			Xanthosine			Thiophosphate		
*Gallate			Uric acid			O-Phospho-L-Threonine		
*Glycine Betaine			D-Serine			Cysteamine-S-Phosphate		
*Choline			D-Valine			Inositol Hexaphosphate		
*Sulfate choline			AABA			3′-TMP		
*Ferulate			a-Amino-N-Valeric acid			5′-TMP		
*Phenylacetate			L-Methionine			Phospho-L-Arginine		
*Vanilate			b-Phenylethylamine					
*Vanilline			D-Asparagine					
*Coniferyl alcohol								
*p-Coumarate								
*Caffeate								
*Nicotinate								

All 96 compounds that were part of the initial 120 knowledge gaps and for which a degradation pathway was ultimately identified are included.These include: 23 BIOLOG and 12 literature-based (indicated by *) carbon sources, 31 BIOLOG nitrogen sources and 29 BIOLOG phosphorus sources. iJP962 was either expanded with the predicted reaction set (Pre), or with the curated reaction set (Cur). The colors represent no-growth (red), growth (green) or growth with the addition of an artificial transporter (orange).

Mechanisms of control of osmolarity

Living in polluted environments, *P. putida* needs to cope with highly variable concentrations of osmolytes. It must therefore build up a matching opportunity to control osmolarity by shuttling between synthesis, degradation and transport of osmolytes. This is reflected in its genome sequence by the concerted presence of genes involved in these biological processes.

Osmoregulation metabolism and transport of osmolytes

Potassium glutamate is a major regulator of osmolarity in a large panel of organisms [157]. The Kdp and Trk transport systems mediate osmoregulatory K+ uptake in a wide range of Bacteria and Archaea. In contrast to what was initially published with the sequence of the genome of *P. putida* KT2440 [290], a complete Kdp system is present in this strain (the *kdpCBAF* operon; table S8). It contains a functional high affinity P-type ATPase-K+ transporter encoded by the *kdpB* gene, previously annotated as a pseudogene. Furthermore, we have identified and annotated a novel gene which encodes the small non-essential KdpF subunit (29 amino acids) that binds and stabilizes the whole protein complex [153]. Expression of this gene is dependent on a two-component regulatory system, encoded by *kdpD* (the sensor kinase component) and *kdpE* (the response regulator component), that activates the expression of the *kdpCBAF* operon under conditions of severe K+ limitation or osmotic upshift [20].

In terms of compatible solutes transport, strain KT2440 has a functional counterpart of the proline/betaine symporter (ProP), a multidrug efflux protein of the major facilitator superfamily (MFS) that mediates the uptake and accumulation of either one of these two osmoprotectants in *E. coli* K-12. ProP allows for adaptation to increasing osmotic pressure by acting as transporter and osmosensor [260]. Exploration of the synteny conservation between *P. putida*, *P. aeruginosa* and *P. syringae* allowed us to identify additional transporters that may operate together to span the whole physiological range of osmolarity and provide optimal uptake of glycine-betaine and choline osmoprotectant molecules from the environment (figure 2.1 and table S8). As reported in experiments performed with *P. aeruginosa* [438], three transporters of the BBCT family (BetT-I, BetT-II and BetT-III) could transport glycine-betaine (BetT-II) and choline (BetT-I and BetT-III), and thus confer osmoprotection (as shown in *P. syringae*, when they are expressed in a hyperosmotic environment [73]). Moreover, a complete choline-betaine-carnitine (CBC) ABC transport system is encoded in the *cbcXWV* operon. The expression of the operon is induced by an AraC-family transcriptional activator (encoded by *gbdR*) in response to glycine-betaine and dimethylglycine [74, 438]. In fact, three different periplasmic substrate-binding proteins in *P. putida*, encoded by the *cbcX*, *caiX*,

and *betX* genes (figure 2.1 and table S8), show high specificity for choline, carnitine, and betaine, respectively [74]. Finally, a small multidrug resistance (SMR) protein, homolog of the *E. coli* K-12 EmrE protein, is also present in *P. putida* KT2440. It could be associated to choline and glycine-betaine export in response to intracellular levels of both osmoprotectants [29].

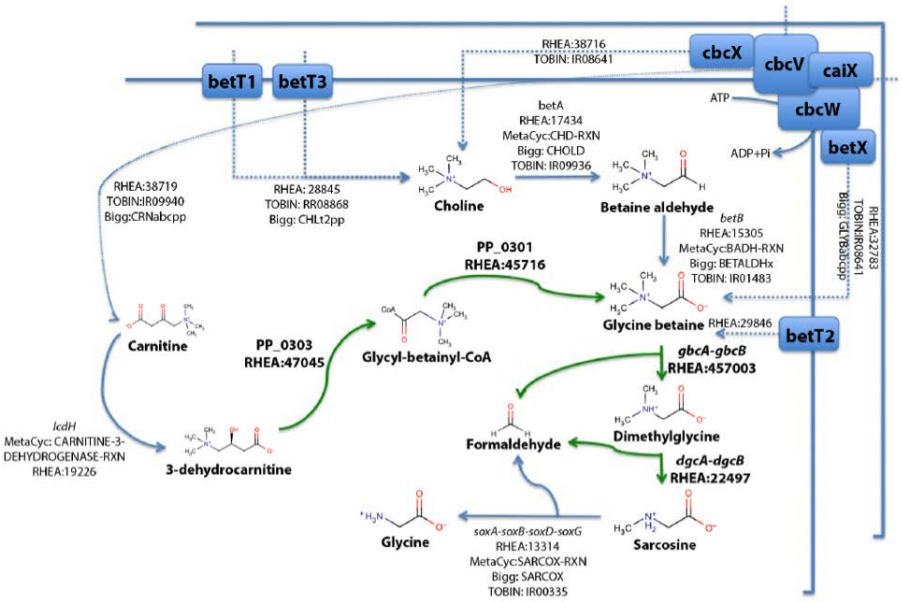


FIGURE 2.1: Schematic representation of the glycine-betaine, carnitine and choline metabolism in *P. putida* KT2440. Blue arrows represent transport reactions (dotted lines) and cytoplasmic reactions (continuous lines) that were present in previous KT2440 GSMs (Bigg = iJN746, TOBIN = iJP962; different namespaces). Green arrows show new GPR associations curated during our reannotation process (they were missing in previous KT2440 GSMs) CoA, coenzyme A.

Glycine-betaine degradation

In addition to the annotation of choline and glycine-betaine transporter genes, the annotation of genes involved in the aerobic degradation of these compounds has also been updated. The *betIBA* operon (figure 2.1 and table S8) encodes a choline-responsive transcriptional repressor (BetI), and two enzymes, a choline oxidase (BetA) and a betaine aldehyde dehydrogenase (BetB), responsible for the two-step conversion of choline to glycine-betaine [366, 434, 468]. As in *P. aeruginosa*, the genes encoding the choline transporter BetT1

and the *betIBA* operon are divergently transcribed in *P. putida* KT2440, allowing rapid transcriptional response to choline [366]. Finally, comparative genomics allowed us to identify orthologs of the *P. aeruginosa* PAO1 genes involved in the three-steps demethylation of glycine-betaine to glycine, a metabolic pathway essential for growth with glycine-betaine as the sole carbon source [440]. This pathway includes a novel demethylase activity associated to the GbcAB enzyme complex that catalyzes the initial demethylation of glycine-betaine to dimethylglycine and formaldehyde. This operates via a process involving a dioxygenase and differs from the process mediated by the betaine-homocysteine S-methyltransferase present in other choline degraders like *Sinorhizobium meliloti* [389, 440]. In *P. putida* KT2440, an heterodimeric flavin-linked oxidoreductase, encoded by the *dgcA* and *dgcB* genes (table S8), catalyzes the second demethylation reaction of dimethylglycine to sarcosine, which is further demethylated to glycine in a reaction catalyzed by a heterotetrameric sarcosine oxidase complex encoded by the gene cluster *soxBDAG* (figure 2.1).

Trehalose-glycerol metabolism

Due to its electroneutral nature and its role as a protein stabilizer, the disaccharide trehalose is a major osmoprotectant in bacterial cells [204, 369]. The *P. putida* genome reannotation process revealed a complex metabolic scenario where trehalose could play a central role both in osmoregulation and in the metabolism of glycogen (figure 2.2). This differs from the metabolic profile present in most γ-Proteobacteria where this role is fulfilled by monosaccharide nucleoside diphosphates [67]. *P. putida* KT2440 lacks the *ostAB* genes encoding enzymes involved in the two-step trehalose biosynthesis pathway from UDP-glucose via a trehalose-6-phosphate intermediate [198]. Rather, it displays two alternative pathways for trehalose biosynthesis (figure 2.2A and table S8). The first one involves the *PP4053* protein (previously annotated as a generic glycosyl hydrolase) that is highly similar to the malto-oligosyl trehalose synthase (TreY) from other *P. putida* strains. Together with the malto-oligosyl trehalose hydrolase (encoded by the *treZ* gene), TreY catalyzes the biosynthesis of trehalose from glycogen [100, 223]. The second pathway is associated to two different trehalose synthases (coded by *treSA* and *treSB* genes) that catalyze the reversible single-step conversion of maltose to trehalose [67, 242, 369]. These enzymes belong to two evolutionary distinct lineages. The corresponding genes do not display sequence similarity and are involved in different genomic and metabolic contexts. The *treSB* gene (*PP4059*; table S8) encodes a fused protein (a trehalose synthase belonging to a family widely

distributed across different bacterial lineages and a maltokinase) and is clustered with genes encoding the glycogen branching enzyme GlgB, and the α-1,4-glucan:maltose-1-phosphate maltosyltransferase GlgE (figure 2.2B). These genes form an operon for a novel glycogen biosynthesis pathway similar to the variant recently discovered in *Mycobacteria* (figure 2.2C), that uses α-maltose-1-phosphate instead of UDP-glucose-6-phosphate as the building block to extend glucan chains [67, 115, 369]. By contrast, the trehalose synthase encoded by the *treSA* gene (*PP2918*; table S8) belongs to a small family of highly active trehalose synthases. It has been biochemically characterized in *P. stutzeri* CJ38 as a biocatalyst of biotechnological interest for the production of trehalose [242]. We propose that this second *P. putida* trehalose synthase may have a role in the control of osmolarity.

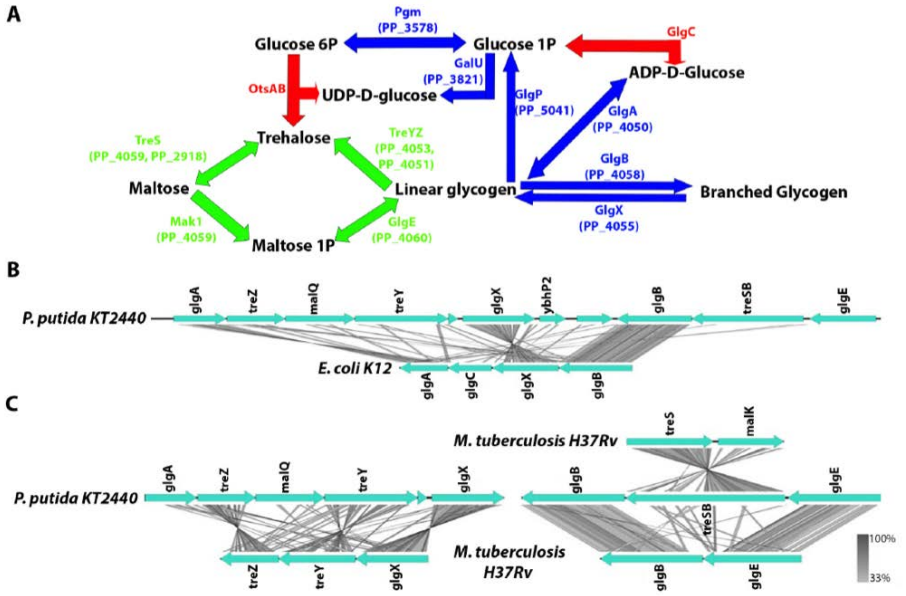


FIGURE 2.2: Trehalose metabolism in *P. putida* KT2440. **(A)** Metabolic pathway of trehalose biosynthesis in *P. putida* KT2440 and *E. coli* K12. Reactions specific to *P. putida* are shown in green and those specific to *E. coli* are shown in red. Shared reactions are represented in blue **(B)** Lineplot showing tblastX similarities between genomic regions containing genes involved in the trehalose metabolism in *P. putida* KT2440 and in *E. coli* K12 **(C)** same as (B) between *P. putida* KT2440 and *Mycobacterium tuberculosis* H37Rv. The gene cluster is splitted in three different genomic regions in *M. tuberculosis* H37Rv. In this organism, *treS* and *malK* genes are not fused. ADP, adenosine diphosphate; UDP, uridine diphosphate.

Control of the proton gradient

P. putida KT2440 is an obligate aerobe that uses the EDEMP cycle (composed by activities from the Entner-Doudoroff, the incomplete Embden-Meyerhof-Parnas, and the pentose phosphate pathway) to process glucose [295]. Furthermore, it lacks the glucose-specific phosphoenolpyruvate:carbohydrate phosphotransferase system (PTS) that usually fuels the EMP pathway in other bacteria, such as *E. coli*. Yet, apart from sugars its growth environment provides a considerable number of compounds that may enter its metabolism at various points. This, in turn, requires the presence of a large number of transport systems, as illustrated by the coding capacity of its genome. The processes encompassing oxygen availability and utilization, carbon catabolism and transport suggest that a considerable amount of protons are involved: they could be channeled during respiration to form ATP and in proton/metabolite co-transport activities. *P. putida* KT2440 possesses counterparts of the cytochrome bo oxidase and the cytochrome bd-I oxidase found in many bacteria. However, it does not have a counterpart of *E. coli* cytochrome bd-II oxidase (AppCD). The activity of cytochrome oxidases contributes to build up a proton motive force [40]. The proton gradient thereby generated is challenged when the pH of the environment varies. We thus wanted to explore the way the bacterium maintains proton homeostasis through critical examination of its genome sequence. *P. putida* is a neutrophilic organism and harbors a standard version of most of the general processes involving protons (ATP synthase, assembly of flagellar motor, NADH/NADPH balance, etc.). It differs, however, from other classes of γ-Proteobacteria such as *Enterobacteria* in the way it manages the acid resistance response and the transport of protons.

Acid resistance response

The acid stress response involves many different processes in species having a periplasm [257], where some enzymes may have an acidic optimum pH for activity (e.g. AppA in *E. coli* [155], a gene not found in *P. putida*). The results of the functional reannotation shows that *P. putida* KT2440 has orthologs of the *E. coli* K-12 genes encoding the alternative sigma factor RpoS and the cAMP receptor protein Crp, that constitute the glucose-repressed Acid Resistance system AR1 allowing cell survival at pH 2.5 [143, 279] (table S9). In *P. putida* RpoS restores the acid resistance phenotype missing in *rpoS*-deficient *E. coli* mutants. Yet, it seems that in *P. putida* the role of RpoS is mainly associated to adaptation to carbon starvation conditions [353]. RpoS and Crp are global regulators which control expression of multiple genes (regulons) under conditions when global resource allocation needs to be modified as the environment changes [184]. However, the regulatory network of both Crp and RpoS activities is noticeably different in *P. putida* when compared to that of *E. coli*, in line with the widely different niches of the organisms [279, 436]. Expression profiling studies in the KT2440 strain using different carbon sources revealed a strong expression of *rpoS* in cells growing with glycine and fructose as carbon sources [145, 214]. A further difference can be pointed out: *P. putida* KT2440 has neither orthologs of the *E. coli* decarboxylase-antiporter systems AR2 (glutamate-decarboxylase isozymes GadA, GadB and 4-aminobutanoate (GABA)-glutamate antiporter GadC), nor of AR3 (degradative arginine-decarboxylase AdiA and agmatine-arginine antiporter AdiC). As far as the Acid Resistance system 4 (AR4) is concerned, the *PP4140* gene, previously annotated as a pseudogene, was now found to be complete. It is similar to the *E. coli* lysine decarboxylase (*ldcC* gene, encoding a constitutive form of the lysine decarboxylase). However, there is no signal of neighbor cadaverine-lysine antiporter CadB characteristic of the AR4 system [143]. Overall *P. putida* lacks most of the acid stress response present in *Enterobacteria*. This may contribute to its recognized lack of pathogenicity, but it needs to be taken into account when *P. putida* is used for biocatalysis in a reactor as well when the organism is used for *in situ* or *ex situ* bioremediation of polluted environments.

Otherwise, *P. putida* KT2440 has functional alternative pathways for the degradation of both L-arginine and GABA, which involve enzymatic activities induced in high pH conditions in *E. coli*. The *P. putida* annotated homologous genes are listed in table S9.

Transport of protons

Protons are involved in many transport systems, including vectorial transport for ATP synthesis, as well as in the mechanical rotation of flagella. *P. putida* KT2440 has two counterparts of the Na+/H+ antiporter NhaA (table S9), the best-understood antiporter which helps maintain the internal pH, protecting cells from excess sodium at high pH [108, 396]. This species also harbors a putative multidrug efflux protein MdfA that extends the pH tolerance range up to pH = 10 in *E. coli*, taking over when NhaA is deleted [247]. However, we did not identify a homolog to the positive regulator NhaR, which controls NhaA activity during exponential growth [64, 349]. This may be compensated for by a possible activity under the control of the functional RpoS sigma factor together with genes of the RpoS regulon also involved in pH homeostasis in the stationary growth phase [108]. The *P. putida* KT2440 genome also harbors a second pH-independent Na+/H+ antiporter, NhaB [325, 335], as well as five proton-sodium antiporters of the monovalent cation:proton antiporter (CPA) families CPA1 and CPA2 (*nhaB* and *nhaP* genes; table S9). Three additional glutathione-gated K+ efflux systems (kef genes) of the CPA2 family have also been found. They are likely to be important for coping with mechanical stress induced by the considerable variations of metabolites charges and concentrations associated with *P. putida* metabolism in a chemically polluted environment. In *B. subtilis* the essential operon *mrpABCDEFG* encodes for a transport system of the CPA3 family, which provides Na+/H+ antiport activity and functions in resistance toward several different compounds and pH homeostasis [55, 199]. As in bacteria from great many other clades (e.g. in *Bdellovibrio bacteriovorus*, *Bordetella pertussis*, *Deinococcus radiodurans*, and *Mycobacterimum smegmatis*), but not in *E. coli*, *P. putida* KT2440 has a complete operon counterpart of *phaABCDEFG* (table S9). This system is widely present in *Pseudomonadales*, where its organization differs slightly from that of *Firmicutes*: MrpA counterpart is fused to MrpB (PhaAB protein). Moreover, 22 additional Major Facilitator Superfamily transporters (MFS) were identified in the *P. putida* KT2440 proteome; they could contribute to pH homeostasis through additional Na+/H+ or K+/H+ antiporter activities, as it is the case for the multidrug efflux protein MdfA in *E. coli* or the tetracycline resistance protein TetL in *B. subtilis* [325] (Padan et al., 2005).

Finally, the annotation of five genes encoding periplasmic and outer-membrane proteins associated to pH homeostasis has been updated in *P. putida* KT2440 (table S9). These genes include the extreme base-induced membrane-bound redox modulator Alx [396], as well as the peptidyl-prolyl cis-trans isomerase SurA, which is necessary for proper folding of outer membrane proteins and whose inactivation is lethal in stationary phase under elevated pH conditions [142, 426].

Aromatic compounds degradation pathways

One of the most relevant metabolic features of *P. putida* KT2440 is its ability to break down into central metabolic intermediates a wide range of aromatic compounds that are present in the rhizosphere and associated to the recycling of plant-derived material prevalent in the environment [290, 327, 372]. This prompted us to explore the gene complement of aromatic degradation pathways, not only for carbon-only aromatic compounds, but also aromatic heterocycles, including purines and pyrimidines.

Degradation of carbon-skeleton aromatics

In addition to the outcome of BIOLOG experiments, much experimental evidence has been reported since the first publication of the genome sequence of *P. putida* KT2440. This had impacted the annotation of genes involved in the degradation of aromatic compounds [216, 299, 372, 453]. The present upgrade includes genes involved in the central aromatic compounds degradation pathways as well as a variety of connected pathways. A summary of the new vision of the aromatic catabolism of strain KT2440 is represented in figure 2.3, and detailed in the Supplemental Results file (see table S10 for a complete description).

We propose a candidate gene for the orphan enzyme (i.e., a defined enzyme without assigned sequence) responsible for the first redox step of the two-step degradation of coniferyl alcohol to ferulate [193, 298]. The *PP2426* gene, corresponding to the alcohol dehydrogenase activity CalA (EC 1.1.1.194; table S10), is likely to encode a coniferyl dehydrogenase, with somewhat promiscuous activity. This candidate gene shows significant similarity with cinnamyl-alcohol dehydrogenases of plant origin (about 50% amino acid identity over the whole protein length), which are also able to act on coniferyl alcohol (see IUBMB annotation, EC 1.1.1.195). The proposed coniferyl dehydrogenase CalA (PP2426) would work together with coniferyl aldehyde dehydrogenase CalB (PP5120, EC 1.2.1.68) [193, 323]. However, an experimental validation is necessary to substantiate this prediction.

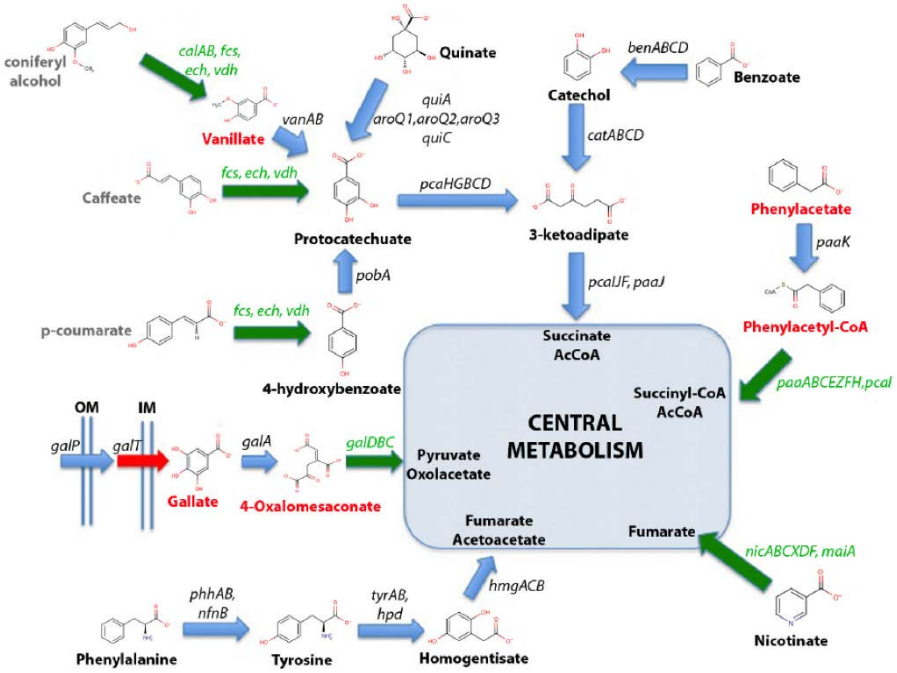


FIGURE 2.3: Schematic representation of the catabolism of aromatic compounds to central metabolites in *P. putida* KT2440 (adapted from [298]). Red compounds correspond to dead-end metabolites in iJP962 (isolated compounds that are either only consumed or only produced by the model). Gray compounds represent aromatic compounds absent in iJP962. Green arrows and green genes represent new GPR associations curated during the *P. putida* KT2440 genome reannotation process (they were absent in the original iJP962). Flux Balance Analysis (FBA) simulations over the extended iJP962 gave rise to a functional phenotype in terms of biomass production with all these aromatic compounds as external carbon sources. OM, outer membrane; IM, inner membrane; CoA, coenzyme A; and AcCoA, acetyl-coenzyme A.

Degradation of nucleotides and other heterocyclic aromatics

The positive redox phenotypes observed in BIOLOG experiments using uracil and thymine as nitrogen sources led us to reannotate a gene cluster which contains all the genes involved in the reductive pathway of pyrimidine nucleotides [378, 445] (table S10). This pathway starts with the reduction of uracil and thymine to the corresponding 5,6-dehydro-derivatives by a type II NADPH-dependent dihydropyrimidine dehydrogenase (DPD) enzyme complex PydXA [179, 321]. The dehydropyrimidines are subsequently hydrolyzed

by a bifunctional D-hydantoinase/dihydropyrimidinase (*pydB* gene) and a β-ureidopropionase (*hyuC* gene) into β-alanine and 3-amino-isobutyrate respectively [378, 445]. The *PP4036* gene (*pydB*), was originally annotated as a pseudogene (sequencing error), but it is likely to be fully functional as it encodes a protein highly similar to the experimentally characterized D-hydrantoinase/dihydropyrimidinase from *P. putida* (*Arthrobacter capsulatus*) [75]. The gene cluster also encodes a permease commonly present in β- and γ-*Proteobacteria* (*pydP* gene), as well as a transcriptional regulator (the *PP4039* gene is similar to the *E. coli rutR* gene) (table S10).

In the same way, the positive redox phenotype observed in BIOLOG experiments when using xanthine, urate or allantoin as nitrogen sources, allowed us to upgrade the annotation of a gene cluster involved in the transport and degradation of purine nucleotides (table S10). It includes the xanthine dehydrogenase enzyme complex XdhABC that catalyzes the NAD+-dependent oxidation of hypoxanthine and xanthine to urate [328], and two of the three enzymes involved in the degradation of urate to S-allantoin [351]: the hydroxyisourate hydroxylase (PucM) and the 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline (OHCU) decarboxylase (PucL). These proteins belong to two chromosomal clusters and share homologies with eukaryotic and prokaryotic proteins (COG2351 and COG3195, respectively). They also display similar co-evolution phylogenetic profiles [117, 432]. This suggests a common evolutionary gain and loss history, as illustrated in other organisms harboring this pathway [351]. Furthermore, S-allantoin can be degraded in four steps to glyoxylate via S-ureidoglycine as an intermediate, releasing ammonia and urea. The first step involves a novel metal-independent allantoinase encoded by the *puuE* gene that differs from the *E. coli* K-12 allantoinase (*allB* gene) [352].

Finally, the annotation of the *nic* gene cluster (*nicPTFEDCXRABS*), responsible for the aerobic degradation of nicotinate to fumarate, has also been updated. It allows *P. putida* KT2440 to grow with nicotinate as both nitrogen and carbon source [192].

Towards an extended view of the KT2440 metabolic model

The updated genome annotation provided us with a list of functions, e.g., chemical conversions, that were not previously identified in *P. putida*. However, the effect of an individual function on systems-wide behavior is not straightforward. For example, a candidate degradation pathway can eventually be deemed non-functional if its by-products cannot be further processed. We decided to assess the full impact of the updated annotation by complementing an existing genome-scale metabolic model with the new reactions. This allowed us to check whether the identified enzymatic conversions could

truly function in the context of the former knowledge of *P. putida* metabolism, and to pinpoint additional knowledge gaps to be addressed in future studies. Specifically, for 96 out of the 120 defined knowledge gaps, we identified a probable degradation pathway during the targeted manual annotation process. Together, these pathways comprised a total of 253 reactions, 234 of which have been assigned to one or more genes and integrated into MicroScope. Moreover, 43 new ChEBI compounds and 73 new RHEA reactions were created during this curation process.

To assess whether these reactions indeed coped with the knowledge gaps, we expanded the iJP962 metabolic model with the degradation pathways and mimicked *in silico* the BIOLOG experiment (see Experimental Procedure section). Surprisingly, this expansion led to an *in silico* positive phenotype for only 20 compounds (out of 96). However, it is important to recall that the BIOLOG setup does not measure growth per se but the integrated activity of redox networks [44]. This relatively small improvement prompted us to inspect the remaining cases in more detail (table S12). A major issue turned out to be the difficulty in identifying transport proteins for specific compounds; our list of curated reactions only contained 23 transporters. To further test the existence of degradation pathways, we complemented the GSM with *ad hoc* transport reactions that behaved as passive diffusion reactions. This improved the outcome of the model, as 72 out of 96 degradation pathways were now functional. Interestingly, even in the original model the addition of *ad hoc* transporters resolved 10 of the knowledge gaps, indicating that for some compounds the lack of a transport reaction was the only functional step preventing *in silico* growth. This procedure also led to *in silico* positive growth phenotypes for 14 compounds with a negative BIOLOG phenotype (table S12), demonstrating that it is essential to get experimental evidence for transport systems. Although such results require future *in vitro* confirmation, they suggest that the range of suitable substrates for *P. putida* may be increased with the sole identification of the corresponding transporter proteins. This observation highlights an essential area for future research that will lead to improve GSMs.

Still, successful *in silico* metabolite degradation was yet to be achieved for 24 out of the 96 compounds with identified degradation pathways. The underlying causes of these remaining knowledge gaps may be roughly divided into 4 categories (table S13):

(i) *Level of detail*. The degradation pathways for 7 compounds involved ill-defined metabolite classes, such as 'NADPORNOP', and 'Oxidized-cytochromes'. Where possible, we replaced these with specific instances of these classes, such as NAD and Ferricytochrome.

(ii) *By-product accumulation*. The degradation pathways for 6 compounds resulted in by-products that the *in silico* cell was unable to dispose of. In particular, 5 degradation pathways led to an accumulation of sulfur-containing compounds. We complemented the model with sulfate, hydrogen sulfide, and sulfite exporters, which allowed successful degradation of 4/5, 1/5 and 5/5 compounds. We show below that *P. putida* KT2440 has 11 candidate *tauE* genes, which may encode a sulfite exporter. The sulfite export reaction and the 11 corresponding genes were thus added to the curated reaction list.

(iii) *Reaction reversibility*. The degradation of one compound, D-glucosamine-6-phosphate, was hampered by a reaction that was irreversible in the model, but reversible according to external sources such as MetaCyc [66] and Brenda [70]. We adjusted the reaction accordingly.

(iv) *Open issues*. Ten out of the degradation pathways led to the production of dead-end metabolites in the model. Dead-end metabolites are metabolites that can either only be produced, or only consumed in the model. We were unable to link possible degradation pathways for these compounds to *P. putida* genes. These non-functioning degradation pathways and the corresponding metabolites highlight a remaining knowledge gap in *P. putida* metabolism to be addressed in future studies.

In addition, we assessed how the expanded model performs in a broader *in silico* growth analysis including both wild-type and mutant growth predictions. We distinguished between predictions for wild-type growth and for mutant growth because these reflect different qualities of a GSM. Wild-type growth predictions indicate whether the GSM includes any pathways that can convert a specific combination of medium constituents into biomass. In contrast, mutant growth predictions assess the quality of the Gene-Protein-Reaction (GPR) associations and the appropriate inclusion or exclusion of alternative pathways. The wild-type growth dataset consisted of the full BIOLOG dataset and 12 additional compounds with literature back-up. The mutant growth dataset was made of a combination of two external datasets: the original test-set for the iJP815 model [341] and experimental data that was published later on [11]. As expected, the accuracy of wild-type growth predictions increased marginally when the GSM was expanded with the automatically predicted reaction set (0.59 to 0.66), but increased substantially when expanded with the curated reaction set (0.59 to 0.8) (table 2.3 and table S12). In contrast, the accuracy of mutant growth predictions decreased considerably when the model was expanded with the automatically predicted reaction set (0.78 to 0.67), but remained stable when expanded with the curated reaction set (0.78 to 0.78), although specificity and sensitivity did change (table 2.3 and table S12). Overall, these results indicate that the curated reaction set is a solid expansion of the existing GSM, while the predicted reaction set reveals discrepancies between the updated annotation and the existing GSM.

TABLE 2.3: Model evaluation. iJP962 as well as the extensions based on predicted (Pre) and curated (Cur) degradation pathways were tested in terms of phenotype predictions (growth/no-growth).

	iJP962	iJP962 + Pre	iJP962 + Cur
Metabolites	980	1375	1122
Reactions	1066	1533	1256
Genes	949	1203	1053
Wild-type predictions			
Specificity	0.90	0.86	0.88
Sensitivity	0.42	0.55	0.75
Accuracy	0.59	0.66	0.79
Mutant predictions			
Coverage	0.70	0.68	0.70
Specificity	0.74	0.56	0.72
Sensitivity	0.72	0.71	0.80
Accuracy	0.73	0.60	0.75

We used both wild-type and mutant growth data [11, 341]. The experimental mutant data comprised gene knockout data in defined media as well as experimentally verified auxotrophies.

We expect future work on *P. putida* metabolic modeling to use the predicted reaction list in conjunction with the available GSMs in order to identify faulty reactions or GPR associations in both our reaction list and the existing GSMs. Although we used the iJP962 GSM as the current knowledge on *P. putida* metabolism in order to contextualize the annotation, it is possible that there are errors in this GSM that became apparent upon expansion with the predict reaction set. Using algorithms such as Growmatch [235], reactions can be selectively included or excluded in a GSM in order to increase the correspondence between *in vivo* observations and *in silico* predictions. For example, the yeast GSM was recently updated to version 6.0 by removing ill-supported GSM reactions and adding reactions based on updated annotation and experimental literature. This led to a substantial increase in accuracy for predicting mutant growth phenotypes [170]. We anticipate that the predicted reaction set for *P. putida* based on the updated annotation will facilitate a similar improvement of the *in silico* mutant growth predictions.

Comparison to other *Pseudomonas putida* strains

Uduando *et al.* [431] have recently reported a comparative analysis of the genomes of nine *P. putida* strains aimed at determining the core collection of genes that give identity to this species. Although the number of strains examined is somewhat limited, the results revealed the lack of pathogenic traits (e.g. exotoxins and type III secretion systems are absent in all cases) and the centrality of the Entner-Doudoroff pathway as the key route for consumption of carbohydrates. Such a core genome (paleome [1, 457]) of *P. putida* consisted of approximately 3,380 genes, a good share of which encoded transporters, both for nutrients and for electrons, which seemingly enable aerobic metabolism under different oxygen regimes. Other genes of the core set determined the pentoses phosphate cycle, arginine and proline metabolism, and different routes for degradation of aromatic chemicals. Amino acid metabolism (synthesis and degradation) was very conserved as well and encoded in each case complete sets of transporters, enzymes and regulators. Flagellar biosynthesis and genes for biofilm formation belong to the *P. putida* core genome as well.

Despite a large number of differences between strains, the wealth of information on strain KT2440 discussed above makes this specimen the reference for the whole group. Many of the general traits discussed above that make special strain KT2440 can be properly extended to other members of the *P. putida* group [293], with the caveat that the *P. putida* group is somewhat fuzzy, strain 2440 lying slightly distant from the reference type strain DSM291 [460].

Conclusions

In this work we have coupled re-sequencing of the *P. putida* KT2440 genome to a complete upgrade of its sequence annotations as means to provide a standard for use of this organism as a versatile chassis for both fundamental and biotechnological endeavors. Over the last few years, *P. putida* strains have been increasingly recognized for their potential to host bioreactions that other model bacteria fail to execute (e.g., strongly oxidative biotransformations). An attractive trait of strain KT2440 that makes it adequate for such applications is the fact that this bacterium harbours a large number of metabolic and stress-endurance properties optimal for biotechnological needs. In our present study, we further highlight the potential of *P. putida* for biotransformations and biodegradation by disclosing mechanisms controlling osmolarity and pH homeostasis. While resequencing per se only provided marginal improvement in the sequence, the update of the annotation allowed us to propose a consistent picture of *P. putida* metabolism. Coupled with experimental

data using the BIOLOG setup this enabled us to considerably improve the outcome of a systems biology approach where GSM predictions could be matched with experimental data. The present state of affairs demonstrates that while there remain some knowledge gaps in the *P. putida* metabolism, we now have a clear picture of its overall functioning. Our approach pinpointed a specific deficiency in our knowledge: we need to considerably improve explicit identification of transport systems. This should be a major task for the immediate future of studies with the *P. putida* chassis, but also for other chassis as well.

In this respect, the present update of the genome sequence of *P. putida* and its annotation emphasized considerable differences with the ubiquitous model organism used as a chassis in many studies, *E. coli* K-12. Indeed, *Enterobacteria* and related bacteria differ considerably from *Pseudomonadales*, and *P. putida* may be an excellent reference model of this clade. Beside metabolic differences that have been outlined in the present article, the way DNA is handled is quite different in these clades, and this may be of importance for studies involving DNA constructs meant to provide novel metabolic engineering approaches. An example of this are the different contingents of DNA polymerase III proteins in different species. In *P. putida* one finds four different DNA polymerase III proteins, three variants of DnaE (DnaE1, DnaE2 and DnaE3) and a second type, PolC [423]. Organisms such as *B. subtilis* combine DnaE1 and PolC [117]. By contrast, *E. coli* has only DnaE1. A second DnaE variant appears as a heterologous subunit of the enzyme when the length of the genome sequence increases. Furthermore, the presence of DnaE2 together with DnaE1 is linked to bacteria featuring large GC-rich genomes and living in aerobic environments [423], as in the case of *P. putida* (*dnaEA*: *PP1606* and *dnaEB*: *PP3119*). Analysis of the co-evolution of the genes that are present in parallel with DnaE2 will certainly help identification of functions that are highly relevant both to the ecological niche of the organism and to its use as a cell factory.

Materials and Methods

***P. putida* sequencing.** The genome of *P. putida* KT2440 DSM 615 was sequenced using Illumina sequencing technology. Paired-end libraries were prepared with fragment size of 300-600 bp and sequenced on HiSeq2000 (100 nt length). The total of 8,786,896 sequence reads produced were processed to remove low-quality reads and mapped over the *P. putida* KT2440 reference genome sequence.

SNPs/InDels detection strategy. High Throughput Sequencing (HTS) data were analyzed using the PALOMA pipeline (Cruveiller S., unpublished) implemented in the Microscope platform [432]. The current pipeline is a "Master" shell script that launches the various modules of the analysis (i.e., a collection of in-house software written in C) and controls for all tasks having been completed without errors. In a first step, the HTS data quality was assessed by including options like reads trimming or merging/split paired-end/mate-paired reads. In a second step, reads were mapped onto the original sequence of *P. putida* KT2440 (Accession Number NC_002947; AE015451.1) using the SSAHA2 package [297]. Unique matches having an alignment score equal to at least half of their length were retained as seeds for full Smith-Waterman realignment [391] keeping at both sides a region of the reference genome extended by five nucleotides. All computed alignments were then screened for discrepancies between read and reference sequences and in fine, a score based on coverage, allele frequency, quality of bases and strand bias was computed for each detected event to assess its relevance. The results generated are available at the MicroScope platform (http://www.genoscope.cns.fr/agc/microscope).

Consensus sequence correction. To correct the original sequence of *P. putida* KT2440, the PALOMA pipeline was run with stringent parameters for the « SNP calling » step (allelic frequency set to 0.8 with at least ten reads mapping the position, a balance of forward reads to reverse reads set to 0.33). This analysis led to a relatively small amount of variations compared to the original one, showing that the 2002 sequence was of excellent quality [290]. An automated process was subsequently implemented to generate a new version of the sequence of the *P. putida* strain KT2440 genome using both the original sequence and the list of detected variations as inputs. During the process, uncovered areas of the reference genome were reported as well, corresponding either to repeats (discarded by default during the reads mapping step) or potentially large deletions in the re-sequenced genome.

RNA-seq Analysis. The complete transcriptome high-throughput sequencing data published in [214] was retrieved from the GEO database [25] (accession no. GSE42491). Data were then analyzed in the MicroScope platform with the workflow TAMARA [432]. The current pipeline is a "Master" shell script that launches the various parts of the analysis (i.e. a collection of Shell/Perl/R scripts) and checks that all tasks are completed without error. Reads pre-processing and mapping steps are performed in the same way as the PALOMA pipeline (see SNPs/InDels detection strategy section for details). After reads were mapped on the newly annotated *P. putida* str. KT2440 genome, we minimized the false positive discovery rate using

SAMtools (v.0.1.8; [250]) to extract reliable alignments from SAM-formatted files. The number of reads matching each genomic object of the reference genome was then calculated with the Bioconductor-GenomicFeatures package [240]. When reads matched several genomic objects, the count number was weighted so as to keep the total number of reads constant. Finally, the Bioconductor-DESeq package [9] was used with default parameters for the analysis of raw count data and to determine whether expression levels differed between conditions.

Structural reannotation of the *P. putida* genome. The corrected genome sequence was subsequently processed by the MicroScope pipeline for complete structural and functional annotation [432]. Gene prediction was performed using the AMIGene software [45] and the microbial gene finding program Prodigal [185] known for its capability to locate the translation initiation site with great accuracy. The predicted genes were compared with those listed in the original annotation (AE015451, version: 05-MAR-2010). Manual curation was performed on the two sets of unique genes (see Results) by taking into account transcriptomic information from [145] and [214] experiments, as well as conservation of sequence similarity and genomic context with homologs in other genomes. A total of 80 unique GenBank genes, and of 296 unique AMIGene CDSs were considered false positive predictions and discarded from the final annotations (artifact status). The 309 unique AMIGene predictions considered as newly predicted *P. putida* genes are numbered starting from the last original annotation (*PP5420*) (i.e., *PP5421*, table S4). The RNAmmer [237] and tRNAscan-SE [255] programs were used to predict rRNA and tRNA-encoding genes respectively, whereas other RNA structures like small RNAs and riboswitches were identified using the RFAM database [60] (n = 65) and from publications (n = 3) [145]. Finally, intra-chromosomal repeats were detected using the method described by [2].

Functional automatic annotation. The predicted/annotated genes were subjected to sequence similarity searches using the gapped blastP algorithm against the UniProtKB protein sequence knowledgebase [88] and several protein family resources: COG [148], HAMAP [331] and FIGfam [276]. They were also processed using the InterProScan software to predict potential sequence motifs, patterns and protein family assignments compiled in InterPro [281]. In addition, genes encoding enzymes were also classified using the PRIAM profiles [85]. In terms of predicted structural features, α-helical transmembrane regions were searched with the TMHMM program [230] and signal peptides with SignalP [332]. Finally, to predict probable subcellular localization of the annotated protein in the cell, PSORTb predictions were also carried out [289]. Using the MicroScope platform, *E. coli* K-12 expert annotation is already an

ongoing process since the work described in [427], with a main focus in the curation of Gene-Protein-Reaction (GPR) associations coming from EcoCyc [209] and literature data. Then, in order to (re)assign functions to each *P. putida* KT2440 annotated genes, bi-directional best-hit (BBH) between *P. putida* KT2440 and *E. coli* K-12 genes were first identified by BLASTP, and annotation transfer from *E. coli* K-12 to *P. putida* KT2440 genes was carried out based on this BBH relationships and the following similarity thresholds: 50% identity on 80% of the length of the longest protein, or 40% identity on 80% of the length of the longest protein in case of shared genomic context or FIGfam protein families assignments [276]. *P. putida* KT2440 annotation transfer includes the transfer of these GPR associations from *E. coli* K-12 counterpart, a feature that improves the subsequent genome-scale metabolic network reconstruction (see below). A total of 706 genes were reannotated using this process. *P. putida* genes escaping the *E. coli* K-12 functional annotation transfer were annotated following the standard MicroScope procedure [432]. Finally, during the curation process of gene function, chosen gene names conform to the nomenclature conventions derived from [101].

Automatic genome-scale metabolic network reconstruction. The metabolic network of *P. putida* KT2440 was reconstructed from the reannotated genome sequence stored in PkGDB using the MicroScope automatic reconstruction pipeline, which is based on the BioCyc pathway reconstruction software [202]. Pathway Tools uses the set of genome annotations as input data to automatically project the set of reference metabolic pathways stored in MetaCyc database [66], generating a specific Pathway Genome Database (PGDB) in a two-step process: first, the Reactome projection step, where associations between genes and metabolic reactions are inferred from gene annotations, and the Pathway projection step where reference pathways are projected based on these gene-reaction associations (see [202] for further details on the algorithm). The Reactome projection step in MicroScope is enhanced by implementing an export procedure from MicroScope PkGDB to Pathway Tools input format that directly associates the genes to MetaCyc reaction identifiers coming from manual validation by MicroScope curators or from automatic reaction transfer from reference organisms [432]. This allowed minimization of the over-prediction or missing of relevant enzymatic reactions resulting from inaccurate or unclear textual annotations. The reconstructed genome scale metabolic network of *P. putida* KT2440 is included in the MicroCyc repository available at http://www.genoscope.cns.fr/agc/microcyc.

Namespace conversion. To integrate the novel GPRs into the GSM iJP962, the GSM was first converted into the standardized MNXref namespace [36] at MetaNetX.org [150] to facilitate integration of reaction sets from external sources. Subsequently, the novel reaction sets were also converted into the MNXref namespace using a custom script that takes into account the metabolites that are already included in the converted GSM. This was done in order to account for possible differences in level of detail between the different sources. For example, the compound glucose can correspond to D-glucose or L-glucose, which in turn can correspond to α-D-glucose, β-D-glucose, α-L-glucose, or β-L-glucose. In order to correctly connect new reactions to an existing GSM, their metabolites thus need to not only be converted into the same namespace, but also at the correct level of detail. Metabolite names that had multiple plausible alternatives in the GSM were manually checked following the logic of metabolic reactions [97].

Model extension. For each predicted or curated reaction from the functional reannotation process (hereafter: new reactions), the GSM iJP962 was first scanned to search for reactions involving the same set of metabolites as the new reactions. Only if no such reaction existed was the new reaction added to the model. Otherwise, the existing and new reactions were compared in terms of reaction directionalities and associated genes. If the new reaction had been manually curated, the GSM reaction was updated in terms of both reaction directionality and gene associations. However, if the new reaction had not been manually curated the GSM reaction directionality was left unchanged. In addition, the gene associations of the GSM reaction were only updated if it was an orphan reaction.

Growth phenotype data using BIOLOG experiments. *P. putida* KT2440 DSM 6125 was tested for its ability to utilize different carbon (C), nitrogen (N) and phosphorus (P) sources, using BIOLOG PM01, PM02A, PM03B and PM04A MicroPlates [43]. Bacteria were grown overnight on nutrient agar plates (DSMZ medium 1) at 28°C. Biolog experiments were performed according to the modified protocol "PM Procedures for *E. coli* and other GN Bacteria" (Biolog, Inc. 16-Jan-06). Subsequently, for PM1 and PM2A experiments cells were transferred and suspended into 20 ml of Inoculating Fluid IF-0 to achieve 85% T (transmittance) in the BIOLOG Turbidimeter. 240 μl Dye Mix A and 3760 μl H2O were added to a final volume of 24 ml. Each well of PM01 and PM02A MicroPlates (carbon sources) were inoculated with 100 μl of the 85% T cell suspension. PM3B and PM4A experiments require an appropriate carbon source, and a stock solution of 2 M sodium succinate and 200 μM ferric citrate was used as an additive as recommended in the PM procedures for gram-negative bacteria. Initial experiments with 85% T resulted in a strong metabolic response for both, the different substrates but also the negative control (PM3B – A1 [nitrogen]; PM4A – A1 [phosphorus], F1 [sulfur]). Accordingly, the amount of cells was successively reduced in a series of test experiments to a turbidity of 98%, which resulted in sufficient signal strength of the tested substrates combined with a comparably low conversion of the dye in the negative control. For the nitrogen plate PM3B, the optimized inoculation fluid contained 10 ml IF-0, 120 μl Dye Mix A, 60 μl additives and 1820 μl H2O, whereas the inoculation fluid of the phosphorus and sulfur plate PM4A contained 10 ml IF-0, 120 μl Dye Mix A, 120 μl additives and 1760 μl H2O. All PM plates were sealed with parafilm and inoculated in the OmniLog plate reader at 28°C. The conversion of the tetrazolium dye was measured and monitored all 15 minutes at OD590 for four days (96 h). The read-outs were analyzed with MicroLog software applying the automatic threshold option. BIOLOG measures above/below the threshold were considered as positive/negative phenotypes respectively. The reading of plates involving sulfur compounds did not provide reliable results, presumably because the BIOLOG set up does not measure growth per se, but, rather, reflects an integrated view of the redox network of the cells in a particular environment.

***In silico* growth simulations.** FBA was performed using the Cobra Toolbox [376] with MatLab [265] and the gurobi solver [163]. The simulations were performed based on the GSM iJP962 of *P. putida* KT2440 [311]. Each well of the BIOLOG Microplates was simulated by adjusting the *in silico* medium to the available C-, N-, P- and S-sources. Specifically, the *in silico* media contained: bicarbonate, CO2, cobalt, dihydrogen, iron, magnesium, nickel, oxygen, potassium, H+, sodium, water, succinate (not in C-source tests), ammonia (not in N-source tests), phosphate (not in P-source tests), sulfate (not in

S-source tests), and the compound specific to the BIOLOG well (see tables S12 and S14). We discriminated between growth and non-growth phenotypes based on a threshold value of 10^-6 [gdw gdw^{-1} h^{-1}].

***In silico* gene essentiality analysis.** The Cobra Toolbox [376] and the Gurobi solver [163] were used for FBA simulations with Matlab [265] based on the GSM iJP962 of *P. putida* KT2440 [311]. We expanded the original gene essentiality test-set of iJP815 [341] by adding new experimental data including auxotrophies [11]. Each knockout gene was simulated by blocking its associated reactions using the 'deleteModelGenes' function of the COBRA toolbox. The *in silico* media were adjusted to the minimal media described for the experiments in [341] and [11] (see tables S12 and S14). We discriminated between growth and non-growth phenotypes of the mutants based on a threshold value equal to 50% of the wild-type growth rate in the same conditions.

Metabolic network curation process. Positive phenotypes in BIOLOG experiments not supported by metabolic model simulations were manually curated using tools and curation interfaces available in MicroScope [432], in order to find potential catabolic pathways for the corresponding compounds. This includes the analysis of pre-computed results of several computational methods used in the functional annotation process (see above). In addition, genome-context methods available in MicroScope were also used in order to guide functional annotation curation and pathway hole filling: they are based on co-evolution of phylogenetic profiles with functionally related genes [117] and the conservation of genomic and metabolic context through the CANOE strategy used to find candidate genes for orphan enzymatic activities [388]. The outcome of these methods was further improved by extensive manual literature searches to add additional support to functional assignments. Biochemical reactions and Gene-Protein-Reaction associations resulting from the curation process were manually validated in MicroScope using the MetaCyc [66] and the Rhea [286] reaction databases. Rhea was mainly used to manage biochemical reactions that are absent from the current MetaCyc repository. This implies the creation of new reactions directly in the Rhea database, starting with chemical compounds defined in the Chemical Entities of Biological Interest ontology (ChEBI) [168]; reactions are stoichiometrically balanced for mass and charge at pH 7.3 [286]. Similarly, in case of missing compounds in ChEBI with correct 2D structure at pH 7.3, the corresponding compounds were created *de novo* in ChEBI using the Marvin suite of tools from ChemAxon (http://www.chemaxon.com).

Supplementary files

The supplementary files of this work can be online found at `http://onlinelibrary.wiley.com/doi/10.1111/1462-2920.13230/abstract`.

Acknowledgments

We would like to thank Brendan Ryback for his help with adding reactions to the existing GSM, Victoria Michael for her support with BIOLOG experiments, and Kristian Axelsen for reading this manuscript.

Funding

This work was supported by European Union's 7th Framework Programme Microme FP7-KBBE-2007-3-2-08-222886.

Chapter 3

Consensus metabolic model generation

Adapted from:
Ruben G. A. van Heck*, Mathias Ganter*, Vítor A. P. Martins dos Santos†, and Joerg Stelling†. "Efficient Reconstruction of Predictive Consensus Metabolic Network Models". In: PLoS Comput Biol 12(8) 2016.

*,†Equal contributions

Abstract

Understanding cellular function requires accurate, comprehensive representations of metabolism. Genome-Scale, constraint-based Metabolic models (GSMs) provide such representations, but their usability is often hampered by inconsistencies at various levels, in particular for concurrent models. COMMGEN, our tool for COnsensus Metabolic Model GENeration, automatically identifies inconsistencies between concurrent models and semi-automatically resolves them, thereby contributing to consolidate knowledge of metabolic function. Tests of COMMGEN for four organisms showed that automatically generated consensus models were predictive and that they substantially increased coherence of knowledge representation. COMMGEN ought to be particularly useful for complex scenarios in which manual curation does not scale, such as for eukaryotic organisms, microbial communities, and host-pathogen interactions.

Introduction

Genome-Scale constraint-based Metabolic models (GSMs) are curated species-specific knowledge repositories [418]. They integrate many distinct (bio)chemical entities and typically account for thousands of metabolites, reactions and genes. When assuming that metabolism is in a steady state, GSMs also enable metabolic simulations with applications in genome annotation [149, 312], analysis of omics data [69, 87, 473], phenotype predictions [39, 284, 341], organism comparison [17, 27, 284, 311], drug discovery [39, 189, 337], and metabolic engineering [341, 461]. GSMs thereby quantitatively reconstruct the internal metabolic and transport wiring of the modeled organism and thus increase our systems level understanding.

Genome-scale metabolic reconstructions consist of metabolites, metabolic reactions (including boundary reactions and a biomass reaction), cellular compartments, and genes [47, 418]. The reactions are organized according to the cellular compartments in which they are active. Enzyme-driven (as opposed to spontaneous) reactions are associated with Gene-Protein-Reaction rules (GPR), which include one or more genes. For multiple genes, the GPR indicates whether alternative isozymes or enzyme complexes catalyze the reaction [358]. A reaction's equation consists of substrates and products with their corresponding stoichiometries. A reaction's reversibility describes whether the reaction operates forward, backward, or bi-directionally. The reaction flux bounds specify the reaction's capacity, that is, the absolute upper and lower bounds of the reaction flux. Transport reactions transfer metabolites between cellular compartments, whereas boundary reactions define nutrient uptake and secretion. The biomass reaction, finally, reflects the molecular composition of a cell or organism and represents cell or organism growth. Together, these entities and their encoding in a GSM aim to represent the current knowledge of the organism's metabolism.

However, even for well-studied organisms such as *Saccharomyces cerevisiae* or *Bacillus subtilis*, many uncertainties remain during GSM construction. These uncertainties are typically manually addressed based on expert knowledge and scientific literature, which involves a laborious iterative process that can take several years, for example, for eukaryotes [418]. The main sources of uncertainties are: (i) incomplete and erroneous information from heterogeneous and potentially contradictory data sources such as insufficiently curated and inconsistent gene annotations [399], alternative naming and spelling variants of metabolites (different namespaces) [177, 178, 348, 399], and conflicting reaction reversibilities [130, 149]; (ii) subjectivity in interpreting literature sources; (iii) integration of qualitative and quantitative data (e.g., inconsistent growth data); and (iv) incompatible levels of detail between and among

(reference) databases; for example, databases may represent metabolic pathways by detailed individual reactions or by a single lumped reaction [399], and they may use varying structural definitions for metabolite classes such as lipids and polymers [233, 348].

As a consequence, when several GSMs for the same organism are developed independently, they are complementary and only partially overlapping [363, 421]. The extent of variation between models for the same organism can be dramatic. For example, the well-established human and yeast GSMs agree only on 3% [399] and 35% [177] of their reactions, respectively, when ignoring electron, proton, and water imbalances. Differences between GSMs resulting from different modeling frameworks and model authors can even be more substantial than biological differences between organisms [283]. Any GSM-driven analysis, which needs to (somewhat arbitrarily) select one GSM when several are available, thus, only operates on a subset of the available information.

To represent metabolism more comprehensively, and thereby improve our understanding of a target organism, alternative GSMs of a target organism can be integrated into a so-called consensus model of the respective organism, one per organism. Consensus models have an increased scope (by combining unique parts of initial GSMs) and they are more consolidated (by identifying shared parts of initial GSMs that are likely to be reliable). When discrepancies exist between GSMs, these must be carefully examined to select the most appropriate modeling alternative. However, while consensus models have been generated successfully for several (model) organisms such as budding yeast and human, this required extensive manual curation by communities of domain experts [177, 311, 363, 420, 421]. To alleviate this bottleneck and render GSMs truly useful for the understanding of cellular function and evolution, community function, and host-pathogen interactions, semi-automatic consensus model generation approaches have been proposed. It has been shown that the combination of complementary GSMs of the same organism reduces existing gaps in individually reconstructed GSMs [15, 76].These approaches focused mainly on reconciling namespaces (a particularly important challenge for matching metabolites) or on curating the underlying databases [348, 399]. Thereby, existing methods address only a small subset of the problems in consensus model generation described above. For example, they do not identify and curate cases when two initial GSMs represent the same metabolic process at different levels of granularity [400].

Here, we present COMMGEN, a tool for COnsensus Metabolic Model GENeration that reconciles two or more distinct GSMs of the same organism beyond a common namespace. COMMGEN automatically identifies similarities, dissimilarities, and complements of the metabolic networks based on

an extensive classification of problems that typically arise during GSM integration and on novel algorithms to resolve these problem classes. For several model organisms, we show that semi-automatically created consensus GSMs in a standardized namespace [36] are substantially more consolidated than achievable by a common namespace alone, and that they retain or even improve on the initial GSMs' predictive capabilities. Because the consensus GSMs contain the information from each initial GSM, they comprehensively represent our best understanding of the organisms' metabolic networks.

Results

Our analyses addressed model building, testing and refinement in a stepwise fashion. We started by identifying the classes of inconsistencies that exist between models for four widely different albeit representative microbes. We subsequently set up the framework for COnsensus Metabolic Model GENeration, and tested it on the four case studies for functionality and predictability.

Inconsistency classes arising in model merging

To systematically resolve inconsistencies between two or more Initial GSMs (IGSMs) to be integrated, we defined three main (coupled) inconsistency categories: metabolites, reactions, and compartments. We explain these categories and the inconsistency classes they contain using examples from four sets of IGSMs that cover gram-positive and gram-negative bacteria as well as yeast (figure 3.1a).

Metabolites

IGMs often represent a specific chemical compound differently because metabolite identifiers are ambiguous and they reside in different namespaces [36]. When one simply merges IGSMs, that is, adds the IGSMs' contents, this leads to redundant pathways (figure 3.1b) that may differ in metabolites, gene associations, stoichiometries, and reversibilities. The essential step of identifying and merging different metabolites that represent the same chemical compound in different namespaces has been emphasized previously [36, 76, 400]. However, more complicated situations exist when different metabolites actually represent different chemical compounds, but these compounds have the same function in their network context. This typically arises when metabolites are modeled at different granularity, for example, as 'iron' and 'Fe2+', or 'glucose' and 'alpha-D-glucose'. Common metabolites may also have different chemical sum formulas in different IGSMs, for example, depending on whether functional groups are specified or not (figure 3.1c), or when polymers

are modeled with a different numbers of subunits (figure 3.1d). In such cases, the merging of metabolites has to prevent stoichiometric inconsistencies in the consensus model: if a merged polymer can be produced from fewer subunits than result from its degradation, mass conservation is violated. Hence, a common namespace is not sufficient to identify common metabolites in IGSMs.

Reactions

A particular biological process is often represented differently in two models because of uncertainties, disagreements, errors, and modeling decisions, resulting in alternative representations of a single reaction or of reaction sets. These alternatives need to be identified and matched to avoid reaction redundancies (figure 3.1b) and violations of mass balances due to inconsistent stoichiometries (figure 3.1c and 3.1d). However, inconsistencies may extend beyond namespaces and stoichiometries. They often result from modeling decisions, both in capturing individual reactions, and in the granularity of representation for metabolic processes. Nested reactions, where one reaction is a perfect subset of another reaction with respect to metabolites, are possible consequences. In the example in figure 3.1e, the cofactor NADH may be used, but it is not required — for a consensus model, a decision between these alternatives eventually has to be made. Alternative modeling decisions on cofactor usage are common in IGSMs as shown in figure 3.1f with a 'choice' between using NADH and NADPH and in figure 3.1g, where the same chemical conversion can either yield NADP from NADPH or NADPH from NADP. More complex cases to resolve are partially overlapping reactions and lumped reactions, where multiple reactions are artificially represented by fewer reactions. Figure 3.1h shows an example of two alternative reactions that generate triphosphate or pyrophosphate and monophosphate, respectively; simply merging the two IGSMs would feed the side-products into different pathways because no reaction exists that interconverts these metabolites directly. Such inconsistencies are not only found between IGSMs, where they are expected, but also within IGSMs, as demonstrated in figure 3.1i. Hence, it is important to consider the network context of the IGSMs and of the merged GSM.

Compartments

IGSMs of the same organism may consider different subcellular compartments (figure 3.1a), affecting the localization and multiplicity of reactions as well as the incorporated transport reactions. For example, in figure 3.1j, the two IGSMs for a gram-negative bacterium have the same net reaction for the

import of cysteine into the cytoplasm. In one IGSM this requires one reaction because the periplasm is not explicitly modeled, whereas the more detailed transport in the other IGSM requires two reactions. After identifying this class of inconsistencies, a consensus model can either replace the transporter connecting the extracellular space with the cytoplasm by two reactions, or remove the entire periplasm and retain a single transport reaction. Because transporters and transport reactions are notoriously difficult to identify and characterize [418], IGSMs are often inconsistent in transport reactions. Figure 3.1k shows an extreme example: a single merging artifact effectively destroys the model of the proton gradient because protons can be transported across the membrane in either direction by simultaneous import and export of putrescine. Inconsistencies in transport reactions can also lead to thermodynamically infeasible cycles [418] such as ATP generation resulting from cycling glycine over the membrane (figure 3.1l). Finally, boundary reactions, which are not mass-balanced because they exchange material with the environment, are sometimes lumped with transport reactions for the same chemical compound and thus first require standardization (figure 3.1m). Overall, therefore, a broad spectrum of unrelated but interconnected inconsistencies at the metabolite, reaction, and compartment levels need to be identified and resolved for consensus model generation.

The COMMGEN framework

COMMGEN is a software tool that is designed to address the above problems in consensus model generation, leading to a semi-automatic reconciliation of two or more GSMs for a given organism. In terms of software architecture, COMMGEN operates on GSMs in SBML format [183], the standard modeling language for systems biology (figure 3.2a). The IGSMs are first converted into a common chemical naming system using the MNXref namespace [36]. Next, COMMGEN combines all reactions of the IGSMs into a Basic Consensus Model (BCM). The BCM is used to identify and reconcile inconsistencies between and within the IGSMs, ultimately yielding a Refined Consensus Model (RCM) in SBML format. Because many inconsistencies are interconnected, it is difficult to identify a consensus between IGSMs, to distinguish between conflicting and complementary model parts, and to resolve all inconsistencies automatically. COMMGEN therefore resolves all unambiguous cases automatically, and it guides the user to decide on the remaining cases. COMMGEN records all changes such that the user can automatically repeat the procedure with minimal effort, including manual alterations of previously made choices.

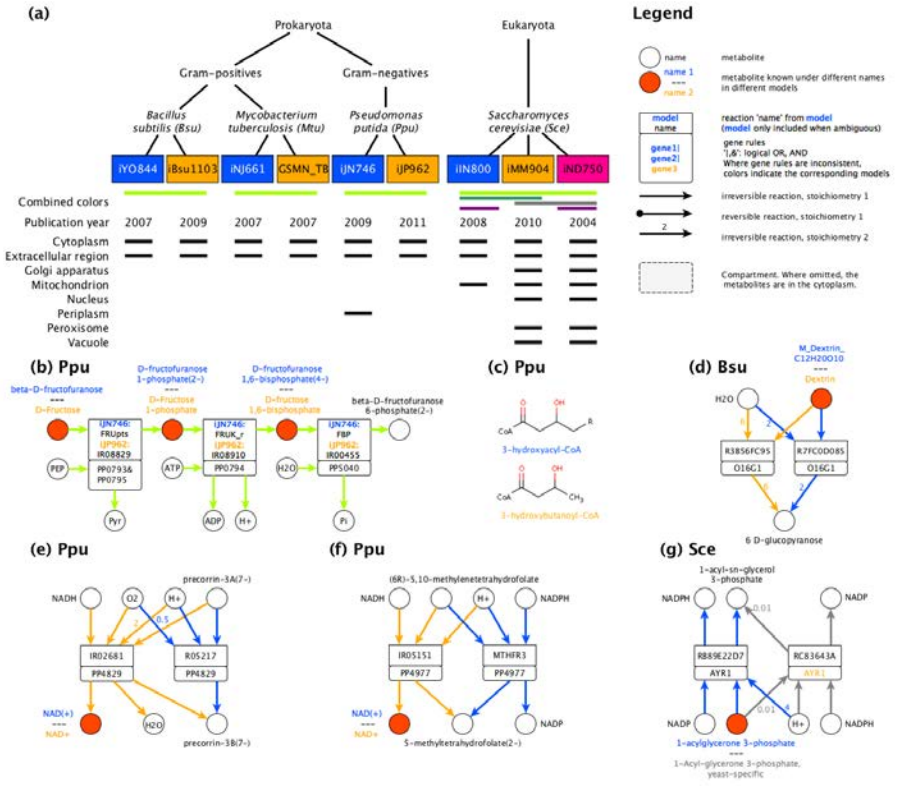


FIGURE 3.1: Models used in this study and classification of inconsistencies. **(a)** Overview of the used initial GSMs. **(b)** Instances of identical metabolites with different MNXref identifiers. **(c)** Non-identical metabolites that perform identical functions in the network context. **(d)** Alternative modeling of polymers. **(e)** Nested and encompassing reactions. **(f)** Alternative usage of redox pairs. **(g)** Alternative reactions with consequences for redox metabolism. Circles represent chemical species, arrows chemical reactions, and grey boxes different compartments. Red nodes indicate instances of identical species within the network context whose alternative names are separated by horizontal lines. Rectangular boxes contain the original reaction names, rounded rectangles their corresponding GPRs, where '&' represents a logical AND, and '|' a logical OR. Continued on next page.

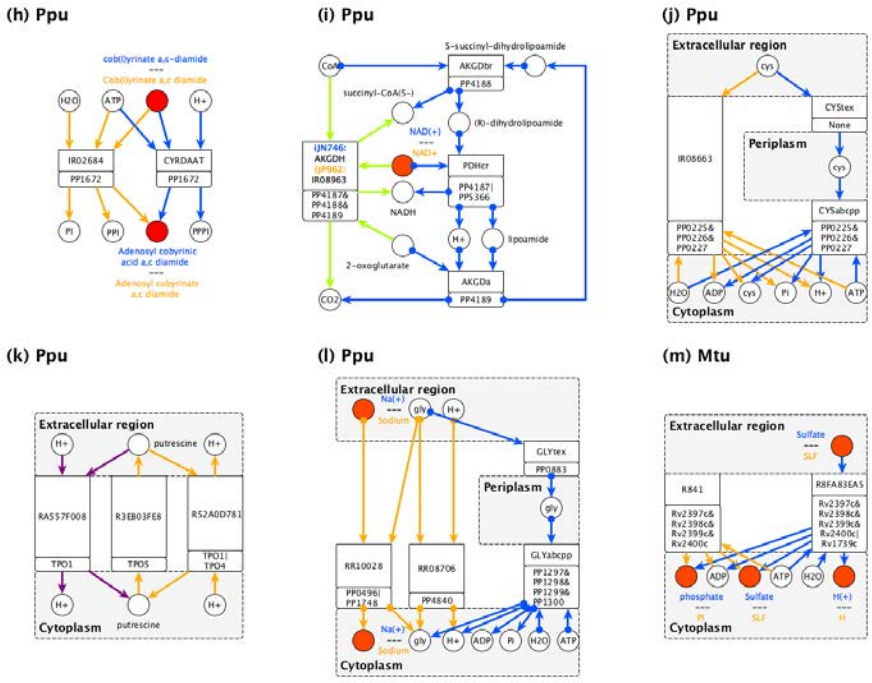
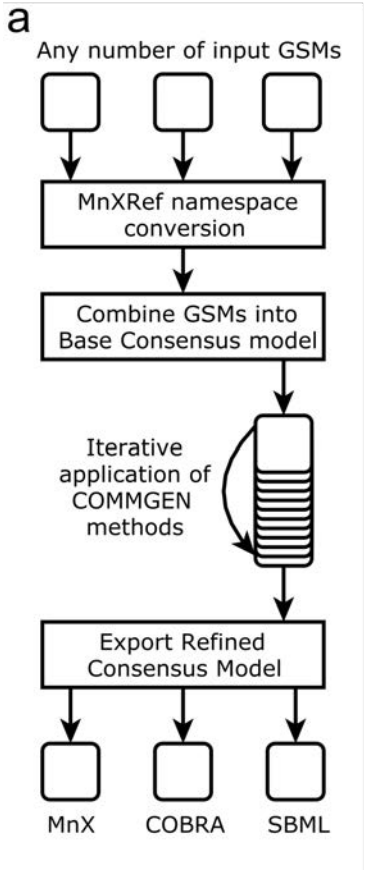


FIGURE 3.1 CONT.: **(h)** Partially overlapping reactions differing in phosphate products. **(i)** Lumped vs. non-lumped representation of a pathway. **(j)** Invalid transport reaction (IR08663). **(k)** Alternative transport reactions for putrescine. **(l)** Alternative transport reactions for glycine. **(m)** Invalid boundary reaction (R841). Edges with filled circles represent reversible reactions. Stoichiometric coefficients unequal to one are indicated at their respective arrows. The shown reactions originate from GSMs of four different organisms: *B. subtilis* (d), as represented in iYO844 [312] (blue) and iBSu1103 [175] (orange); *M. tuberculosis* (m), as represented in iNJ661 [189] (blue) and GSMN_TB [39] (orange); *P. putida* (b,c,e,f,h,I,j,k), as represented in iJN746 [298] (blue) and iJP962 [311] (orange); and *S. cerevisiae* (g,l), as represented in iIN800 [303] (blue) and iMM904 [282, 400] (orange) and iND750 [111] (pink).

To identify and address all the different inconsistency classes described above, COMMGEN iteratively applies a set of independent methods (figure 3.2b). All methods automatically identify instances of their respective inconsistency classes. Metabolite matching is a core element of model merging. We developed a novel algorithm to identify sets of metabolites that represent the same chemical compound based on their network context, that is, their neighboring metabolites and reactions, thereby addressing the issue of different granularity in IGSMs for metabolites (see Methods for details). Performance tests for *P. putida* networks revealed very high sensitivity and specificity of the algorithm, even when only a minority of the network is used to infer matching metabolite sets (figure 3.2c). Metabolite matching allows COMMGEN subsequently to reconcile the associated reactions: metabolites are merged, through which novel pathways and branching points can be formed, and alternative representations of biochemical reactions become apparent. Specifically, COMMGEN matches sets of reactions in the following categories (see Methods for the respective algorithms): (i) reactions with identical metabolites but different stoichiometries; (ii) nested reactions; (iii) reactions that differ only in redox pairs; (iv) partially overlapping reactions; and (v) lumped reactions. Furthermore, it deals with differences in subcellular compartmentalization by (i) facilitating the removal of transporters; (ii) enabling the removal of entire compartments; (iii) resolving differences in the modeling of boundary reactions; (iv) identifying different transport reactions for the same metabolite across the same membrane; and (v) identifying identical biochemical conversions in different compartments.

COMMGEN's methods differ in the extent to which identified inconsistencies can be resolved automatically (figure 3.2b). For some categories, the user can choose to automatically handle inconsistencies, for example, to deal with differences in reaction ality. Conditionally automatic refers to inconsistency classes where some instances can be addressed automatically, but others cannot: if two matched reactions differ only in stoichiometric coefficients, COMMGEN can automatically select the elementally balanced reaction, but only when exactly one reaction is balanced. Manual intervention is always possible, and it is required when inconsistencies are too complex and diverse for a well-performing heuristic for automation. Manual curation is also advisable when an erroneous choice may substantially impact model performance. For example, a single incorrect match between two metabolites with different chemical sum formulas can have severe consequences for the correctness of model predictions. Hence, although the COMMGEN method for network-based metabolite matching performs extremely well (figure 3.2c), we recommend manual confirmation of predicted matches.



b

Category	COMMGEN method	Automatic solving
Metabolites	Metabolite matching 1:1	-
	Metabolite matching 2:2	-
Compartments	Alternative transport reactions	-
	Invalid transport reactions	+
	Alternative compartmentalisation	+/-
	Unknown compartment	+/-
	Invalid boundary reactions	+
	Remove compartment	+
Reactions	Identical net reactions	+/-
	Alternative stoichiometries	+/-
	Lumped reactions	-
	Alternative redox	-
	Nested reactions	+/-
	Partially overlapping reactions	-

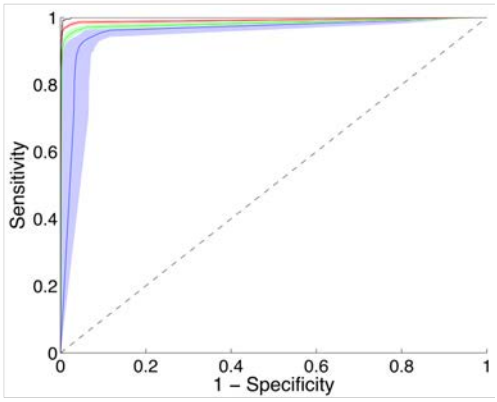


FIGURE 3.2: COMMGEN framework. **(a,b)** Overview of COMMGEN workflow and available methods. The COMMGEN methods are either fully automatic (+), conditionally or optionally automatic (+/-), or they always require manual intervention (-). **(c)** Performance of the metabolite matching methods if run without manual intervention, leading to ROC-curves of the classification of metabolites as identical or non-identical based on their network context. Lines correspond to different fractions of the network information being randomly discarded: black, 0%; red, 30%; green, 60%; blue, 90%. The shades indicate the standard deviations in the classification. The data presented here was obtained using the *Pseudomonas putida* GSMs iJP962 [311] and iJN746 [298]; analysis results for the other sets of GSMs and additional information can be found in S5 Protocol.

Model generation with COMMGEN: Case study for *P. putida*

To describe COMMGEN operation in detail and to evaluate the framework's performance, we focus on consensus model generation for *Pseudomonas putida*, for which the two GSMs iJP962 [311, 341] and iJN746 [298] have been developed independently (figure 3.1a). The initial overlap between these two models is surprisingly low: they only have 58% of their genes, 33% of their metabolites and 2% of their reactions in common. Conversion into the MNXref namespace [36] only increases the common part to 44% for metabolites and 11% for reactions.

To quantitatively determine the occurrences of inconsistencies and their resolution, we classify reactions as consensus reactions (shared between the GSMs) and unique reactions. We further categorize unique reactions according to whether they are unrelated to any inconsistency, related to a single inconsistency, or related to multiple inconsistencies (a reaction may appear in the last category because COMMGEN methods are not mutually exclusive in the inconsistencies they identify). Because the identified inconsistencies ultimately depend on namespace consistency, user-defined settings, and user choices, we quantified the resolution of inconsistencies by automatic processing to remove user bias as much as possible. After creating the BCM from the IGSMs and merging the identical reactions, the fraction of consensus reactions was low (11%) and approximately half of the unique reactions were associated with at least one inconsistency (figure 3.3a; S1 Protocol). The inconsistencies exemplified in figure 3.1 are, thus, not isolated cases; they merely illustrate the main problems in consensus model generation.

Next, we employed a four-step automatic process to reconcile inconsistencies between the IGSMs and to converge to an automatically generated RCM (figure 3.3a). First, COMMGEN increased the namespace consistency through our network context-based metabolite matching method (note that we manually confirmed the proposed matches such that subsequently identified inconsistencies were not overestimated). This increased the overlap to 53% for metabolites and 16% for reactions. In the second step, COMMGEN addressed the difference in cellular compartments in the *P. putida* GSMs (figure 3.1a). In particular, transport reactions from iJP962 that immediately take up metabolites from the extracellular space into the cytoplasm were split such that they match the transport processes from iJN746, and periplasmic instances of the involved metabolites were added. Next, COMMGEN identified and merged sets of reactions with practically (ignoring protons and water) identical net formula. These sets include reactions that have different GPR rules or different reaction directionalities, or that did not have identical net formulas prior to the splitting of transport reactions or the COMMGEN-based metabolite

matching. In this step, we processed inconsistent reaction reversibilities using our previously published method to predict reaction directionalities based on metabolite patterns [149], and we processed inconsistent gene associations by combining the GPR rules with a 'strict' heuristic (see S2 Protocol). Finally, COMMGEN identified and merged reactions that involve the same metabolites, but differ in stoichiometric coefficients; directionality and GPR inconsistencies were handled as above.

The detailed data shown in figure 3.3a emphasize the interdependencies of inconsistencies that may arise in model merging, in particular, that resolving inconsistencies may facilitate subsequent identification of more inconsistencies, resulting in an increased number of identified inconsistent reactions. The four automated steps increased the share of reactions that are consensus reactions originating from both IGSMs from 11% (in the BCM) to 39% (in the RCM), while also substantially reducing the number of reactions associated with inconsistencies (figure 3.3a). We evaluated the significance of the metabolite matching step by re-running the process without it, which lead to only 23% consensus reactions (figure 3.3b). In addition, we used the automatically generated RCM as the starting point for manual curation guided by COMMGEN methods. This allowed us to reconcile most of the remaining inconsistencies and to obtain a consensus for 50% of the reactions (figure 3.3c). In summary, our detailed case study for *P. putida* therefore provides evidence for the efficiency of the COMMGEN framework, and in particular of its novel methods such as network context-based metabolite matching.

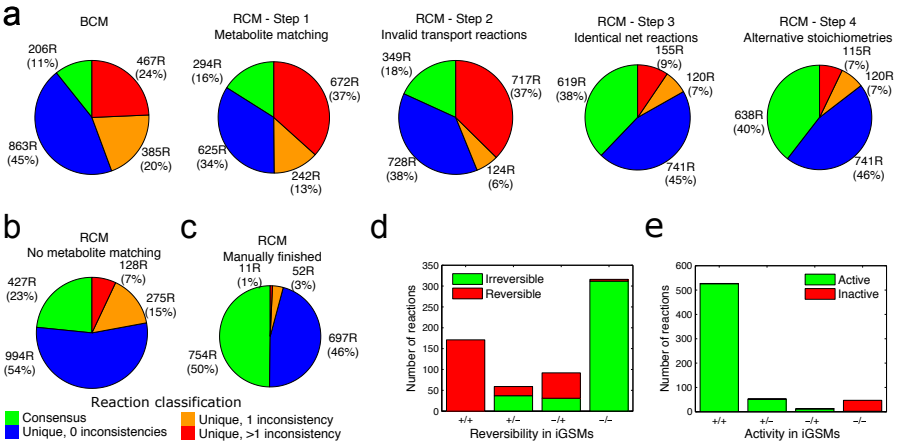


FIGURE 3.3: Application of COMMGEN to *P. putida* GSMs. (a) Automatic inconsistency identification and reconciliation substantially increases consensus and reduces inconsistencies. Reactions are classified into consensus reactions (green) and unique reactions involving no (blue), a single (orange), or multiple (red) inconsistencies. (b, c) Characteristics of the refined consensus model as in (a) without network-based metabolite matching (b), or after manually addressing the remaining inconsistencies (c). (d) Numbers of reversible ('+') and irreversible ('-') reactions in the RCM, grouped by the four possible combinations of reversibilities in the IGSMs. (e) Numbers of active and inactive reactions in the RCM, grouped by being active ('+') or inactive ('-') in the IGSMs.

Automatically generated consensus models are functional and predictive

We next asked, to what extent automated consensus model generation preserved or even extended functionality of the IGSMs, initially focusing on the *P. putida* models. Our automated method involved the probabilistic prediction of reaction directionalities [149] to resolve reaction inconsistencies, instead of simply setting all reactions with conflicting directionalities to reversible, which would tend to overestimate the organism's metabolic capabilities. It maintained reaction directions in case of consensus between the IGSMs, although the prediction method is agnostic to matches between models; it constrained directions in many cases when such constraints existed in only one IGSM (figure 3.3d). The benefits of this approach are best exemplified with a concrete example (figure 3.4a). The *P. putida* BCM contains a small set of reactions that together allow for non-physiological CO2 fixation. This incorrect CO2 fixation cycle was automatically removed when inconsistent directionalities of a reaction present in both IGSMs were processed, thereby preventing

a major error in the RCM. Note that direction prediction also identified a reaction assigned with a direction that is not consistent with the remainder of the network (see also figure 3.1i), namely a directed lumped reaction common to both IGSMs, and a bidirectional non-lumped reaction set present in only one model. Another important aspect of model consolidation is the extent to which active reactions in the IGSMs (that is, reactions that can carry metabolic flux in principle) are preserved. As shown in figure 3.3e, essentially all active reactions in one of the networks remained active in the RCM, and only reactions that were non-functional in both IGSMs remained inactive. In growth phenotype predictions, the RCM occasionally disagreed with all IGSMs, suggesting 'new' metabolic functions. For example, while neither of the IGSMs captured that *P. putida* can grow on L-quinate as sole carbon source, complementation of reactions in the RCM enabled a biologically consistent model behavior (figure 3.4b). These aspects together indicate overall functionality of the automatically generated consensus model.

The performance of GSMs as mathematical models for cellular metabolism is typically evaluated by assessing their ability to correctly predict wild type and mutant growth phenotypes across different growth conditions [235]. We performed corresponding simulations for automatically refined consensus models as well as for their ancestors (IGSMs and BCM) for each of the four evaluated organisms (figure 3.1a). Specifically, we computed sensitivity, specificity, accuracy, and Matthew's Correlation Coefficient (MCC; unlike accuracy it takes the total numbers of true and false test cases into account) [266] for growth phenotype predictions (see S3 Protocol for details). figure 3.5a shows the performance indicators for the IGSMs, the BCMs, and the automatically refined consensus models for each organism. In nearly all metrics, the IGSMs outperformed the BCM (except for *P. putida*), and they were outperformed by the RCM (except for *B. subtilis*). For *B. subtilis*, resolving inconsistencies in the BCM decreased all scores except sensitivity. This can be explained by one IGSM (iBSu1103) being largely based on a predecessor (iYO844); in addition, iBSu1103 was optimized for correct growth predictions using GrowMatch [175, 235]. Information from iYO844 can thus include errors that were deliberately removed from iBSu1103 and it can reverse changes made by the performance optimization. Thus, although the prediction profiles of the RCMs largely resemble the IGSM profiles, RCMs on average outperform both the IGSMs and the BCMs, indicating efficiency of the automated consensus model generation methods in COMMGEN even in terms of prediction capabilities. Notably, user choices of the biomass reaction do not influence the performance substantially (figure 3.5a), pointing to robustness of the methods as well.

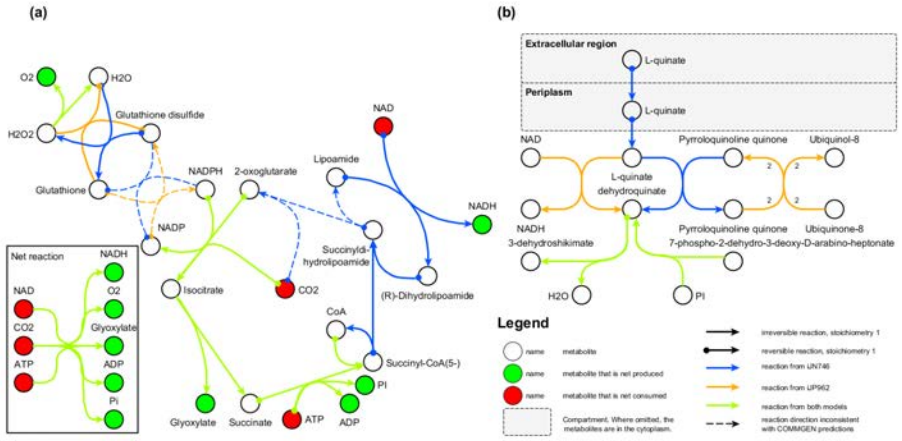


FIGURE 3.4: Subnetwork analysis for *P. putida*. **(a)** Example error of 'naïve' iGSM merging where the initial *P. putida* BCM contains a biologically inaccurate carbon dioxide fixation cycle due to incorrect directionalities in the IGSMs. This error is automatically resolved as COMMGEN assigns reaction directionalities opposite to those shown with dashed reaction arrows. **(b)** Example for a new metabolic function in the consensus model. *P. putida* can grow on L-quinate as its sole carbon source. Neither of the initial models captures this behavior, whereas the consensus model provides the necessary, complementary reactions.

Automatic reconciliation is comparable to manual consensus model generation

Finally, we wanted to evaluate how automatic consensus model generation compares to its (largely) manual counterpart. We focused on the community approach to establish a yeast consensus model [177] based on the IGSMs iMM904 [282] and iLL672 [232] because this first model reconciliation effort is especially well documented. Figure 3.5b shows that transfer of the IGSMs into a standardized namespace alone identifies only small subsets of common metabolites and reactions. COMMGEN's automated reconciliation method, in contrast, achieves nearly the same extent of matching between the IGSMs as reported for the manual curation. The automatically generated RCM showed good performance in mutant phenotype predictions (sensitivity = 0.98, specificity = 0.28, accuracy = 0.87 and MCC = 0.42; note that a comparison to the manual consensus model is impossible because the community effort did not aim at establishing a model suitable for FBA). In addition, COMMGEN directly identifies many inconsistencies between model reactions that result, for example, from different numbers of compartments in the IGSMs (figure

3.5c). These would be clear starting points for domain experts for subsequent COMMGEN-assisted manual curation. We believe that the combination of automated procedures with close-to-manual quality and of support for targeted manual curations would substantially enhance future community efforts.

Discussion

Genome-scale constraint-based metabolic models are both integrated knowledge repositories and predictive mathematical models. In terms of knowledge representation, a consensus model should be more consolidated than individual GSMs due to shared parts, more comprehensive due to unique parts, and more accurate due to reconciliation of inconsistencies in similar parts. A consensus model, however, can propagate errors in the initial models' unique parts, and it may be less consistent than the initial models, especially when inconsistencies in similar model parts were not identified or reconciled.

Inconsistencies in GSMs are typically nested, not mutually exclusive, and therefore difficult to address, which so far prevented the development of methods for the automated generation of consensus models [400]. Manual network reconciliation, the predominant approach applied today, is difficult and cumbersome because the number of inconsistencies between just two or three IGSMs already runs in the thousands. Based on a systematic classification of inconsistencies, COMMGEN automatically identifies and semi-automatically reconciles inconsistencies between and within two or more IGSMs. The inconsistencies could theoretically be reconciled fully automatically, but automated resolution depends on the used reference databases, which vary to a large extent [399]. Therefore, COMMGEN does not entirely remove the need for manual inspection and curation. For example, our framework relies on network similarity between alternative realizations of metabolites and reactions in order to match them. Because the reactions surrounding biomass formation are often implemented very differently in different GSMs, they are not matched. While our implementation lets the user choose one of the IGSM biomass reactions, a manual update seems necessary as long as COMMGEN does not automatically fetch external information that would enable an automatic reconciliation of the biomass reaction. In addition, there exists a trade-off between sensitivity and specificity for the identification of inconsistent reactions, which limits the detection of lumped and non-lumped pathway representations with a different net reaction. Also, the identification of similar or identical reactions in different cellular compartments is difficult to achieve automatically (but an extension of the current framework could progress in this direction by combining the information from metabolite instances in different compartments prior to metabolite matching). COMMGEN

thus forms a necessary bridge between full automation and high-quality manual curation for consensus metabolic model generation.

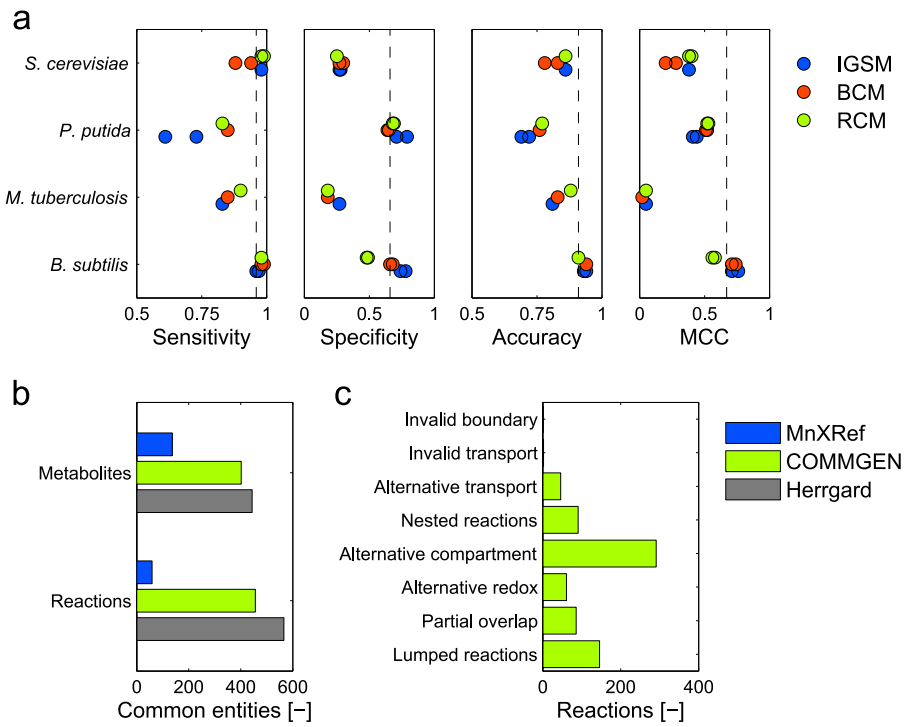


FIGURE 3.5: Performance evaluation of COMMGEN. **(a)** Evaluation of GSM ability to predict growth phenotypes. Predictive ability of initial GSMs (blue), basic consensus models (red), and automatically created refined consensus model (green) according to the metrics defined in the text. The test data comprised gene knockout data (*B. subtilis* [175, 312], *P. putida* [11, 341], *M. tuberculosis* [374], *S. cerevisiae* [111]), biolog data (*B. subtilis* [175, 312], *P. putida* [298, 341]) and auxotrophies (*P. putida* [11]). See S3 Protocol for details. **(b,c)** Comparison of manual yeast consensus model [177] based on the IGSMs iMM904 [282] and iLL672 [232]) with automatic consensus model generation with namespace matching only, or with COMMGEN. **(b)** Numbers of common reactions and metabolites for manual curation, namespace conversion, and automatically created refined consensus model. **(c)** Number of incidences of inconsistent reaction classes identified by COMMGEN.

Regarding a GSM's predictive mathematical model character, it is important to note that remaining inconsistencies in a consensus model can have severe effects, for example, when inconsistencies resulting from model merging are not adequately addressed. As a consequence, individual GSMs may outperform a consensus model in terms of predictive ability even though the latter is more representative of the available information. COMMGEN's aim (and design) is to compare and reconcile IGSMs in order to obtain a high-quality representation of the IGSMs' combined information. In contrast to model optimization methods such as GrowMatch [235], COMMGEN does not create a model optimized for predictive ability, and it does not use corresponding experimental information. However, our example applications also demonstrated that automatically generated consensus models almost always have higher predictive power than the manually curated IGSMs and that these models can be comparable to manually constructed consensus models as shown for yeast. COMMGEN increases coherence with the actual biological system while maintaining predictive power. This balance is of utmost importance for the usability and reliability of GSMs to elucidate cell function interactions.

As demonstrated by our case study for *P. putida*, we argue that (semi-)automatically generated consensus models provide the basis for additional improvements due to their comprehensiveness and standardized naming system. Gap-filling methods [149, 419] may be able to close gaps that are not apparent in the IGSMs. One can use existing methods [149, 304] to re-evaluate reaction directionalities, especially for reactions that differed in the IGSMs. Compartment assignment methods [280] can resolve remaining compartmentalization issues and optimization methods [109, 235] may alter the model to increase its predictive ability. Finally, a good consensus model is a solid foundation for new models by providing a basis for GSMs of similar organisms, and via its integration into multi-scale whole-cell or tissue models [175].

More generally, the systematic integration of heterogeneous information is an essentially unsolved challenge in (post-)genomic biology. For metabolism, consensus GSMs are formalized means for complementing incomplete information, and for identifying and addressing errors through the comparison of independently generated GSMs for the same organism. COMMGEN automatically identifies and semi-automatically resolves widespread and highly interlinked inconsistencies between initial GSMs, thereby moving beyond existing approaches for manual and computer-aided consensus model generation. It can therefore facilitate the construction of new models by comparing and combining information from automatic model construction tools such as the Model SEED [176] and manual model construction efforts, and facilitate GSM updates using a reference — both tasks are analogous to consensus GSM generation.

While we focus here on the reconciliation of multiple GSMs for the same species, we argue that COMMGEN's methods and standardization are more widely applicable. The identification of similar, yet distinct, biochemical entities can help to compare metabolic capabilities of organisms in detail via their GSMs, or even to compare entire pathway databases. However, dealing with different species will require new, systematic preprocessing steps to map gene sets in different organisms functionally to each other (e.g., via orthology or enzyme classification numbers), which is a topic of future research. In addition, COMMGEN's methods for identifying redundancies and hierarchical relationships in networks can be used to further advance standardization of terms and ontologies. We therefore expect COMMGEN to be of substantial aid in future integration of knowledge for metabolic networks, to greatly accelerate model-building processes and to thereby improve subsequent high-throughput model-based network analyses. Although COMMGEN will not directly address the domain-specific problems, these capabilities will lay a solid foundation for the systematic, genome-scale comparison of metabolic spaces within and across genera and will have substantial impact for large-scale evolutionary analyses, design of microbial communities, and understanding of host-microbe (pathogen, microbiome) interactions.

Materials and Methods

Genome-scale metabolic models. iJN746 and iJP962 were requested from and received by email from the first authors of the corresponding papers. GSMN-TB was downloaded from http://sysbio3.fhms.surrey.ac.uk/. iNJ661 was obtained from the supplementary files of the corresponding paper. The remaining models were taken from the model repository at www.metanetx.org. See S1 Dataset for details.

Evaluation of model performance. For comparison to experimental data, the models were loaded into the COBRA toolbox [376]. The bounds of the boundary reactions were adjusted based on the medium composition and, where necessary, additional flexibility was provided to individual models. Gene knockout strains were simulated by removing the reactions requiring the encoded protein. To discriminate growth from no growth for wild type strains a default cut-off value (10^6) was used whereas a minimal relative growth rate (30%) to the wild type was used for mutant strains. See S3 Protocol for details.

Matching metabolites based on network context. In a metabolic network, reaction nodes are only connected to the metabolite and gene nodes that are involved in the corresponding reaction. Similarly, metabolite and gene nodes are only connected to reaction nodes. However, reaction nodes are not informative for the identity of metabolites as two metabolites representing the same chemical compound are non-overlapping in their connected reaction nodes. Therefore, we characterize metabolites by the other metabolite and gene nodes that are connected to the same reactions. We use this information to quantify how similar metabolites from different models are based on their network context. These similarity scores are then compared to the scores of metabolites that are known to match because they are present in both models: pairs of metabolites that score comparable to these shared metabolites may consist of functionally equivalent chemical compounds. We use a user-defined percentile of shared metabolite scores as a threshold to identify similar metabolites. The method is described in the following:

i We create a Boolean metabolite-to-metabolite matrix M_m (m x m) where a 1 indicates that the two metabolites share a reaction.

ii We create a Boolean gene-to-metabolite matrix M_g (g x m) where a 1 indicates that the metabolite and gene share a reaction.

iii We create an attribute matrix M_a ((m + g) x m) by vertically concatenating M_m and M_g.

iv We normalize M_a by dividing each row by its sum such that the numbers in each row sum up to 1. Thereby, the values in M_a reflect both that a metabolite is connected to a metabolite or gene and how rare (defining) this connection is.

v We discard rows from M_a that correspond to metabolites and genes that are not included in both models for these cannot aid in the identification of common metabolites between the models.

vi We discard the columns from M_a that correspond to metabolites that are identified to be the same in both GSMs.

vii We create a scoring matrix M_s (m x m) where the number at position i,j corresponds to the Pearson's correlation coefficient between columns i and j of M_a.

viii We distinguish between similar and non-similar metabolites in M_s using a minimal score. The minimal score equals a user-defined percentile of scores for metabolites that are present in both models.

Identification of lumped reactions A lumped reaction is an artificial reaction that represents the net effect of multiple individual reactions. Therefore, if the lumped and non-lumped representations carry flux in opposite directions, steady state is maintained as they cancel each other out. We use this property to identify lumped reactions by linear programming. The method is described in the following:

i We determine the directionality for each reaction as forward, backward, or reversible.

ii We transform each reaction such that it only runs in the forward direction; backward reactions are reversed and reversible reactions are split into two reactions.

iii We update the stoichiometric matrix S (m x r) accordingly.

iv We remove the boundary reactions from S as these reflect exchanges of metabolites between the organism and the medium.

v We define the linear programming (LP) problem:

$$\begin{array}{ll} \text{Maximize} & \mathbf{c'x} \\ \text{Subject to} & S_{irr}\mathbf{x} = \mathbf{b} \\ & \mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub} \end{array}$$

vi We initiate the variables of the LP problem
c: Vector (1 x r) containing the objective coefficient for each reaction. We set each value to -1 to penalize flux through each reaction; this ensures that the total flux in the network is minimized.
lb: Vector (1 x r) containing the lower bounds of each reaction. As all reactions are forward reactions, every value is set to 0.
ub: Vector (1 x r) containing the upper bounds of each reaction. As all reactions are forward reactions, every value is set to 1000.
b: Vector (m x 1) containing the desired accumulation or dissipation of each metabolite. Each value in this vector is set to 0 to ensure a steady-state flux distribution.

vii We select a reaction LR with index i_{LR} to be considered as a lumped reaction. We set: $\mathbf{c}(i_{LR}) = -1000$ $\mathbf{lb}(i_{LR}) = -1$ LR is thus now allowed to carry flux in the backward direction, which results in a positive contribution to the objective value.

viii We run the LP problem as defined under step v. The LP problem returns a flux distribution x that either only contains zeros (no non-lumped representation available), or contains a flux distribution such that the flux through LR is maximized in the reverse direction while having a minimal flux through the rest of the network. In the first case, we skip steps ix and x. In the latter case, we identified a set NL_1 of corresponding non-lumped reactions.

ix We save the set NL_1 for future reference.

x We modify the LP problem such that any alternative sets NL_x may be identified. $c(NL_1) = 3c(NL_1)$ This effectively further penalizes flux through the reactions of NL1 such that it becomes more 'rewarding' to use other reactions.

xi We repeat steps viii-x (replace NL_1 by NL_2, NL_3, ...) until: No non-zero solution to the problem exists, or The number of reactions in NL_x exceeds a user-defined threshold (default: 5), or There is a recurring set NL_x.

xii We filter the different sets NL_x such that only sets remain that overlap to a pre-defined extent in gene associations with LR.

xiii We repeat steps v-xi such that we obtain sets NL for each reaction in the model.

Identification of alternative transport. Alternative transport reactions result in the transport of a metabolite between two compartments with a different net reaction. We identify metabolites with alternative transport reactions one metabolite at a time. If a metabolite is present in two or more compartments, we identify all transport reactions for this metabolite by selecting reactions where the metabolite is on both sides of the equation. If two of these reactions transport the metabolite between the same two compartments, these reactions are alternative transport reactions.

Identification of invalid transport. Invalid transport reactions are reactions that transport metabolites between two unconnected compartments. We identify these by forming a list of all compartments that are directly connected through transport reactions in the IGSMs and asking the user to indicate if any of these are invalid. For any of the invalid compartment connections, we identify reactions that contain metabolites from both compartments; these reactions are invalid transport reactions.

Identification of alternative compartmentalization. We create a separate stoichiometric matrix S_{cmp} (m x r) for each compartment. These matrices only contain reactions of which all metabolites are in the same compartment. Columns (reactions) that are identical between these matrices represent identical reactions with an alternative compartmentalization.

Identification of unknown compartment. In the MNXref namespace, metabolites with an unclear compartmentalization are placed in the compartment UNK_COMP. For each reaction that contains a metabolite in UNK_COMP, we identify reactions from the other IGSM(s) that involve all metabolites with known compartmentalization similarly to the identification of alternative stoichiometries. These reactions are then filtered for reactions that also involve the metabolite with the unknown compartmentalization.

Identification of invalid boundary reactions. Boundary (exchange) reactions are artificial reactions that represent the exchange of metabolites with the medium. They only involve a single metabolite, and have no metabolites on the other side of the equation. In some models these reactions are lumped together with transport reactions that import metabolites from the extracellular compartment. After the MNXref namespace conversion these reactions are still annotated as boundary reactions, and are thus easily identified in COMMGEN by searching for boundary reactions with non-extracellular metabolites.

Removing a compartment. To combine GSMs with an alternative compartmentalization, it is sometimes most straightforward to remove a compartment 'RC' from a GSM and move its reactions to a different target compartment 'TC'. We defined four categories of reactions in RC, which are treated differently when RC is removed: (i) Reactions that only involve metabolites from RC are moved to TC; (ii) Multi-compartment reactions that transport a metabolite between RC and TC are removed; (iii) Multi-compartment reactions involving RC and TC that involve a chemical conversion are kept, but all metabolites from RC are placed in TC; (iv) Multi-compartment reactions involving RC and a metabolite other than TC are kept, and all metabolites from RC are placed in TC.

Identification of identical net reactions. Identical net reactions are reactions that involve the same set of metabolites in the same stoichiometries, but they may be defined in opposing directions. Therefore, we create a double stoichiometric matrix S_{dbl} (m x 2r) that contains the normal stoichiometric matrix S (m x r), as well as its negative -S (m x r). We then identify columns (reactions) in S_{dbl} that are identical.

Identification of alternative stoichiometries. We convert the S (m x r) matrix to a Boolean (0/1) representation S_{log} (m x r). We then identify columns in S_{log} that are identical; these correspond to reactions involving the same metabolites, but in different stoichiometries.

Identification of alternative redox pairs. GSMs often differ in their involvement of redox pairs in any particular reaction. The first step in identifying these inconsistencies is the creation of a list of redox pairs. COMMGEN comes with a list of commonly used redox pairs in the MNXref namespace, and this list can be expanded by the user. COMMGEN can suggest expansions for this list by selecting metabolite pairs that co-occur frequently ($\geq$ 80% of reactions). We identify reactions that are identical except for their redox pairs by expanding the stoichiometric matrix S (m x r) to S_{rdx} (m+1 x r) by adding an artificial metabolite 'redox pair'. Then, for each reaction that involves a redox pair, we put the stoichiometric coefficients of the redox metabolites in S_{rdx} to '0', and add a '1' in the 'redox pair' row instead. We then use the same approach as for the identification of alternative stoichiometries to identify reactions that only differ in stoichiometries and redox pairs.

Identification of nested reactions. We convert the S (m x r) matrix to a Boolean (0/1) representation S_{log} (m x r). For each column (reaction) we then identify other columns that contain nonzero elements on each row where the respective column has a nonzero element. These sets of columns (reactions) are potentially nested reactions. We then confirm these sets by detecting sets where two or more metabolites that are on the same side of the equation for one reaction, are on the same side of the equation for the other reaction.

Identification of similar reactions. Similar reactions are reactions from different IGSMs that share a predefined number of genes, substrates and products. We identify similar reactions by constructing three sets of pairs of reactions: (i) reactions that originate from different IGSMs, (ii) reactions that share the required number of substrates and products, and (iii) reactions that share the required number of genes. All combinations of two reactions in each of these three sets are considered similar reactions.

Implementation and simulation. All computational simulations and analyses were performed using MATLAB [265]. Gurobi [163] was used as linear programming solver for flux balance analysis.

Namespace conversion. COMMGEN uploads SBML files to MetaNetX.org [150], where the namespace conversion into MNXref [36] is performed, and downloads the resulting model. Because errors may be introduced at this stage (incorrect namespace conversion of individual metabolites) the mapping is presented to the user who can reject incorrect matches. See S4 Protocol for details.

File formats and accessibility. The COMMGEN version used for this paper is freely available as MATLAB code as S6 Protocol. A current version of COMMGEN can be found at https://gitlab.com/Rubenvanheck/COMMGEN.

Supplementary files

The supplementary files of this work can be found online at `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005085#sec031`.

Acknowledgements

We thank Sumana Srivatsa for testing COMMGEN as well as Marco Pagni, Thomas Bernard and Sebastien Moretti for support with MetaNetX and MNXre f.

Author contributions

Conceived and designed the experiments: MG JS VAPMdS.
Performed the experiments: RGAvH MG.
Analyzed the data: RGAvH MG JS VAPMdS.
Contributed reagents/materials/analysis tools: RGAvH MG.
Wrote the paper: RGAvH MG JS VAPMdS.

Funding

We gratefully acknowledge financial support from the Swiss Initiative for Systems Biology (SystemsX.ch, project MetaNetX) reviewed by the Swiss National Science Foundation (SNF), the Wageningen university IPOP project, and the European projects INFECT (Project reference: 305340) and EmPowerPutida (Project reference: 635536). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4

Modeling essentials: The *Pseudomonas* genus

Adapted from:
Jasper J. Koehorst*, Jesse C. J. van Dam*, **Ruben G. A. van Heck**, Edoardo Saccenti, Vitor A. P. Martins dos Santos, Maria Suarez-Diez, and Peter J. Schaap. "Comparison of 432 *Pseudomonas* strains through integration of genomic, functional, metabolic and expression data". In: Scientific Reports 6 2016.

*Equal contributions

Abstract

Pseudomonas is a highly versatile genus containing species that can be harmful to humans and plants while others are widely used for bioengineering and bioremediation. We analyzed 432 sequenced *Pseudomonas* strains by integrating results from a large scale functional comparison using protein domains with data from six metabolic models, nearly a thousand transcriptome measurements and four large scale transposon mutagenesis experiments. Through heterogeneous data integration we linked gene essentiality, persistence and expression variability. The pan-genome of *Pseudomonas* is closed indicating a limited role of horizontal gene transfer in the evolutionary history of this genus. A large fraction of essential genes are highly persistent, still non essential genes represent a considerable fraction of the core-genome. Our results emphasize the power of integrating large scale comparative functional genomics with heterogeneous data for exploring bacterial diversity and versatility.

Introduction

The *Pseudomonas* genus exhibits a broad spectrum of traits and *Pseudomonas* species show a remarkable adaptability to the biochemical nature of the large variety of environments, often extreme, they thrive in [424, 453]. The genus currently includes almost 200 recognized species, which have been clustered into seven groups and into lineages on the basis of a limited set of loci [253]. Some species are well-studied because they are human or plant pathogens, like *P. aeruginosa* or *P. syringae*, or because they are considered harmless and possess interesting biodegradation properties while others can produce a variety of extraordinary secondary metabolites with anti-microbial properties [159]. *P. putida* KT2440 is even Generally Recognized as Safe (GRAS-certified) for expression of heterologous genes and has been transformed into a genetically accessible laboratory and industrial workhorse [290].

A number of comparative genomics studies have been performed in the past [21, 253, 453] but the number of available *Pseudomonas* genomes quadrupled in the last five years due to the widespread use and the advancement of high-throughput sequencing technologies. As of December 2015, the complete and draft genomes of 432 strains distributed over 33 species are publicly available (see Supplementary figure S1). This plethora of data entitles an in-depth comparative re-analysis of *Pseudomonas* genomes to explore their metabolic and ecological diversity.

Large scale functional comparison based on sequence similarity is challenged by methodological problems, such as the need of of defining arbitrarily generalized minimal alignment length and similarity cut-off for all sequence to be analyzed, and it is hampered by the high computational cost, since time and memory requirements scale quadratically with the number of genome sequences to be compared[225]. Many bacterial proteins consist of two or more domains and fusion/fission events are the major drivers of modular evolution of multi-domain bacterial proteins [329]. Interspecies domain variation can thus give rise to an annotation transfer problem: sequence based functional annotation methods use a consecutive alignment to identify common ancestry and therefore may miss domain insertion/deletion, exchange or repetition events, which may lead to functional shifts and promiscuity. Comparisons at protein sequence level should therefore be complemented with comparisons at the protein domain level [225]. In addition, in order to avoid technical biases a biologically meaningful functional comparison requires consistent and up-to-date annotations. Instead, the biological information available in public databases varies in quality due to the use of different databases and annotation pipelines that include different methods and may assign different names, acronyms and aliases to the same protein. Re-interpretation of these predictions in most cases requires reverse engineering as data provenance is usually

not available.

In this paper 432 *Pseudomonas* genome sequences were *de novo* reannotated and the generated annotation information was integrated through a semantic platform with data from six metabolic models, nearly a thousand transcriptome measurements and four large scale transposon mutagenesis experiments. We identified phylogenetic relationships among different species using protein domains and performed extensive analysis of the core- and pan-genomes of the *Pseudomonas* genus and considered the habitat factor while analyzing the pan/core-genome. Finally, we linked domain content and domain variability of persistent and essential genes and their transcriptional regulation.

Results

De novo annotation of *P. putida* KT2440 as a minimal working example

P. putida KT2440 [290] is one of the best-characterized *Pseudomonas* strains. A *de novo* annotation obtained using an in-house annotation pipeline, the annotation deposited in GenBank (NC_002947) and an alternative annotation obtained using RAST [16] were compared, see Table 4.1. The total number of genes identified using three gene calling methods, Prodigal 2.6 (in our pipeline), Glimmer3 (RAST), and Glimmer (GenBank) are very similar, differing less than 4%. However, as each of these algorithms have an intrinsic false discovery rate in start-site prediction, significant differences in the start position of the identified genes were found. The number of exact matches in gene start-sites is only 73% (4073 genes) confirming previous observations [429]. These 5' variations in gene identification can result in a putative gain or loss of biological functions; however, since different naming conventions are used in the different annotation protocols applied, a direct functional comparison to spot possible differences is not possible (figure 4.1).

TABLE 4.1: Annotation results for *P. putida KT2440*. GenBank refers to the original deposited annotation (available at NCBI), whereas RAST and SAPP refer respectively to their annotation.

	#Genes	#Unique start/ end positions	#Unique GO	#Unique domains	#Unique EC
GenBank	5350	170	0	3574	443
RAST	5531	62	726	3631	447
SAPP	5555	252	1403	3636	447

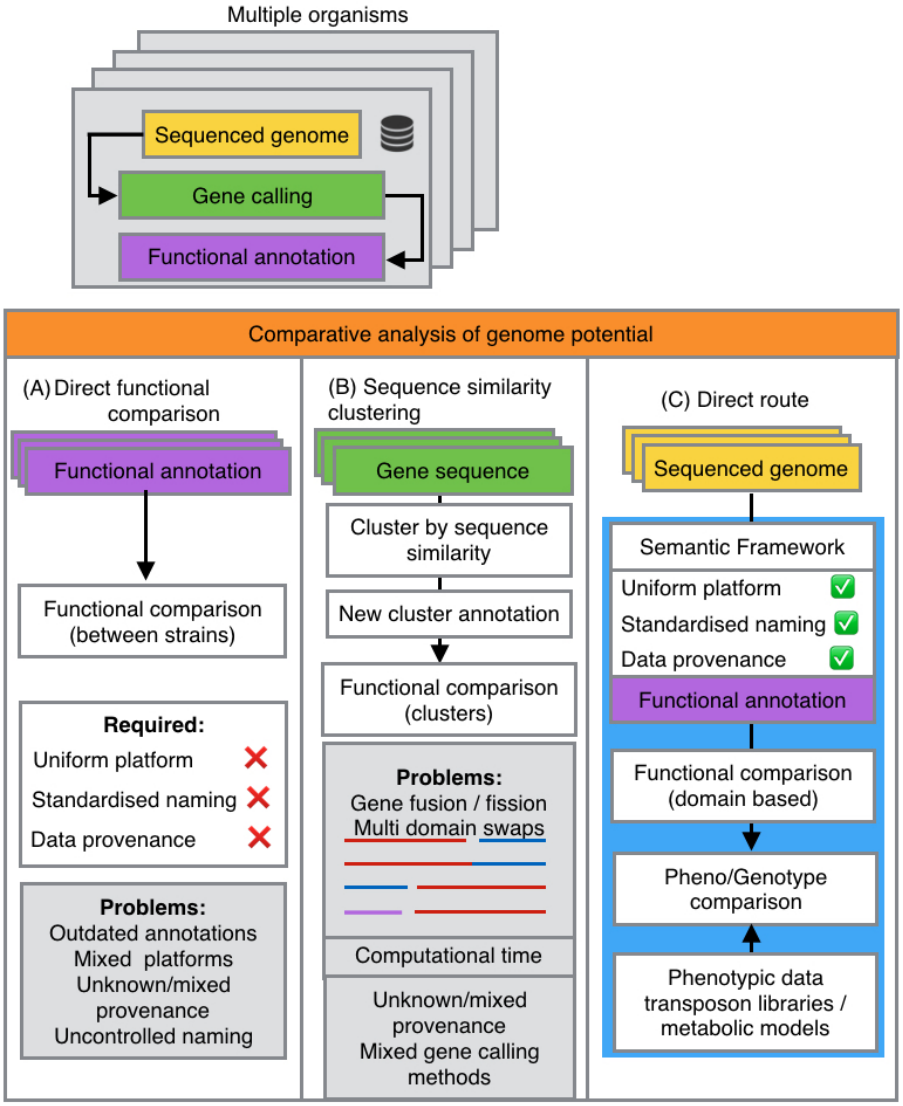


FIGURE 4.1: Alternatives for functional genome comparison: **(A)** Direct comparison of genome potential using existing annotation is often hampered by lack of standardization of gene calling and annotation tools, mixed and unknown data provenance and inconsistent naming of function. **(B)** Sequence similarity clustering bypasses inconsistent functional annotations. Computational time scales quadratically with the number of genome sequences and gene fusion/fission events might be overlooked. **(C)** Usage of standardized annotation tools ensures uniform genome annotation prior to comparison; annotation provenance is stored for all steps.

The use of controlled vocabularies overcomes this issue, so that functional comparison can be performed using gene ontology (GO) terms, Enzyme Commission (EC) numbers and InterPro identifiers. For the GenBank deposited annotation no GO information was available but the difference observed between the RAST and the *de novo* annotation is striking. This minimal working example shows that even for a single genome a comparative analysis of functional annotations derived from three work-flows is almost impossible by computational means due to lack of standardization and data provenance. This example further emphasizes that comparative genomic analysis requires homogeneous annotation.

Comparison of the genomic potential of *Pseudomonas* species

Since for a comparative genomics study a consistent and standardized genome annotation is a prerequisite, we evaluated the impact by comparing the functional annotations of 432 *Pseudomonas* genomes with a *de novo* annotation. We used both complete and draft genomes. According to the quality metric defined by Cook and Ussery, almost 30% of the available draft genomes were of low quality [90]. This was mostly due to a high number of contigs and not to the quality of the assemblies in itself, so they were included in the analysis.

GenBank files were converted into RDF, extracting genome sequences and gene-calls. Genomes were structurally and functionally reannotated. The originally deposited gene-calls were functionally reannotated as well and a pairwise comparison of GO terms, and EC identifiers assigned to the originally deposited and the *de novo* gene-calls was performed at gene and protein domain level. Figure 4.2 summarizes the results for the available 58 complete genomes. Differences in annotations were observed at all functional levels. Per genome on average 38 new genes were predicted while a functional reannotation of the set of complete genomes yielded 838 additional GO-terms and 146 additional domains (For a more detailed overview see Supplementary data S2). Considering the full set of 432 genomes, on average a difference of 153 genes per genome was detected. The results advocate for routine implementation of consistent gene-calling methods combined with an up-to-date functional annotation before performing comparative genomic analyses, as many of these differences will results in gain or loss of biological functions.

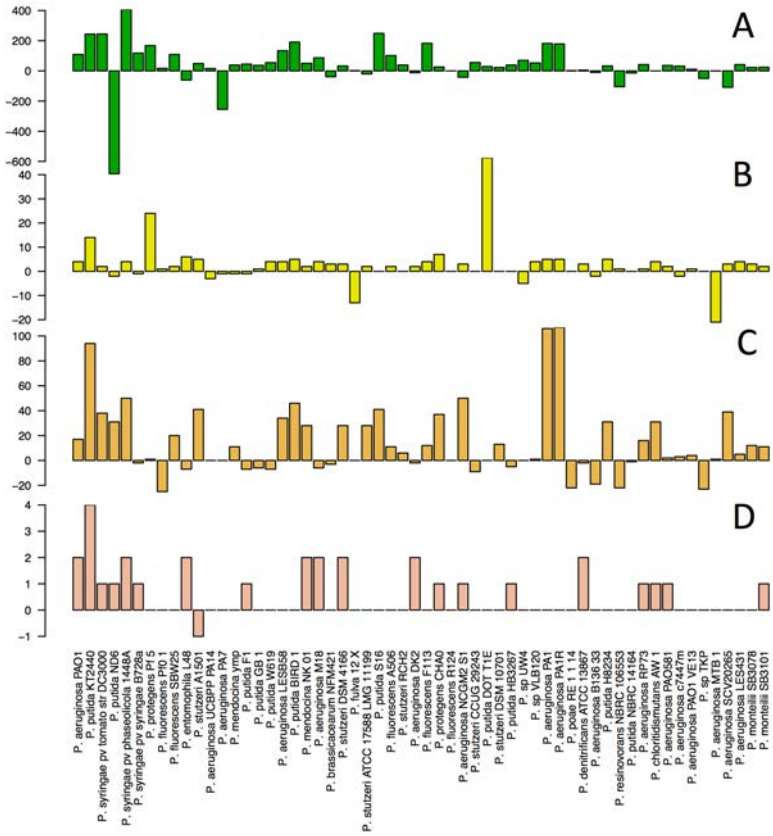


FIGURE 4.2: *De novo* annotation of *Pseudomonas* genomes. Comparison between the original and de novo annotations of 58 completely sequenced *Pseudomonas* genome sequences. Barplots indicate differences in the number of retrieved genome features terms between the *de novo* annotations and the original deposited annotations. A: gene abundance; B: protein domains; C: GO terms, and D: EC identifiers. The genomes are ordered from left to right by deposition date in the NCBI database (from oldest to newest).

Sequence and function based comparative genomics of *Pseudomonas*

Genome-wide comparative analysis usually relies on sequence similarity clustering based on a blast-based all-against-all bidirectional best hit (BBH) heuristic approach. There are several limitations to this approach. Firstly, the runtime increases quadratically with the number and complexity of the species involved. Secondly, clustering is strongly context-dependent as it dramatically depends on chosen cut-off values to define statistical significance of sequence similarity. Problems may arise with in-paralogous sequences that evolve at very similar rates resulting from recent duplication events[307]. Thirdly, protein fusion and fission events are difficult to detect using alignments and thus critical information might be lost.

An alternative approach, already employed in a comparative genomics study of *Escherichia coli* [394], consists of grouping of proteins on the base of domain architectures with a fixed N-C terminal order [241]. Clustering based on domain order is highly scalable and moreover, most protein domains represent structural folds that can be directly linked to function. Here, both approaches were compared. Protein sequence similarity clusters were identified in a BBH approach using orthAgogue [113]. Due to runtime constraints, protein clustering was limited to the analysis of the 58 complete genomes leading to the identification of 14757 protein clusters. For each protein found within a cluster the domain content and N-C terminal domain order ranked by the position of the first detected amino acid of the domain (domain start) in the protein sequence (domain architectures) was analyzed and is summarized in figure 4.3A. 5515 sequence based protein clusters (37%) present a one-to-one correspondence to domain architectures, whereas 3134 (21%) can be associated to two distinct domain architectures. Overall, 93% of the identified clusters can be associated to 4 or less distinct domain architectures. Figure 4.3A also shows the number of proteins in each orthologous cluster. 3162 clusters (21%) contain proteins lacking established domains and almost 75% of them contain less than 10 sequences. These clusters correspond, in their vast majority, to hypothetical proteins. Regarding the core genome, 1618 clusters (11%) were found to be present in all 58 genomes. From these 1618 protein clusters, 242 contained duplication events leaving 1376 distinct single copy gene protein clusters common to all 58 genomes. 543 of those clusters showed a single domain architecture whereas the rest contained domain architecture variations as summarized in figure 4.3B. We noted that such variability was mainly due to swapping or inversion in domains order. In a sequence based approach domain order variation can potentially lead to false negatives, broken clusters and even reduction of the core genome when more genomes are added to the analysis.

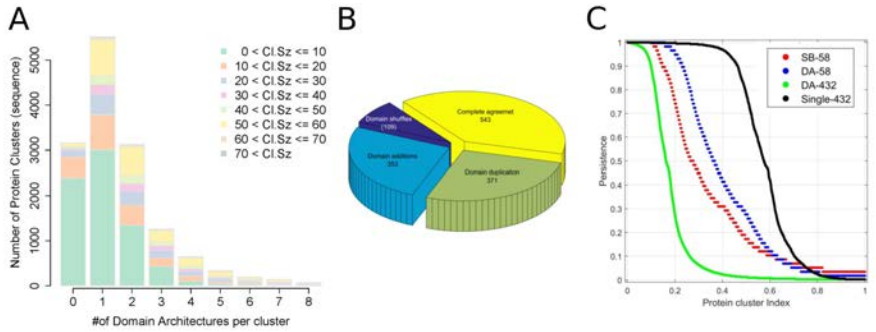


FIGURE 4.3: Domain architectures in sequence based clusters of orthologous proteins. **(A)** Number of distinct domain architectures per cluster **(B)** Variability in domain architectures per gene cluster in core-genome. Complete agreement indicates a unique domain architecture shared by all members of the cluster; For the cases where multiple domain architectures were found in a sequence cluster, the number of cases corresponding to domain duplications, additions and shuffles are indicated. (For A and B only 58 complete genome sequences considered). **(C)** Persistence analysis within the *Pseudomonas* genus. The curves indicate the persistence of each of the cluster. Clusters have been arranged by decreasing persistence values and the x-axis has been scaled to 0-1 range, in this way the cluster with the highest persistence have an x value of 0 and the cluster with the lowest persistence has an x value of 1. The y-axis indicates the persistence of a given cluster (see Equation 1): for instance a persistence of 0.8 indicates that 80% of the analyzed genomes contain sequences in that given cluster. SB-58 refers to the use of sequence based cluster considering the 58 complete genomes; DA-58 and DA-432 refers to the use of protein domains, for 58 and 432 genomes respectively; Single-432 reproduces the analysis for single domain proteins found in the full set 432 genome sequences.

The analysis of 58 complete genome sequences showed that domain architectures retain enough information for functional characterization and that they can be used as a fingerprint for a functional cluster. Since the computational cost for obtaining protein domain identification scales linearly with number of genomes and can be easily distributed over multiple machines, we used these functional fingerprints to extend the analysis to all 432 *Pseudomonas* genomes. Over two million (2,704,339) genes were identified coding for over one million (1,196,884) unique protein sequences of which 85.6% (1,024,877) contain known protein domains. Figure 4.3C shows the results of persistence analysis, reporting the fraction of the total number of analysed genomes in which the corresponding cluster/protein domain/domain architecture was found; 40% the protein domains are persistent in the genus, showing that the functional information at domain level is preserved.

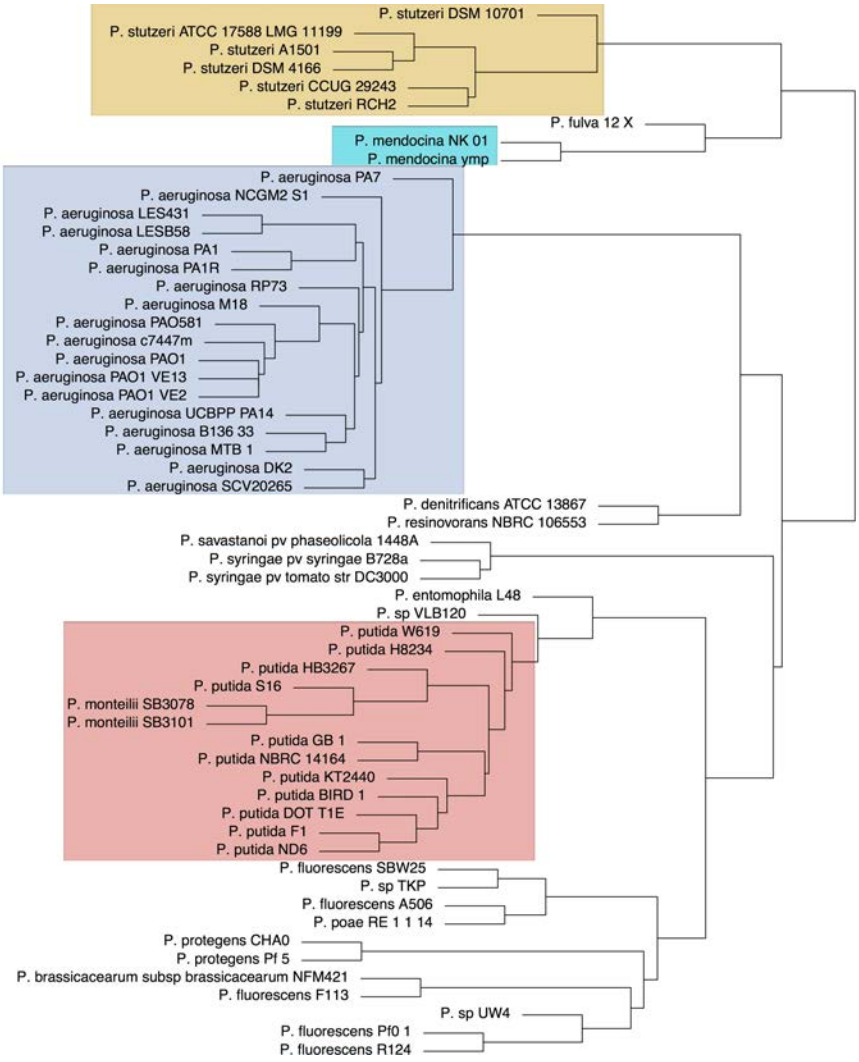


FIGURE 4.4: Domain based distance tree of 58 *Pseudomonas* strains. The tree was built considering the pattern of presence/absence of protein domains using an average clustering approach. Only completely sequenced genomes are considered. The phylogenetic clusters corresponding to the most abundant species (*P. stutzeri, P. mendocina, P. aeruginosa and P.putida*) are color-shadowed.

Classification of *Pseudomonas* strains based on genome potential

Patterns of protein domain presence/absence can provide an alternative and complementary way for assessing strain diversity [5, 458]. There are still many unclassified *Pseudomonas* strains and there is a continuous development on assessing the phylogeny using various approaches [37]. Figure 4.4 shows a distance tree of genome potential based on presence/absence of protein domains for the 58 complete *Pseudomonas* genomes. We found excellent agreement between this distance tree and the taxonomic classification based on 16S sequences indicating that binary patterns of protein domains retain enough information to reconstruct evolutionary history. The positioning of *Pseudomonas* sp. UW4 within the clade of *P. fluorescence*, confirms a previous observation based on 16S and three housekeeping genes (*gyrB*, *rpoB* and *rpoD*) [110]. *P. aeruginosa* and *P. stutzeri* clades are conserved while *P. putida* and *P. fluorescence* clades shows the addition of different species.

We further extended the domain based distance analysis to include all 432 *Pseudomonas* strains (see Supplementary figure S3). The majority of the strains cluster in accord with their taxonomic classification. Many of the unclassified strains could be classified either in *P. aeruginosa* (4) or *P. putida* (13).

Exploring the pan- and core-genome of *Pseudomonas* at protein domain level

The core-genome of a taxon level is defined as the genes persistently present in the population, while the pan-genome is essentially the amount of different genes found within a population at the specified taxonomic level [392]. The currently available genomes allow to measure the pan- and core-genome sizes, however these sizes change upon the addition of new sequences. The core-genome is usually reduced and the pan-genome increases mostly due to the discovery of novel accessory genes that accumulate by lateral transfer, forming new trait combinations until saturation has been reached. Saturated pan-genomes with a stable core-genome are called closed. From the currently available genomes an estimation can be made, using mathematical modelling [392], of the size of the pan- and core- genomes that are expected if the sequences of every existing strain were to be included in the analysis. We refer to these estimations as estimated pan- and core- genome sizes.

Genome potential of the genus *Pseudomonas* is reflected in its metabolic diversity which allows individual species to inhabit a wide variety of environments. With the current set of 432 (draft) genomes we studied whether the observed diversity in genome potential reflects a closed pan-genome. We initially considered the 58 complete genomes. Observed core-genome of 2687

protein domains was to be confronted with an estimated size of 2681. For the pan-genome we found 6472 protein domains (observed) versus 6541 (estimated). Since these measures depend on the number of genomes considered, we explored how these measures vary by using a different number of genomes (from 5 to 58). This was achieved by applying a 10-fold random re-sampling from the 58 genomes to obtain an indication of the possible variability (figure 4.5). As expected the size of the core-genome of the genus decreases with the number of genomes considered while that of the pan-genome increases. The observed and estimated sizes of both the pan- and core-genome are rather stable with respect to the number of genomes used in the calculation, except for small sample size (< 15).

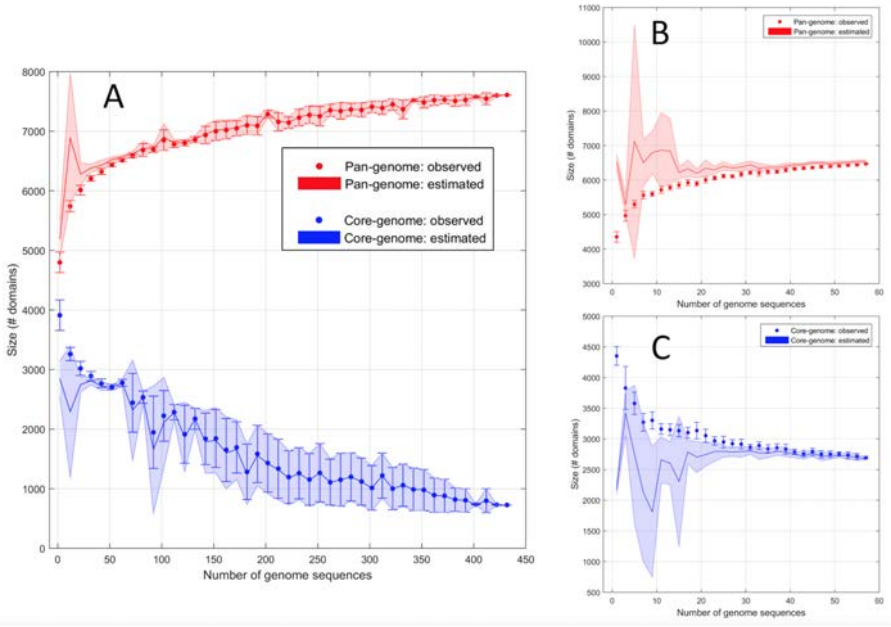


FIGURE 4.5: *Pseudomonas* pan- and core-genome defined on the base of protein domains. **(A)** Complete overview of the distribution of the size of the pan- and core-distribution of protein domains. Error bars correspond to standard deviations based on 10 measured random realizations of the indicated number of genomes whereas the shadowed area is the estimated standard deviation using the same approach. **(B)** Pan-genome of the 58 fully circular genomes. **(C)** Core-genome of the 58 fully circular genomes.

Including draft genomes in the calculations resulted in a dramatic reduction, up to the 73%, of the size of the core-genome both observed and estimated, which dropped to 726 and 720 protein domains architectures, respectively. Interestingly, this reduction does not lead to a loss of functional information since single domains are highly persistent as previously stated (40%).

We observed a large variability for both measures. The reduction of the core size and its variability can be partly explained due to the inclusion of draft genomes with a high number of gaps containing non-sequenced genes. The difference between observed and estimated sizes reduced to only one protein domain for both the pan- and core-genome, indicating saturation. Addition of new genome sequences to the analysis will most likely not lead to the identification of a significant set of new domains within this genus. This saturation effect does not depend on the particular estimation model used. Saturation of the pan-genome was also seen through a heap model ($\alpha = 1.30 \pm 0.05$). In this analysis values > 1 indicate a closed pan-genome [413].

Essentiality analysis of domains in the core-genome

From a functional point of view, the core-genome of a genus is most likely enriched in essential genes necessary for (long term) viability and adaptation to ever changing environmental conditions. Since persistence can be used to identify genes required for survival [1, 273], a positive correlation between persistence (the number of genomes sharing a given gene) and essentiality can be hypothesized. To verify this hypothesis we combined gene essentiality measures with gene persistence in the genus. Gene essentiality was defined from experimental results available for two *P. aeruginosa* strains (PAO1 and PA14)[243, 251] and from *in silico* predictions. For the latter, we considered 6 genome-scale constraint-based metabolic models which rely on functional annotation to uncover the metabolic potential of biological systems and are able to accurately predict gene essentiality in a large variety of growth conditions [318].

We observed that essential genes show higher persistence values than non essential ones: this relationship is conserved when persistence is computed either using a sequence similarity based approach on 58 completely sequenced genomes or for 432 genomes by using a domain architecture approach as shown in figure 4.6A.

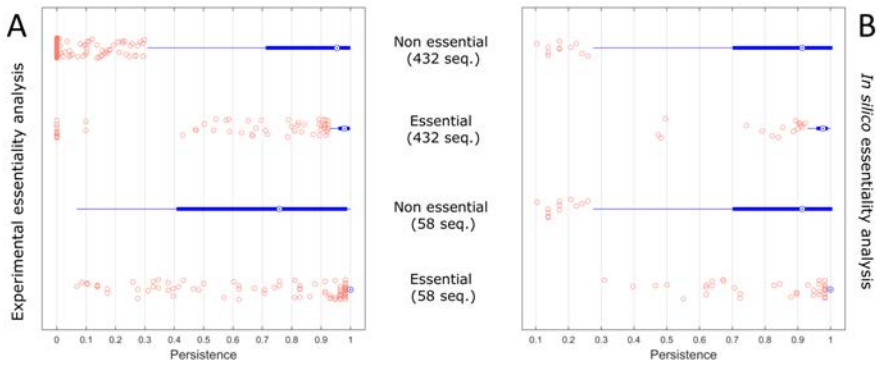


FIGURE 4.6: Persistence of (non-)essential genes. **(A)** Persistence of essential and non-essential genes as derived by experimental investigations. **(B)** Persistence of essential and non-essential genes as derived by *in silico* modeling using genome based constrained metabolic modeling. Results shown pertain the use of the iMO1086 model for *P. aeruginosa* PAO1. In both cases persistence is calculated using the 58 completely sequenced *Pseudomonas* genomes and the complete set of 432 genomes sequences. Magenta (circle) dots indicate outliers.

A comparison of gene persistence and essentiality for the two strains showed that 65% of genes found to be essential for PA14 growth on LB are also essential for growth of PAO1 on either LB, minimal with pyruvate or sputum agar, but only 39% of genes reported to be essential for PAO1 growth were found to be essential for PA14 (see Supplementary figure S4). This difference could be due to the smaller set of tested conditions. We used a less stringent cut-off for persistence: 0.95 instead of 1 to allow for non-sequenced genes due to incomplete draft genomes. Therefore, we observed that a small fraction of persistent genes is present in only one of the two strains (0.016% and 0.025% for PA14 and PAO1, corresponding to 75 and 47 genes respectively) which are likely to have been lost through evolution.

Analysis of the complete pan-genome revealed that 1252 single copy genes are persistent. Of these, almost one third (404) were found to be essential *in vivo* under three growth conditions (LB, minimal-pyruvate or sputum agar) for *P. aeruginosa* PAO1 strain [243]. Similar ratios were observed for strain PA14.

1112 unique domains were identified in the 404 essential persistent genes and 1340 unique domains in the non-essential but persistent genes. 203 domains were shared between essential and non-essential persistent genes. Essential genes contain a larger repertoire of unique, single copy domains: 404 essential persistent genes contained, on average, 1.53 single copy domains whereas for non essential persistent genes, the average was 0.82.

In vivo essentiallity analysis were limited to four conditions. Using metabolic models a wider range of conditions can be explored albeit the analysis is restricted to metabolic genes. We considered six genome scale constraint based metabolic models describing the metabolism of *P. aeruginosa* PAO1 (models iMO1056[310] and iMO1086 [311]), *P. fluorescens* SBW25 (iSB1139 [49]) and *P. putida* KT2440 (iJN746 [298], iJP815 [341], and iJP962 [311]).

We explored a wide range of growth conditions with varying carbon, nitrogen, phosphorus and sulphur sources and for each medium composition, gene essentiality predictions were performed using Flux Balance Analysis (FBA) and are summarized in Table 4.2. Figure 4.6B shows results for *P. aeruginosa* model iMO1086, confirming what was observed for experimental data. Of the 750 essential metabolic genes that were identified under 3366 media compositions for iMO1086, 169 genes were identified to be essential under experimental conditions whereas 42 genes were essential but not *in silico* (25%). Average persistence over the 58 complete genomes was 0.96 ± 0.14 for predicted essential genes and 0.85 ± 0.24 for non-essential, which we found to be significant (p-value < 0.01 for a Wilcoxon test). When considering the 432 genomes, we still observed difference in the persistence of predicted essential and non essential genes 0.95 ± 0.12 versus 0.89 ± 0.21, p-value < 0.01). Similar results were also obtained when using essentiality predictions for the other metabolic models.

TABLE 4.2: Conditional gene essentiality predictions using six metabolic models from three *Pseudomonas* species.

Organism	*P. aeruginosa*		*P. putida*			*P. fluorescens*
Model	iMO1056	iMO1086	iJN746	iJP815	iJP962	iSB1139
Medium sources						
#Carbon	49	51	60	40	43	44
#Nitrogen	32	33	22	25	27	19
#Sulfur	4	1	10	1	1	6
#Phosphor	2	2	1	1	1	2
Genes						
#Essential/persistent*	115/106	149/132	118/104	112/100	162/148	117/95
#Conditional/persistent*	591/278	601/278	389/170	113/64	495/252	615/290
#Non-essential	348	336	253	593	305	407
#Overlapping genes	95		68			

*Persistence was computed for each essential and conditional essential genes over the 58 *Pseudomonas* genomes

Using metabolic models to simulate media compositions we identified additional genes that were essential in a number of conditions, retrieving on average 1.47 single copy domains per gene, consistently with what observed for essentiality experiments. We further combined the models' predictions and we inspected genes predicted to be essential in all the tested conditions. For *P. putida*, the three models showed an overlap of 68 essential genes. Interestingly, these genes contained 2.53 single copy domains on average, underpinning previous results. Non-essential genes contain domains that are shared with other genes. This can result in the presence of isozymes or of potentially moonlighting enzymes which can step in for essential functions in the case of deletions or mutations.

Variability of gene expression and its association to persistence and essentiality in *Pseudomonas*

Associations between gene essentiality and low variation in protein abundance have been observed in *E. coli* [407]. We hypothesized the existence of an association between gene persistence and expression level variation. We analysed gene expression variability in *P. aeruginosa* using a gene expression compendium containing over 900 samples and 100 datasets regarding *P. aeruginosa* PAO1 genes [406]. Each gene was assigned a score, Variability, for transcriptional variation. Persistent genes tend to show significantly lower degree of variation in expression level than non persistent ones (p-value < 0.01); this holds true also for essential genes (figure 4.7). Similar results are obtained when analysing a more limited dataset containing RNAseq measurements of *P. aeruginosa* PA14 in 14 growth conditions [106] (see Supplementary methods S5) This association between low expression variability and persistence/essentiality could indicate that expression of genes in the core-genome is likely to be buffered and independent from environmental growth conditions. To the best of our knowledge such associations have never been established on such large scale due to the limitations associated to comparing hundreds of genome sequences.

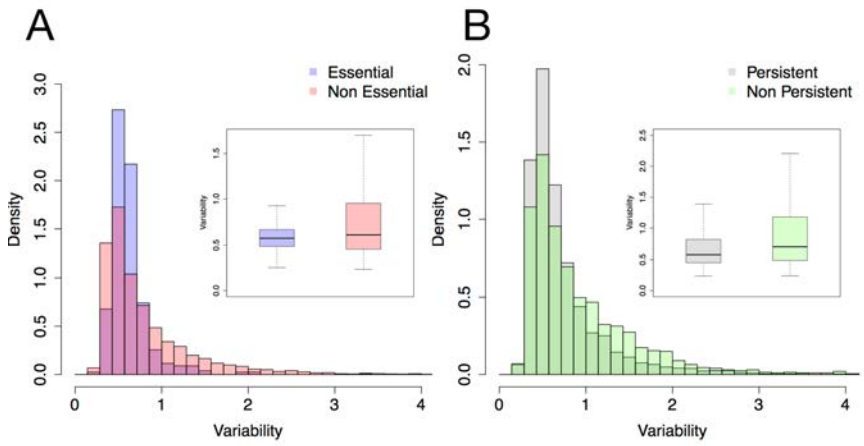


FIGURE 4.7: Variability of gene expression levels and its association with persistence and essentiality. **(A)** Distribution of variability score for (non) persistent genes (genes with persistence lower or higher than 0.95, respectively). Box plots show variability values for both groups. Difference between mean values is significant (p-val < 0.01). **(B)** Distribution of variability score for essential and non-essential genes with gene essentiality derived experimentally [243]. Box plots show variability values for both groups. Difference between mean values is significant (p-val < 0.01).

Discussion

For our analysis we did not rely on previously existing annotations, but we performed a consistent reannotation of all the sequences using a standardized approach that ensured coherence and uniformity. A sequenced based approach was used for a prior comparative analysis to define clusters of orthologous proteins in the smaller dataset of 58 complete genomes. Due to polynomial growth of computational time, this approach is not feasible for large data sets. Mining a gene sequence for domain occurrences is less computationally demanding, which provides an effective scalable approach.

Sequence based approaches are used to identify clusters of orthologous proteins, however the analysis of domain architectures is targeted towards the identification of groups of functionally equivalent proteins. Protein domains provide a standardized way to assess sequence variation and its impact in function, since every amino acid has a characteristic weight in the domain model. Protein domains are more strongly associated to protein structure than protein sequences, thereby providing a closer link to function that can bridge over larger evolutionary distances, which is essential to comparative functional analysis. Still there is a need for improving how protein domain

are defined to accommodate similar models arising from, possibly different, databases and to take into account positional variations that might lead to spurious domain inversions.

When applied to the inferred proteomes of the 58 complete genomes, both clustering methods yield similar results. The same clusters were obtained in 40% of the cases meaning that each of these clusters contained an equal number of proteins, captured the same strains and shared the same domain architectures. In 20% of the cases, very similar but numerically distinct clusters were obtained, as a given sequence similarity cluster had captured two distinct domain architectures. In most of these cases variability in domain architecture were caused by changes in domain order due to small variations in the start position of overlapping domains. Approximately, 20% of identified proteins have no recognizable functional domains. As most of these proteins are hypothetical they were not considered for functional analysis. When only proteins containing domains are considered, over 90% of the clusters identified using sequence comparisons contain 4 or less distinct architectures.

The differences in the persistence curves shown in Fig. 3C show that the way the clusters are defined, either using sequence similarity or protein domains, impacts the calculation of gene persistence: this has repercussions on the definition of the core genome and its size. We found these differences to be larger when more genomes are considered. This is more likely linked to the broader range of phylogenetic distances among considered genomes: this is explored in more detail in Koehorst *et al*.[225]

Our analysis resulted in the identification of the pan- and core-domainome of 432 *Pseudomonas* which is closed according to the heap model as also recently noted for the *P. aeruginosa* species [288]. This suggests that sequencing additional strains will fail to add new genes to the pan-genome: however, this is likely an oversimplification. Here, we understand closeness of the pan-genome as measure of the genus ability to acquire exogenous genes and as a proxy for the ratio between vertical and horizontal gene transfer indicating that horizontal gene transfer has not played a major role in shaping the genome content of the genus.

Key characteristics of *Pseudomonas* must be located in the genus core-genome, however comparison with metabolic models shows that identified core is not autonomously functional. Not all the genes in the core-genome seem to be essential (under given tested conditions), however essential genes represent $\approx 40\%$ of the core-genome, in agreement with previously reported ratios for other species/genus [459]. The remaining 60% contain unique features defining the genus.

We found a strong association between gene essentiality and protein domain properties. We observe an inverse correlation between the number of proteins in the genome containing the considered domain and essentiality,

with average number of domains uniquely present in the considered protein going from 1.5 to 0.8 when non essential/essential genes in the core-genome are considered. The average number of single copy domains per gene further increases when stricter criteria for gene essentiality are applied, namely that genes should be essential in all the simulated media.

Accurate algorithms to predict gene essentiality from genomic features have been also developed and domain enrichment score has been shown to have a high predictive power [102] which is computed based on the ratio of occurring frequencies of a particular domain between essential genes and the total genes in the whole genome of already characterized species. Here we have established a link between the number of copies of a domain in a genome and gene essentiality that can be used to complement essentiality predictions.

The extensive use of metabolic reconstructions allowed us to identify conditionally essential genes, and a large number of single copy domains is also observed in these genes. This supports the idea that protein domains are the driving force behind gene essentiality which is preserved through protein domains rather than through the conservation of entire genes [103].

We have shown that lower fluctuations in gene expression are associated to essential and/or persistent genes. Further work is required to clarify the overlap and intertwining between both gene categories (essential/persistent) and to clarify the (possibly different) regulatory mechanisms stabilizing their expression levels.

Materials and Methods

Genome retrieval. Genbank files containing genome sequences and existing annotations for 58 circular genomes and 374 draft genomes of the *Pseudomonas* genus were downloaded from the GenBank database in June 2015. Annotation of *Pseudomonas* KT2440 was also downloaded from RAST [16]. A detailed list of the included strains is available (see Supplementary figure S1 and Supplementary data S2).

Genome *de novo* annotation. To perform the re-analysis of the 432 gemomes sequences we used an in-house pipeline for annotation and data storage[225]. Similar to existing annotation pipelines such Prokka [383], it relies on external feature prediction tools to identify the coordinates of genomic features within genome sequences. The pipeline consists of a number of python modules that execute annotation applications and convert results and provenance directly into the RDF data model with a self defined ontology (the complete description of the implemented ontology can be obtained using RDF2Graph [96]) using the RDFLib library. For genetic elements determination a variety

of tools is implemented such as Prodigal [185] for gene prediction. The main difference is that results are stored as Turtle files [32] containing an RDF model which allows simultaneous exploration of annotation data of multiple genome sequences, greatly facilitation multiple comparison and the integration of heterogeneous source of information. Since it deploys semantic features allowing the storage of data provenance, we refer to it as SAPP (Semantic Annotation Pipeline with Provenance). Annotation can be exported to other formats for downstream processing with other tools such as Roary [326]. Each genome sequence was converted to the RDF data model using the EMBL/GBK to RDF module. The FASTA2RDF, GeneCaller (a semantic wrapper for Prodigal 2.6 [185]) and InterPro (a wrapper for InterProScan [194]) modules were used to handle and annotate the genome sequences. Results were retrieved with SPARQL queries.

Protein domain presence and phylogenetic analysis. A SPARQL query was used to extract the presence of protein domains for all 432 genomes. Data were stored in a 432 (genomes) by 7608 (protein domains) binary matrix (0/1 for absence/presence). Protein domains were identified by their INTERPRO identifiers. Phylogenetic trees based on protein domains were created taking as input the domain presence/absence matrix. The R package `pvclust` was implemented in R (version 3.3.1)[410] with a binary distance and average clustering approach with a bootstrap value of 10 [404].

Protein domain architecture based clustering. The positions (start and end on the protein sequence) of domains having InterPro [194] identifiers were used to extract domain architectures (*i.e.* combinations of protein domains). Protein domains were retrieved for each protein individually. The domain starting positions were used to assess relative position in the case of overlapping domains; alphabetic ordering was used in the case of domains with the same starting position. Labels indicating N-C terminal order of identified domains were assigned to each protein so that the same labels were assigned to proteins sharing the same domain architecture. Here we have followed a strict approach and two domain architectures were considered different whenever they had different domains or they appeared in different order. For more details see Koehorst *et al.*[225].

Estimation of pan- and core-genome size. The estimated number of domains in the pan- and core-genomes expected if the sequences of every existing strain were to be included in the analysis were computed using binomial mixture models as implemented in the `micropan` R package [393] using the domain presence/absence matrix previously defined and default values for the parameters. *Pan-* and *core-* analysis was initially performed on the

87 genomes with a maximum of 3 contigs to avoid bias due to incomplete genome sequences. Analysis was extended to the remaining 374 draft genome sequences available. To obtain an indication of the variability of these measures as function of the number of sequences used, these were calculated by a 10 fold random sampling from the full set. Heap analysis as implemented in the `micropan` R package was used to estimate openness or closeness of the pan-genome [413] using 500 genome permutations and repeating the calculation 10 times. Final measure is given as the mean $\pm$ standard error.

Orthologous gene detection. Orthologous genes were calculated initially for the set of 58 completely sequenced genomes. Protein sequences predicted using Prodigal 2.6 were extracted using a SPARQL query and used in a Best Bidirectional Hit approach [409]: using an all-versus-all BLASTP comparison and an E-value threshold of 10^{-5} and a maximum target sequence of 10^5. OrthAgogue [113] was used to convert BLAST results into a weighted graph. The MCL [119] clustering algorithm was applied, using an inflation value of 1.5, on the graph to define protein clusters. The results were then extrapolated to the full set of 432 genomes using cluster specific domain fingerprints. Specifically, the sequence clusters obtained through MCL clustering on the 58 complete genomes were used to define sets of protein domains (each sequence cluster was mapped to a set of domains). The remaining genomes were then looked for any given domain set defined on the 58 genomes to define their presence/absence in the draft genomes.

Comparison of gene expression profiles. A publicly available gene expression compendium for *P. aeruginosa* was retrieved [406]. Briefly, this dataset contains a collection of gene expression datasets (950 individual samples pertaining 109 distinct datasets) measured using Affymetrix platform GPL84 and processed using a common normalization and background correction protocol. The final dataset contains expression measurements (in a log_2 scale) for 5549 genes from *P. aeruginosa* PAO1. For every gene we considered its expression profile in this compendium and a Variability value was calculated as the ratio between the standard deviation and the mean.

Persistence and essentiality analysis. The persistence of a gene can be defined as

$$Persistence = \frac{N(\text{orth})}{N} \tag{4.1}$$

where $N(orth)$ is the number of genomes carrying a given orthologue and N is the number of genomes searched [125]. For the 58 completely sequenced genomes, orthologous genes were inferred using a BBH approach. For the full set of 432 sequenced genomes orthologous genes were inferred by making use of protein domain arrangements.Locus tags for predicted proteins were inferred from the original annotation through SPARQL. Locus tags were linked to gene essentiality as defined in experimental studies available for *P. aeruginosa* PAO1 [243] and PA14 [251]. For each of the predicted proteins with inferred locus tag the corresponding protein cluster was initially calculated for the 58 genomes. The domain architecture corresponding to each cluster was extracted and subsequently scanned against all 432 available sequences. We used the MCL clusters as a reference set for the identification of domain architecture variations which were then extrapolated over the 432 genomes. The persistence for each locus tag was calculated and compared against the essentiality score obtained from two experimental studies.

Metabolic model essentiality analysis. We considered six genome scale constraint based metabolic models describing the metabolism of *P. putida* KT2440 (models iJN746 [298], iJP815 [341], and iJP962 [311]), *P. aeruginosa* PAO1 (models iMO1056[310] and iMO1086 [311]) and *P. fluorescens* SBW25 (model iSB1139 [49]). For each genome-scale metabolic model we performed a single gene essentiality analysis in a large number of growth media varying in carbon (C), nitrogen (N), phosphorus (P) and sulphur (S) source. To define the growth media we first identified candidate C, N, P, and S sources in each model independently. Because chemical sum formulas were not always available, we considered each compound for which an exchange reaction was present as a candidate C, N, P and S sources. We changed the *in silico* medium composition to a minimal salts medium containing glucose as C source, ammonia as N source, phosphate as P source, sulphate as S source, in addition to oxygen, water, H^+, and a variety of salts depending on the particular model considered. The potential of each candidate C, N, P, and S source was then evaluated by adding it to the *in silico* medium while omitting the default C, N, P, or S sources. Growth predictions were performed using Flux Balance Analysis [318] as implemented in the Matlab COBRA Toolbox [376]. This provided 4 lists of compounds that were suitable as C, N, P or S sources which were then combined into a single list of growth media by taking all combinations of compounds from the 4 lists. For each medium, we then used the *singleGeneDeletion* function from the COBRA toolbox to determine the growth rate

of the mutant strains. If a gene knock-out reduced the *in silico* growth rate below $10^{-}6$ we considered the gene as essential. Models and Matlab scripts used in this analysis are available in Supplementary data S6.

Availability of Data and Materials. The annotation pipeline framework is distributed under the MIT license. The pipeline all genomic data, data provenance and computational results associated with this study are freely available at `http://semantics.systemsbiology.nl`. Additionally, the data associated to this study are provided in turtle format as an RDF serialized dump. This dataset is made available under the Open Database License: http://opendatacommons.org/licenses/odbl/1.0/.

Supplementary files

The supplementary files of this work can be found online at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5138606/`.

Acknowledgements

The research leading to these results has received funding from the European Union FP7 and H2020 under grant agreements No. 305340 (INFECT) and No. 635536 (EmPowerPutida). This work was carried out on the Dutch national e-infrastructure with the support of the SURF foundation.

Author contributions

J.J.K., J.v.D., V.M.d.S. and P.J.S. participated in the conception and design of the study. J.J.K. and J.v.D. were responsible for the code and design of the semantic framework. R.v.H. performed model-based essentiality analysis. M.S.-D. performed the integration of expression data. J.J.K., E.S., V.M.d.S., M.S.-D. and P.J.S. wrote the manuscript. All authors critically revised the manuscript.

Chapter 5

Breathless: *Pseudomonas putida* redesigned

In silico design of anaerobic *Pseudomonas putida*. **Ruben G. A. van Heck***, Stefano Donati*, Linde Kampers, Pablo I. Nikel, Maria Suarez-Diez, Vitor A. P. Martins dos Santos, Peter J. Schaap, Edoardo Saccenti. Unpublished work.

*Equal contributions

Abstract

Pseudomonas putida is a metabolically versatile bacterium long attracting widespread interest for bioremediation and biotechnology. Underlying this interest are many physiological features including its large reducing power, tolerance to toxins and solvents, and genetic accessibility. However, unlike many other *Pseudomonas* species, *P. putida* is not able to grow anaerobically which limits its potential applications. The obligate aerobic nature of *P. putida* has been attributed to its inability to produce sufficient ATP and maintain redox balance without molecular oxygen. These bottlenecks have been experimentally addressed in *P. putida*, but anaerobic growth was not observed. In this work, we used a combination of genome-scale metabolic modeling and comparative genomics to determine why *P. putida* does not grow in the absence of O_2. We pinpointed several essential O_2-dependent processes that prevent anaerobic growth. Following a model-driven approach we designed *P. putida* strains that can theoretically grow anaerobically via either fermentation or alternative respiration.

Introduction

The *Pseudomonas* genus of gram-negative bacteria is renowned for its metabolic versatility, large reducing power, and tolerance to toxins and solvents [84, 285]. In particular, *P. putida* KT2440 (hereafter: *P. putida*) has a GRAS status, is genetically accessible, has been extensively modeled [33, 171, 298, 311, 341, 395], and has been successfully engineered to produce various compounds of industrial interest [252, 338]. Therefore, *P. putida* is a recognized synthetic biology and industrial workhorse [33, 252, 264, 293, 338]. However, *P. putida* also has a major drawback for industrial processes in its obligate aerobic nature [292, 395]. The requirement for O_2 in large-scale cultivations results in increased expenses and reduced homogeneity due to aeration and stirring, and excludes the use and production of O_2-sensitive enzymes, pathway intermediates, and products. The obligate aerobic nature of *P. putida* is not shared with several other species in the *Pseudomonas* genus, suggesting that *P. putida* can be redesigned as an anaerobe with a limited number of genetic modifications. This has been experimentally attempted several times by either enabling anaerobic fermentation [292, 395] or anaerobic respiration [238, 377, 398].

The first attempt to obtain an anaerobically fermenting *P. putida* was by Sohn *et al.* in 2010. They created a Genome-Scale Metabolic model (GSM) of *P. putida* KT2440, from which they concluded that *P. putida* can not grow anaerobically due to insufficient anaerobic ATP generation. Anaerobic ATP generation was then enhanced by heterologously expressing acetate kinase, which resulted in increased anaerobic survival of *P. putida*, but not in growth [395]. In a later study, Nikel *et al.* reasoned that in the absence of respiration there is not only a lack of ATP generation, but also an accumulation of NADH that can not be oxidized to NAD via the electron transfer chain. Therefore, they expressed acetate kinase, pyruvate decarboxylase and alcohol dehydrogenase II to facilitate both energy generation and redox rebalancing. This approach further increased anaerobic survival, but still did not enable growth [292].

The approaches to create an anaerobically respiring *P. putida* began with the introduction of nitrate and nitrite respiration machinery in *P. putida* in 2013. Nitrate and nitrite respiration are common anaerobic alternatives to O_2 respiration in other *Pseudomonas* species [147], but the machinery is completely absent in *P. putida*. Therefore, Steen *et al.* separately expressed the *nar* and *nir-nor* operons from *P. aeruginosa* in *P. putida* to enable nitrate and nitrite respiration. Both led to increased anaerobic survival in *P. putida*, but growth was not observed [398]. Another approach for anaerobic respiration in *P. putida* was the use of phenazines to transfer electrons from the cell to an electrode. Schmitz *et al.* expressed a phenazine biosynthesis cluster from *P. aeruginosa* in *P. putida* and observed that the phenazines facilitate electron

discharge to the electrode, although the wild type *P. putida* unexpectedly also showed a limited ability to interact with the electrode [377]. Similarly, Lai *et al.* showed that several redox mediators can be added to the culture medium to enable wild-type *P. putida* F1 to discharge electrons to an electrode. Nonetheless, the use of redox mediators and an electrode did not enable *P. putida* to grow anaerobically [238, 377].

The aforementioned studies focused on anaerobic energy generation and redox balancing in *P. putida*. These efforts consistently resulted in increased anaerobic survival but anaerobic growth was not observed. This suggests an additional role for O_2 in essential cellular processes.

In this study, we set out (i) to re-evaluate previous designs for an anaerobic *P. putida in silico*, and (ii) to identify additional limitations to anaerobic growth, with the ultimate goal (iii) to design an anaerobically growing *P. putida*. In our pursuit of this goal we took advantage of both the extensive knowledge on *P. putida* metabolism, as well as of the wealth of genomic data on *P. putida* and other *Pseudomonas* species [224]. Specifically, we used a GSM to probe *P. putida* metabolism, and we used comparative genomics to pinpoint the distinct genetic features of *Pseudomonas* species capable of growing anaerobically. These *in silico* approaches elucidated several limitations to anaerobic growth in *P. putida*, and thereby enabled the design of both fermenting and respiring anaerobic *P. putida* strains.

Results

No anaerobic *in silico* growth for WT *P. putida* and previous designs of anaerobic *P. putida* strains

The previous designs of anaerobic *P. putida* strains were conceptually based on insufficient anaerobic energy generation and redox balancing [238, 292, 377, 395, 398]. Although the designs tackled these problems, they were mostly developed independently from the rest of *P. putida* metabolism [238, 292, 377, 398]. Therefore, we first re-evaluated them in the context of the *P. putida* GMSs iJP962 [311] and iJN746 [298]. iJP962 and iJN746 describe the known metabolism of *P. putida*, as well as its requirements for survival and growth. When analysed with Flux Balance Analysis (FBA) [318], these GSMs can predict whether or not *P. putida* grows in various conditions [298, 311]. The analysis of the two GSMs led to similar results and hereafter we will specifically report those obtained with iJP962.

iJP962 was first confirmed to correctly predict the obligate aerobic nature of wildtype *P. putida* via FBA. FBA was used to predict the maximally achievable anaerobic growth rate of *P. putida* in both a minimal glucose medium and a

rich medium. In both media iJP962 predicted the complete absence of growth. It thus correctly describes the obligate aerobic nature of wildtype *P. putida*.

iJP962 was then used to contextualize the previous anaerobic *P. putida* designs. The GSM was expanded with reactions corresponding to the heterologously expressed genes for each previous experimental design. These expanded GSMs still predicted that anaerobic growth was not possible in neither the minimal glucose nor the rich medium. These predictions are consistent with the experimental observations that none of the previous designs enabled *P. putida* to grow anaerobically. In addition, these predictions imply that iJP962 captures one or more previously undescribed limitations to anaerobic growth in *P. putida*.

Identification of limitations to anaerobic growth

To identify these additional limitations, we performed two independent *in silico* approaches: Genome-scale metabolic modeling and comparative genomics. The metabolic modeling approach focused on identifying the essential O_2-dependent metabolic reactions in iJP962, whereas the comparative genomics approach focused on pinpointing the genetic differences between obligate aerobic *P. putida* strains and other facultative anaerobic *Pseudomonas* species (see figure 5.1).

The obligate aerobic *in silico* phenotype in iJP962 suggests that there are one or more reactions in the GSM that involve O_2 and are essential for growth. To identify these reactions, we first set the *in silico* growth medium to an aerobic minimal glucose medium. Then, we iteratively deleted each reaction that involves O_2 and predicted whether or not *in silico* growth was possible. *In silico* growth was no longer possible upon the deletion of either (i) protoporphyrinogen oxidase, (ii) L-aspartate oxidase, or (iii) dihydroorotate dehydrogenase. These reactions are required for the biosynthesis of heme, NAD/NADP, and pyrimidines respectively.

Next, we evaluated whether the lack of anaerobic alternatives to these three reactions is the only limitation to *in silico* anaerobic growth. We expanded iJP962 with anaerobic alternatives for L-asparate oxidase, dihydroorotate dehydrogenase, and protoporphyrinogen oxidase and again optimized for growth using FBA. iJP962 now predicted anaerobic growth of the modified *P. putida* in the glucose minimal medium, suggesting that the lack of anaerobic alternatives to the aforementioned three reactions is the only limitation to anaerobic growth.

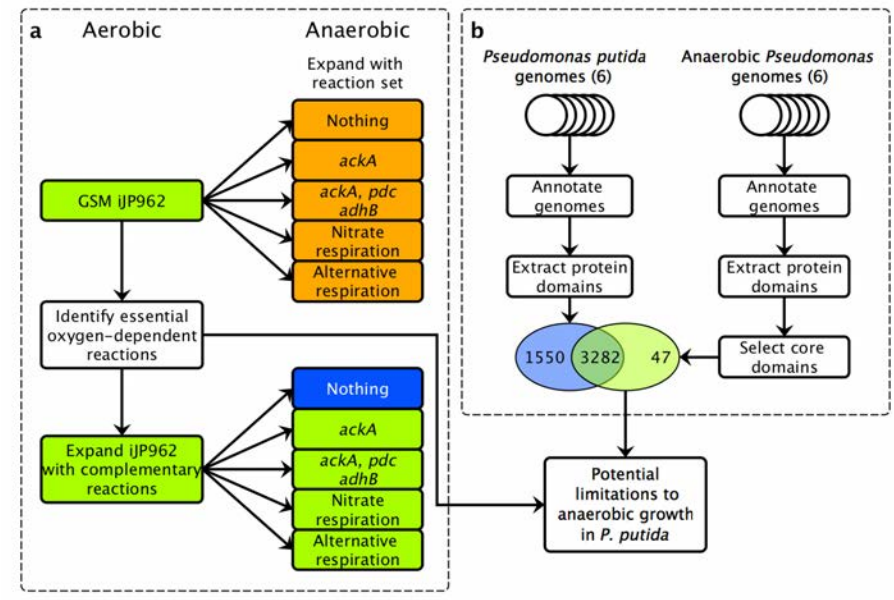


FIGURE 5.1: Overview of *in silico* approaches to identify limitations to anaerobic growth in *P. putida*. **(a)** Genome-scale metabolic modeling. The colors indicate no-growth (orange), poor growth (blue), and growth (green) predictions of iJP962 [311] given an (an-)aerobic environment and expansion with indicated reaction sets. **(b)** Comparative genomics. All genomes of the *P. putida* group and the anaerobic *Pseudomonas* group were annotated using SAPP [224], the annotated protein domains were extracted, and the domains common to all anaerobic *Pseudomonas* species (the core domains) were selected.

The predicted anaerobic growth rate was, however, very low (0.007 [h^{-1}]) compared to the aerobic growth rate (0.450 [h^{-1}]). This difference suggests that the anaerobic ATP generation is extremely inefficient. Therefore, we expanded iJP962 with both the complementary reactions to the newly identified limitations, as well as the reactions corresponding to the previous anaerobic *P. putida* designs that deal with anaerobic ATP generation [238, 292, 377, 395, 398]. Indeed, anaerobic ATP generation proved a limiting factor as the combined designs resulted in predicted growth rates from 0.014 [h^{-1}] with the addition of acetate kinase, and up to 0.171 [h^{-1}] with the addition of nitrate respiration machinery.

Although the GSM-based approach successfully identified several limitations to anaerobic growth in *P. putida*, this identification is restricted to metabolism as described in iJP962. Other cellular processes that may rely on

O_2 are not described. Therefore, we also used comparative genomics to pinpoint genetic differences between select groups of obligate aerobic *P. putida* strains and facultative anaerobic *Pseudomonas* species. The *P. putida* group consisted of the *P. putida* strains KT2440, F1, S16, W619, GB1, and BIRD1, and the facultative anaerobic group consisted of *P. aeruginosa* PAO1 and M18, *P. stutzeri* DSM10701 and A1501, *P. denitrificans* ATCC13867 and *P. fluorescens* F113.

The genomes of the aforementioned *Pseudomonas* strains were *de novo* annotated to avoid artifacts from differences in the annotation procedures. The annotated protein domains were extracted from each genome to compare the presence of functional protein domains. The domains of the anaerobic *Pseudomonas* group were then further filtered for those domains that are shared by all members of the group; the core domains. These core domains were contrasted to the domains present in any *P. putida* to identify those domains that are common to all selected anaerobic *Pseudomonas* species and absent from all selected *P. putida* strains. This resulted in a shortlist of 47 anaerobic-only protein domains (see figure 5.1 and table 5.1).

The 47 anaerobic-only protein domains represent the biological functions that are both common and exclusive to the members of the anaerobic *Pseudomonas* group. These domains thereby pinpoint the genetic makeup of the anaerobic lifestyle. They can roughly be divided in three categories: (i) domains of unknown function (13), (ii) domains related to nitrate and nitrite respiration (16), and (iii) remaining domains (18).

The remaining domains can be further separated as (i) domains specific to several enzymes such as acetate kinase (IPR000890), aspartate decarboxylase (IPR003190), a restriction enzyme (IPR007409), phosphoserine phosphate and homoserine phosphotransferase (IPR011863), ribonucleotide-triphosphate reductase (IPR012833, IPR012840), and molybdenum cofactor sulfurase (IPR015808); (ii) siderophore transport (IPR003538), (iii) structural domains (IPR004443, IPR023218), (iv) iron-sulfur cluster domains (IPR007202, IPR018298), (v) pilus assembly domains (IPR008707, IPR013362, IPR013374, IPR025746), and (vi) a protein kinase domain (IPR017441), as well as (vii) a zinc finger domain (IPR020460).

All domains have been manually inspected using the information on them at Uniprot [89], as well as using scientific articles regarding their encompassing genes. A large fraction of the domains relates to nitrate respiration (see table 5.1), or ATP generation via acetate kinase (IPR000890). For the majority of the remaining domains, it is not directly clear how they would contribute to the desired anaerobic lifestyle. For example, several domains associated with siderophore transport (IPR003538) and pilus assembly (IPR008707, IPR013362, IPR013374, IPR025746) have been related to virulence in *P. aeruginosa* [164],

which is part of its anaerobic pathogenic lifestlye. Other domains may be beneficial but not essential for anaerobic growth. For example, iron-sulfur clusters (IPR007202, IPR018298) are typically O_2-sensitive, which explains their absence in the obligate aerobic *P. putida* strains. Nonetheless, the identification of the two domains that are related to ribonucleotide-triphosphate reductase (IPR012833, IPR012840) has been crucial for this work, as discussed further below.

TABLE 5.1: Overview of identified anaerobic-only protein domains.

Interpro ID	Interpro ID description	Additional comment
Domains of Unknown Function (DUF)		
IPR000952	Uncharacterised protein family UPF0017, hydrolase-like, conserved site	
IPR001602	Uncharacterised protein family UPF0047	
IPR002798	Protein of unknown function DUF95, transmembrane	
IPR007156	LemA	Protein function unknown.
IPR008523	Protein of unknown function DUF805	
IPR010879	Domain of unknown function DUF1508	
IPR014958	DGC	Probably Zinc-binding
IPR018706	Protein of unknown function DUF2214, membrane	
IPR021268	Protein of unknown function DUF2845	
IPR023353	LemA-like domain	Protein function unknown.
IPR024612	Domain of unknown function DUF3749	
IPR025403	Domain of unknown function DUF4129	
IPR025646	Domain of unknown function DUF4350	
Nitrate and Nitrite respiration machinery		
IPR002324	Cytochrome c, class ID	
IPR003143	Cytochrome cd1-nitrite reductase, C-terminal domain	
IPR003765	Nitrate reductase chaperone, NarJ	
IPR003816	Nitrate reductase, gamma subunit	
IPR005346	RnfH protein	Electron transport chain.
IPR006468	Nitrate reductase, alpha subunit	
IPR006547	Nitrate reductase, beta subunit	
IPR008719	Nitrous oxide reductase accessory protein NosL	
IPR010207	Electron transport complex, RnfB	
IPR013615	CbbQ/NirQ/NorQ C-terminal	
IPR020945	DMSO/Nitrate reductase chaperone	
IPR023644	Nitrous-oxide reductase	
IPR023992	Heme D1 biosynthesis, radical SAM protein NirJ	
IPR026464	Nitrous oxide reductase family maturation protein NosD	
IPR028189	Nitrate reductase, alpha subunit, N-terminal	
IPR029263	Respiratory nitrate reductase beta, C-terminal	
Other		
IPR000890	Aliphatic acid kinase, short-chain	Domain of acetate kinase
IPR003190	Aspartate decarboxylase	
IPR003538	Gram-negative bacterial TonB protein	Siderophore transport/binding
IPR004443	YjeF N-terminal domain	
IPR007202	4Fe-4S domain	Potentially O_2-sensitive
IPR007409	Restriction endonuclease, type I, HsdR, N-terminal	
IPR008707	PilC beta-propeller domain	Virulence-related [164]
IPR011863	Phosphoserine phosphatase/homoserine phosphotransferase bifunctional protein	
IPR012833	Ribonucleoside-triphosphate reductase, anaerobic	
IPR012840	Ribonucleoside-triphosphate reductase, anaerobic-like	
IPR013362	Pilus modification type IV, PilV	Virulence-related [164]
IPR013374	ATPase, type IV, pilus assembly, PilB	Virulence-related [164]
IPR015808	Molybdenum cofactor sulfurase, C-terminal-like	
IPR017441	Protein kinase, ATP binding site	
IPR018298	Adrenodoxin, iron-sulphur binding site	Potentially O_2-sensitive
IPR020460	Zinc finger, DksA/TraR C4-type, bacteria	Virulence-related [197]
IPR023218	UPF0291 structural domain	
IPR025746	Type 4 fimbrial biogenesis protein PilX, N-terminal domain	Virulence-related [164]

Analysis of limitations to anaerobic growth

Together, the *in silico* approaches provide a holistic view on the cellular processes to consider in order to design an anaerobic *P. putida* (see table 5.1 and figure 5.2). These include: (i) anaerobic energy generation, (ii) L-aspartate oxidase (*nadB*), (iii) protoporphyrinogen oxidase (*hemY*), (iv) dihydroorotate dehydrogenase (*pyrD*), and (v) ribonucleotide-triphosphate reductase (IPR012833,IPR012840).

The requirement for increased anaerobic energy generation was previously identified [395] and reconfirmed here via both metabolic modelling and comparative genomics. In particular, ATP generation is seen to be growth-limiting according to iJP962, and the comparative genomics pinpointed anaerobic-only protein domains for both acetate kinase (IPR000890) and nitrate/nitrite respiration machinery (see table 5.1). Acetate kinase and nitrate/nitrite respiration machinery have both been successfully expressed in *P. putida* in earlier efforts to enable fermentation or anaerobic respiration [292, 395, 398]. Both fermentation and anaerobic respiration are attainable anaerobic lifestyles in *P. putida* according to our results with iJP962 (see figure 5.1).

L-aspartate oxidase (*nadB*) catalyzes the conversion of L-aspartate to iminoaspartate, a precursor in NAD/NADP biosynthesis. In iJP962 this conversion requires O_2 as electron acceptor. However, in other organisms *nadB* is known to use both O_2 and fumarate as possible electron acceptors [227]. Therefore, we have experimentally assessed the O_2-dependence of *P. putida nadB* through heterologous expression in a *nadB*-deficient *E. coli* strain. We have determined that *P. putida nadB* functions anaerobically, although we have not confirmed that fumarate is the used electron acceptor [data not shown].

Dihydroorotate dehydrogenase (*pyrD*) produces orotate, which is required for pyrimidine biosynthesis, and thus ultimately for the synthesis of RNA and DNA. In iJP962 this enzyme interacts directly with O_2. However, *P. putida pyrD* encodes a membrane-bound class II dihydroorotate dehydrogenase [33], which interacts with quinones rather than directly with O_2 [306]. The re-oxidation of the quinones requires the flow of electrons towards the terminal electron acceptor O_2. Thereby, *P. putida pyrD* is still indirectly dependent on O_2 via the electron transfer chain. To circumvent the need of the electron transfer chain, a class I dihydroorotate dehydrogenase that uses fumarate, FAD, or NAD [306] can be introduced in *P. putida*.

Protoporphyrinogen oxidase (*hemY*) converts protoporphyrinogen IX to protoporphyrin IX, which is further converted to heme. Heme is involved in many cellular processes – including respiration – and is essential for most organisms, excluding some anaerobic fermenters [79, 95]. It is unclear whether heme is essential for an anaerobically fermenting *P. putida*, but it is most likely

required for anaerobic respiration. There are two known alternative protoporphyrinogen oxidases to *hemY* among gram negative bacteria: *hemG* and *hemJ*. The gene *hemG* is quinone-dependent instead of O_2-dependent, but is not found in other *Pseudomonas* species. The gene *hemJ* is found in *Pseudomonas species*, including *P. aeruginosa* [54] and *P. putida*, but it has only recently been identified and its exact mechanism, including whether or not it relies on O_2, is currently unknown [79].

Ribonucleotide-triphosphate reductases (RNRs) are required for the biosynthesis of nucleotides and, thus, DNA. There are three classes of RNRs: Class I is strictly aerobic, class II is O_2-independent, and class III is O_2-sensitive [93, 387]. *P. putida* only has a class I RNR and is not able to produce DNA in anaerobic conditions. In contrast, *P. aeruginosa* has access to all three RNR classes [387]. RNR knockout experiments in *P. aeruginosa* revealed that the class II enzyme contributes most to anaerobic dNTP production in *P. aeruginosa* [387], and, in a different study, the class II RNR was reported to be essential for anaerobic growth in *P. aeruginosa* [135]. Therefore, a class II RNR seems the most promising candidate to enable anaerobic nucleotide production in *P. putida*.

Designs of anaerobic *P. putida*

The identified limitations and alternative options enable the theoretical design of anaerobically growing *P. putida* strains. Here we present two possible designs of an anaerobically growing *P. putida*. These designs are based on the previously successfully expressed acetate kinase for fermentation [292, 395] and nitrate respiratory machinery for respiration [398]. Note that L-aspartate oxidase does not need to be replaced in either design as the endogeneous enzyme was experimentally determined to function anaerobically [data not shown].

The designed anaerobically fermenting *P. putida* is not capable of reoxidising its quinones. Therefore, it employs: (i) a quinone-independent class I dihydroorotate dehydrogenase, which can be found in most gram-positive bacteria [229], (ii) the class II ribonucleotide-triphosphate reductase (RNR) from the closely related *P. aeruginosa*, and (iii) acetate kinase also from *P. aeruginosa* (see figure 5.3). We have not included an alternative biosynthesis gene for the heme precursor protoporphyrin IX. It is likely that protoporphyrin IX can be synthesized by the endogeneous *P. putida hemJ*, although it is currently unknown whether this directly or indirectly depends on O_2. In addition, heme is not required for all anaerobic fermenters [79, 95].

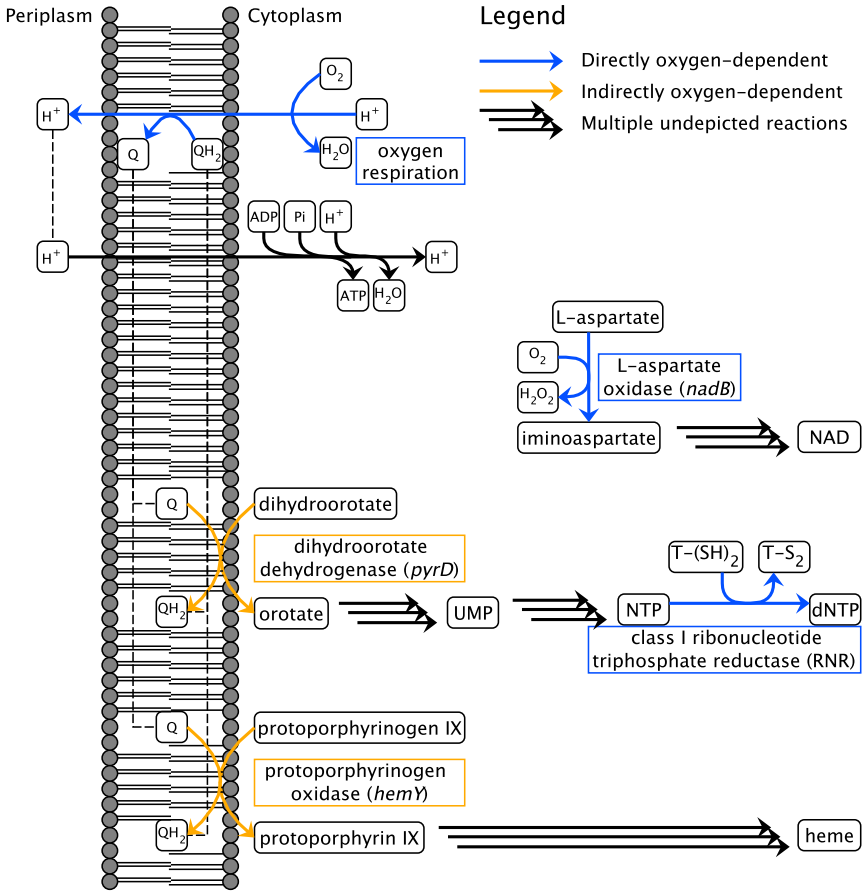


FIGURE 5.2: Computationally identified limitations to anaerobic growth in *P. putida*. Pi: Phosphate; QH2: quinol; Q: quinone; T-$(SH)_2$: thioredoxin; T-S_2: thioredoxin disulfide; NTP: nucleoside triphosphate; dNTP: deoxynucleoside triphosphate; UMP: uridine monophosphate

The designed anaerobically respiring *P. putida* is capable of re-oxidising its quinones, and thus does not need an alternative enzyme for the biosynthesis of orotate or protoporphyrin IX. This designed consists of: (i) the class II ribonucleotide-triphosphate reductase (RNR) from the closely related *P. aeruginosa*, and (ii) the nitrate respiration machinery from *P. aeruginosa* (see figure 5.4).

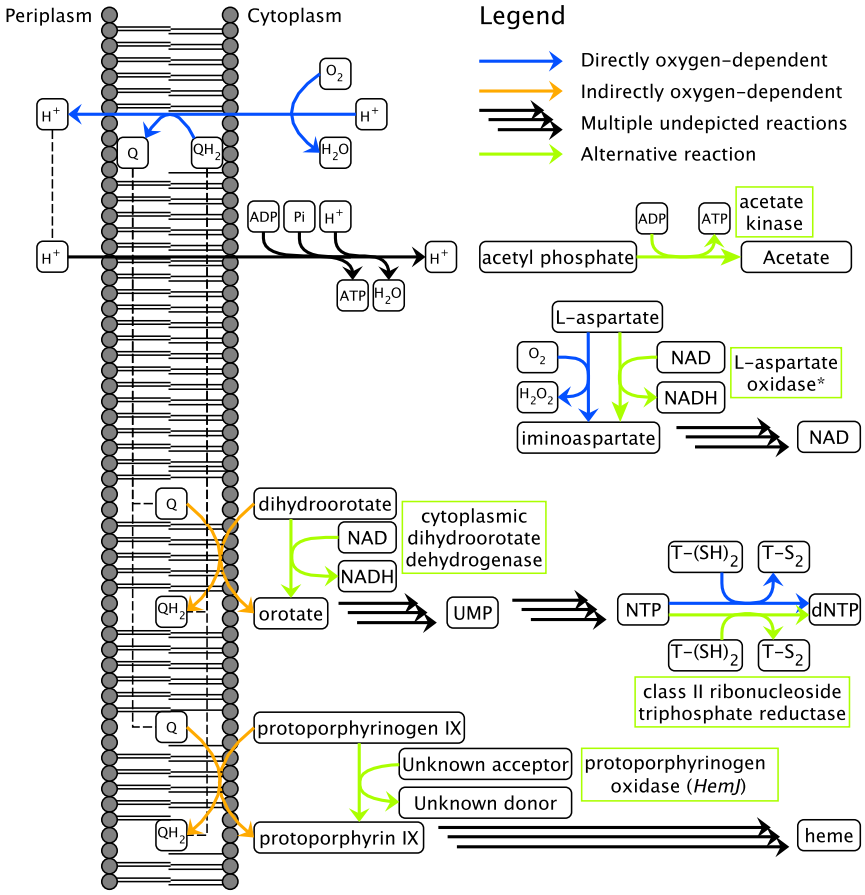


FIGURE 5.3: Design of anaerobically fermenting *P. putida*. *The endogeneous L-aspartate oxidase has been experimentally determined to function anaerobically, but the proposed co-factor has not been confirmed. Pi: Phosphate; QH2: quinol; Q: quinone; T-$(SH)_2$: thioredoxin; T-S_2: thioredoxin disulfide; NTP: nucleoside triphosphate; dNTP: deoxynucleoside triphosphate; UMP: uridine monophosphate

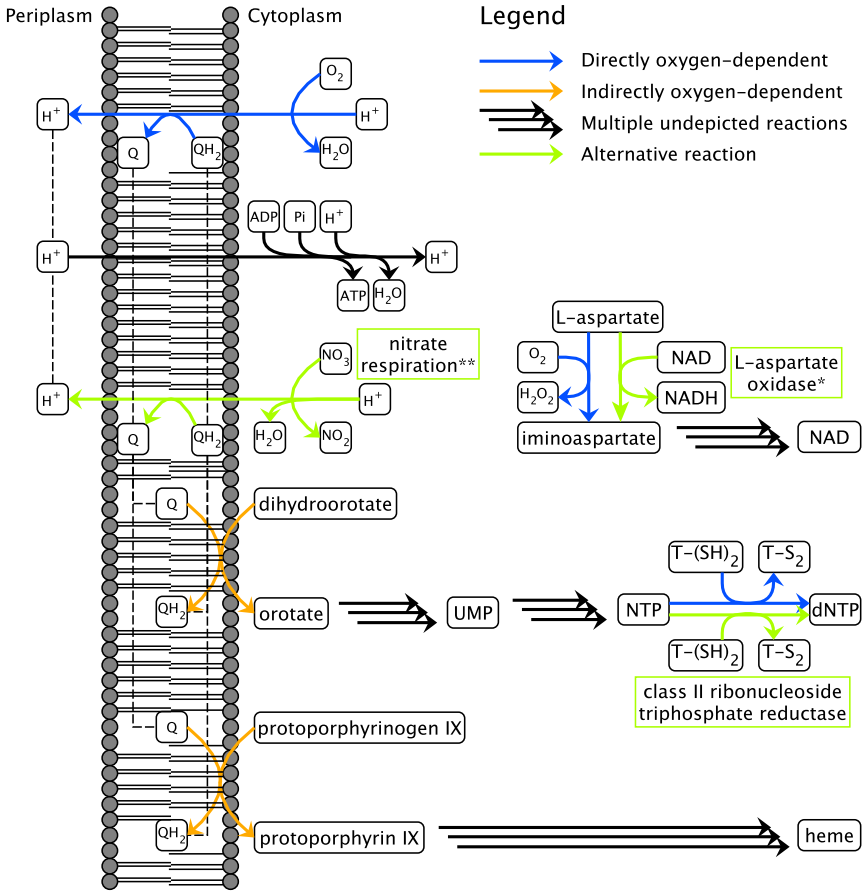


FIGURE 5.4: Design of anaerobically respiring *P. putida*. *The endogeneous L-aspartate oxidase has been experimentally determined to function anaerobically, but the proposed co-factor has not been confirmed. Pi: Phosphate; QH2: quinol; Q: quinone; T-$(SH)_2$: thioredoxin; T-S_2: thioredoxin disulfide; NTP: nucleoside triphosphate; dNTP: deoxynucleoside triphosphate; UMP: uridine monophosphate

Discussion

Previous efforts to design an anaerobically growing *P. putida* [238, 292, 377, 395, 398] were based on intuitive differences between aerobic and anaerobic conditions. The prime consideration was that without access to aerobic respiration, *P. putida* could not generate sufficient energy or fully balance its redox state. The various approaches to complement these processes with anaerobic alternatives resulted in increased anaerobic survival, but not in anaerobic growth [238, 292, 377, 395, 398]. This prompted us to explore whether these designs should have worked based on the current theoretical understanding of *P. putida* metabolism as represented in iJP962 [311]. Consistent with the previous experimental results, iJP962 predicted that the previous designs do not enable anaerobic growth, thereby implying that additional limitations exist.

We identified these limitations *in silico* using a combination of genome-scale metabolic modeling and functional comparative genomics. iJP962 was used to identify the essential reactions of *P. putida* metabolism that rely on O_2 to function, and to evaluate the inclusion of alternative reactions. Comparative genomics was used to pinpoint the genetic elements that distinguish the facultative anaerobic *Pseudomonas* species from *P. putida*. Together, these *in silico* approaches provided a holistic view on the limitations for anaerobic growth in wildtype *P. putida*.

We have used these newly identified limitations to design both respirative and fermentative anaerobic *P. putida* strains. In either case, the class I RNR has to be complemented by a class II or class III RNR. Anaerobic growth via respiration can then be obtained via the expression of, for example, nitrate respiration machinery. Instead, anaerobic growth via fermentation requires the expression of a class I dihydroorotate dehydrogenase and a fermentative ATP-generating enzyme such as acetate kinase. According to the understanding of *P. putida* metabolism as represented in iJP962, both of these designs enable anaerobic growth in *P. putida*.

The respirative design appears easier to realise than the fermentative design for several reasons: (i) The sheer number of limitations to complement is lower. One of the two respirative design complementations does regard the expression of the 10-gene *nar* operon, but this operon has succesfully been expressed in *P. putida* before [398]. (ii) There is little documentation on fermentative anaerobic growth of other *Pseudomonas* species, although fermentation using arginine [135, 433] and pyruvate [122, 135] have been described for *P. aeruginosa*. (iii) Not all electron transfer chain-dependent processes may be identified. iJP962 might be missing or misrepresenting these processes as it is certainly incomplete [33], and redox metabolism is typically inconsistently represented in GSMs [171]. In addition, the comparative genomics analysis would only pick up alternative processes if they are shared by all members

of the anaerobic *Pseudomonas* group. However, some of these may only be capable of anaerobic respiration, implying they have no need for these alternative processes. (iv) Replacing one type of respirative growth with another will require a less drastic adaptation of native *P. putida* metabolism.

The more drastic adaptation towards fermentative growth is, however, more interesting to entertain from a theoretical and synthetic biology viewpoint. The current theoretical understanding of *P. putida* metabolism as described in iJP962 supports fermentative growth in *P. putida* following but a small number of heterologously expressed genes. The fermentative design of *P. putida* as detailed in figure 5.3 is currently being experimentally evaluated in our laboratory with very promising preliminary results [data not shown].

Conclusions

The results of this study demonstrate the benefit, or even necessity, of computationally exploring the requirements and challenges in synthetic biology. This computational exploration revealed that previous approaches on enabling anaerobic growth in *P. putida* [238, 292, 377, 395, 398] did not address all limitations to anaerobic growth. It is well-recognized that aerobic energy generation needs to be replaced with an anaerobic alternative, but other less intuitive limitations to anaerobic growth in *P. putida* were not previously identified and, thus, not addressed. We expect that our approach of combining metabolic modelling and comparative genomics can find widespread use as the methods are inherently complementary. GSMs, already available for many organisms [456], describe one organism in high detail; whereas comparative genomics pinpoints the genetic elements underlying the differences between organisms.

The newly identified limitations to anaerobic growth in *P. putida* provide clear directions towards the creation of an anaerobically growing *P. putida*. Wildtype *P. putida* requires O_2 not only for energy generation, but also for the biosynthesis of heme, pyrimidines, and nucleotides. Once these processes are complemented by anaerobic alternatives, *P. putida* theoretically has all the required tools for an anaerobic lifestyle. This anaerobic lifestyle opens up novel biotechnological applications for *P. putida*, which may be exploited in future work. Regardless, the model-driven redesign of an obligate aerobe into a facultative anaerobe constitutes an important fundamental step in the rational redesign of biological systems.

Materials and Methods

Genome-scale metabolic models. In this study we used the *P. putida* genome-scale metabolic model iJP962 [311]. iJP962 was previously constructed in our group and directly obtained from the authors.

Metabolic modelling. iJP962 was analyzed using Flux Balance Analysis (FBA) [318] in the CobraPy toolbox [112]. The minimal glucose medium conditions were set by allowing 'unlimited' (-1000 [mmol gdw^{-1} h^{-1}]) uptake of copper, cobalt, iron, protons, water, sodium, nickel, ammonia, phosphate, sulfate, and nitrate. In addition, the glucose uptake rate was set to be maximally 6.14 [mmol gdw^{-1} h^{-1}], based on experimentally measured uptake rates in [295]. The rich medium conditions were set by allowing unlimited ([mmol gdw^{-1} h^{-1}]) uptake of all compounds for which exchange reactions are present in the GSM. In aerobic conditions, the O_2 uptake rate was set to maximally 18.5 [mmol gdw^{-1} h^{-1}]. Reactions for previous designs or to complement the newly identified issues were added using the CobraPy 'addReaction' function.

Genome annotation. We selected 6 genomes of facultative anaerobic organisms from the *Pseudomonas* genus (*P. aeruginosa* PAO1 and M18, *P. stutzeri* DSM10701 and A1501, *P. denitrificans* ATCC13867 and *P. fluorescens* F113) and 6 genomes of obligate aerobic *P. putida* strains (KT2440, F1, S16, W619, GB1, and BIRD1). All genomes were obtained from the EnsemblBacteria [208] repository on March 2015 and annotated in SAPP [224] using Prodigal (version 2.6)[185]. The annotated genomes were run using InterProScan version 5.4-47.0 [194] with the selected applications: TIGRFAM, ProDom, SMART, PROSITE Pattern, PfamA, PRINTS, SUPERFAMILY, Coils, Gene3d.

Acknowledgements

We thank Ruud Weusthuis for insightful discussions on anaerobic growth regarding the continuation of this project, and thank Dorett Odoni for feedback on the manuscript.

Author contributions

Conceived and designed the experiments: RGAvH PS.
Performed the metabolic modelling: RGAvH SDo.
Performed the comparative genomics: SDo ES.
Performed the experimental work: SDo LK.
Analyzed the data: RGAvH SDo ES PN LK MSD.
Supervised the work: RGAvH ES MSD PN VMdS.
Wrote the manuscript: RGAvH.
Arranged funding: VmdS.

Funding

We gratefully acknowledge financial support from the Wageningen University IPOP project, and the European Horizon2020 project EmPowerPutida (Project reference: 635536). The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

Chapter 6

Fix it! CO_2 fixation redesigned.

In silico design of species-specific synthetic CO_2 fixation pathways. **Ruben G. A. van Heck***, Henri van Kruistum*, Maria Suarez-Diez, Nico J. Claassens.

*Equal contributions

Abstract

Carbon fixation refers to the conversion of inorganic CO_2 into organic compounds. Microbial CO_2 fixation followed by bioproduction of chemicals is a promising alternative to the fossil-based industry. However, CO_2 fixation pathways found in nature are relatively inefficient, thereby limiting autotrophic growth and production. Efforts have been made to design more efficient synthetic CO_2 fixation pathways and transplant them in model organisms. However, these pathway designs often overlook the metabolic context of the production organism, leading to pathways requiring a too large number of heterologous enzymes to be expressed. Here, we present CO2FIX, an algorithm to design CO_2 fixation pathways optimally tuned to the native metabolic capabilities of the target organism. Genome-Scale Metabolic models are used as a representation of the organisms metabolic capabilities. We deployed CO2FIX to design species-specific CO_2 fixation pathways in eight microbes. Design specifications include a limited number of non-native reactions, ATP efficiency, thermodynamic feasibility and kinetic efficiency. For each of the investigated organisms, CO_2 fixation pathways were designed with five or less non-native reactions while resulting in a high predicted growth rate on CO_2. Analysis of the designed pathways showed them to be more efficient, within the target organism, than both naturally occurring CO_2 fixation pathways and recently described synthetic pathways. We highlight three newly identified CO_2 fixation pathways with particularly promising features.

Introduction

Carbon fixation is the metabolic process converting inorganic CO_2 into organic compounds; a key process for the global ecosystem, agricultural production, and with great potential for the biobased production of fuels and chemicals. Direct conversion of CO_2 into chemical compounds of interest by autotrophic microorganisms is an important avenue to explore for sustainable production. In nature, a wide variety of chemoautotrophic and photoautotrophic microorganisms is available, but their potential for biobased production remains limited, partly due to limitations in genetic accessibility and genetic toolboxes for most microbial autotrophs [83].

Additionally, autotrophic growth and production is often hampered by the limited availability of external energy inputs and the inefficiency of their energy-harvesting systems, such as light and photosystems for photoautotrophs, and hydrogen and hydrogenases for chemoautotrophs [83]. Those energy-harvesting systems provide energy carriers (e.g. ATP) and electron donors (e.g. NADP(P)H or ferredoxin), primarily to drive CO_2 fixation.

The naturally dominant autotrophic CO_2 fixation pathway, the Calvin cycle, requires a relatively high ATP input, while relying on a notoriously slow carboxylase: RuBisCO [120]. Apart from the ubiquitous Calvin cycle, only five autotrophic pathways for CO_2 fixation have been discovered in nature: the Wood-Ljungdahl pathway (WL), the reductive tricarboxylic acid cycle (rTCA), the 3-hydroxpropionate-4-hydroxybutyrate cycle (3HP-4HB), the dicarboxylate-4-hydroxybutyrate cycle (DC-4HB) and the 3-hydroxypropionate bi-cycle (3-HP) [35]. Some of these pathways have advantageous features relative to the Calvin cycle such as lower ATP-requirements and faster kinetics.

In addition to these natural pathways, there have been synthetic biology efforts to design novel pathways by exploiting the repertoire of characterized enzymes [53, 82]. By mixing and matching either natural or engineered enzymes attractive synthetic CO_2 fixation pathways can be designed and realized [23, 121, 380]. An extensive exploration of the known repertoire of natural enzyme reactions by Bar-Even *et al.* [23] led to the identification of a large number of potential synthetic CO_2 fixation pathways. Promising pathways were further selected based on energetic efficiency, favorable thermodynamics, predicted pathway kinetics and the number of enzymes involved. Pathways identified in this work include the oxygen-tolerant malonyl-CoA-oxaloacetate-glyoxylate (MOG) pathways, such as the C4-glyoxylate pathway involving the kinetically fast phosphoenol pyruvate carboxylase (PEPC). A few more ATP-efficient, but oxygen-sensitive, ferrodoxin oxidoreductase-based pathways were identified in this study as well, such as the pyruvate synthase - pyruvate carboxylase - glyoxylate bi-cycle

(PyrS-PyrC-glx). Another promising synthetic pathway based on the fast crotonyl-CoA-carboxylase, the CETCH-cycle was recently designed and realized *in vitro* [380].

These synthetic and alternative natural CO_2 fixation pathways may prove advantageous for industrial biobased production due to their advantageous features compared to the Calvin cycle. By replacing the Calvin cycle, the native CO_2 fixation can be enhanced in autotrophs. However, metabolic engineering usually has an extensive list of requirements and often biotechnological applications favor well-characterized heterotrophs over autotrophs. The efficient exploitation of alternative CO_2 fixation pathways may thus require these pathways to be introduced into biotechnologically relevant heterotrophs. This arduous task also requires the challenging transplantation of energy-harvesting systems such as photosystems or inorganic electron uptake mechanisms [83, 422].

The performance of alternative natural and synthetic pathways will depend on the metabolic context of the used organism. Furthermore, functional engineering of synthetic pathways *in vivo*, designed without considering the metabolic context, will often require the introduction of a large number of non-native enzymes in most microbial hosts. Heterologous expression and optimization of a large number of non-native enzymes for a functional CO_2 fixation pathway has been demonstrated to be challenging [267]. In contrast, the successful engineering of a functional Calvin cycle in *Escherichia coli*, requiring only the expression of three non-native enzymes and subsequent experimental evolution, has recently been demonstrated [13].

Efficient design of synthetic pathways for CO_2 fixation in biotechnologically relevant microbes would greatly benefit from a holistic systems biology approach able to explicitly account for the native metabolic context of the host. This holistic approach has the potential to identify the best suited non-native enzymatic reactions and combine those into efficient CO_2 fixation pathways for efficient autotrophic growth and production.

Here, we present CO2FIX, an algorithm to design species-specific CO_2 fixation pathways requiring the expression of a limited number of non-native genes to enable efficient CO_2 fixation. CO2FIX uses (i) Genome-Scale Metabolic models (GSMs) to describe the metabolism of the target organism, (ii) Mixed-Integer Linear Programming and parsimonious Flux Balance Analysis (pFBA) [249] to identify energy-efficient and species-specific candidate CO_2 fixation pathways, (iii) Max-min Driving Force (MDF) [305] to assess the thermodynamic feasibility of those candidate pathways, and (iv) Pathway Specific Activity (PSA) [23] to evaluate the kinetic feasibility of those candidate pathways.

We have deployed CO2FIX to design pathways for eight biotechnologically relevant microorganisms, including heterotrophs and autotrophs, as well

as aerobes and anaerobes. For all organisms we identified several not-yet-described synthetic CO2FIX pathways that satisfy all design requirements. Several of these pathways are ATP-efficient, thermodynamically feasible, and kinetically attractive, while requiring the introduction of surprisingly few non-native enzymes.

Results

CO2FIX

We developed CO2FIX, a computational method to design species-specific metabolic pathways endowing the target organism with CO_2 fixation capabilities or improving existing capabilities. The metabolism of an organism is represented by a Genome-Scale Metabolic model (GSM). Within CO2FIX a CO_2 fixation pathway is defined as the set of reactions required to produce 1 pyruvate from 3 CO_2, excluding reactions representing ATP generation, cofactor regeneration, cross-membrane transport, and exchange with the medium. CO2FIX consists of four distinct phases, namely: (i) GSM expansion, (ii) Candidate pathway elucidation, (iii) Thermodynamic evaluation, and (iv) Kinetic evaluation (see figure 6.1).

In the GSM expansion phase, reactions are added to the GSM in order to enable or improve CO_2 fixation. Autotrophic growth requires a mechanism to generate ATP from either light or an inorganic electron donor and a mechanism to generate reducing equivalents (NAD(P)H and reduced ferrodoxin) using an external inorganic electron donor. Therefore, in the initial step of this phase, reactions representing an anoxygenic photosystem and hydrogenases are added to the GSM if such systems are not already natively present. Then, a reference reaction database compatible with the metabolite naming system in the model (see materials and methods) is mined to identify a predefined number of reaction additions enabling the highest possible growth rate. Optimal database exploration is ensured by a Mixed-Integer Linear Programming (MILP) approach. Iteration of this phase, excluding previously identified reaction sets, leads to identification of additional reactions sets.

Candidate synthetic CO_2 fixation pathways are formed by some native GSM reactions and the newly added reactions. In the second phase, reactions in the candidate CO_2 fixation pathway and their pathway coefficients are determined. To this task CO2FIX uses parsimonious FBA (pFBA) [249], but instead of the original approach (maximal flux through biomass reaction), a flux of 1 [$mmol\ gdw^{-1}\ h^{-1}$] is set for pyruvate production. Reactions required to carry flux for pyruvate production and not associated to transport, cofactor regeneration, or ATP generation, constitute the candidate CO_2 fixation pathway.

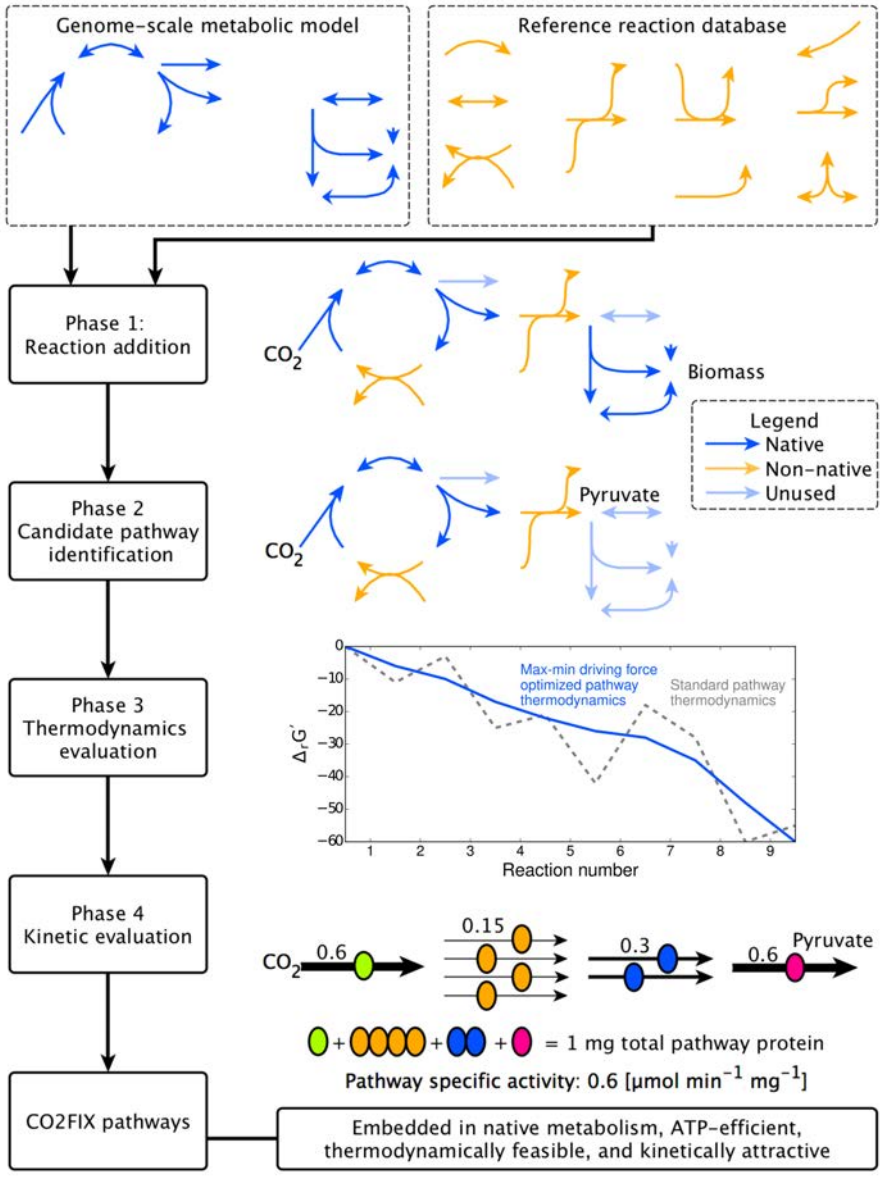


FIGURE 6.1: The CO2FIX process.

The third phase represents the thermodynamic evaluation of the candidate pathways. Thermodynamic feasibility of a biological pathway implies a negative Gibbs free energy change, or a positive thermodynamic driving force for each reaction under physiological conditions, defined as $-\Delta_r G'$. Feasibility of all individual reactions in a pathway is preferably evaluated in the context of a pathway in the cellular environment, as product and substrate concentrations of all reactions determine actual Gibbs free energies and driving forces. Such an analysis can identify the bottleneck reaction in a pathway, i.e. the specific reaction that has the minimal thermodynamic driving force of a complete pathway. CO2FIX evaluates candidate pathways using the Max-min Driving Force (MDF) method [305], which optimizes the concentrations of involved metabolites throughout the pathway, within physiological ranges, in order to maximize the minimal driving force. Pathways for which no positive MDF can be obtained are deemed not feasible and discarded.

In the last phase, kinetics of the candidate pathways are evaluated using Pathway Specific Activity (PSA) [23]. PSA is defined as the maximal pathway flux attainable by 1 mg of total pathway protein. In other words, PSA is a weighted average of the specific activities of the involved enzymes. The higher the PSA is, the faster the pathway is expected to operate given the same amount of enzymes.

CO_2 fixation requires few non-native reactions

The CO2FIX algorithm is employed to identify promising CO_2 fixation pathways for eight microorganisms with biotechnological relevance; each represented by a species-specific GSM. Our selection includes the most common facultative anaerobic heterotrophic workhorses *E. coli* (GSM: iJO1366 [319]), *Bacillus subtilis* (iYO844 [312]), and *Saccharomyces cerevisiae* (iMM904 [282]). In addition, we included the obligate aerobic heterotrophic biotechnological workhorse *Pseudomonas putida* (iJN746 [298]). Furthermore the strictly anaerobic heterotrophic thermophilic bacterium *Thermotoga maritima* (iLK478 [466]) is included, as thermophilic cell factories are desired for several biotechnological applications. Potential autotrophic cell factories included are the strictly aerobic photosynthetic cyanobacterium *Synechocistis* sp. PCC6803 (hereafter: *Synechocystis*) (iJN678 [301]) and *Rhodobacter sphaeroides* (iRsp1095 [187]), which can grow both phototrophically with hydrogen in anaerobic conditions and chemolithoautotrophically with hydrogen in aerobic conditions. The last organism included is the strict anaerobic facultative autotrophic *Geobacter metallireducens* (iAF987 [132]), which can also respire using iron as an electron acceptor and formate as an electron donor.

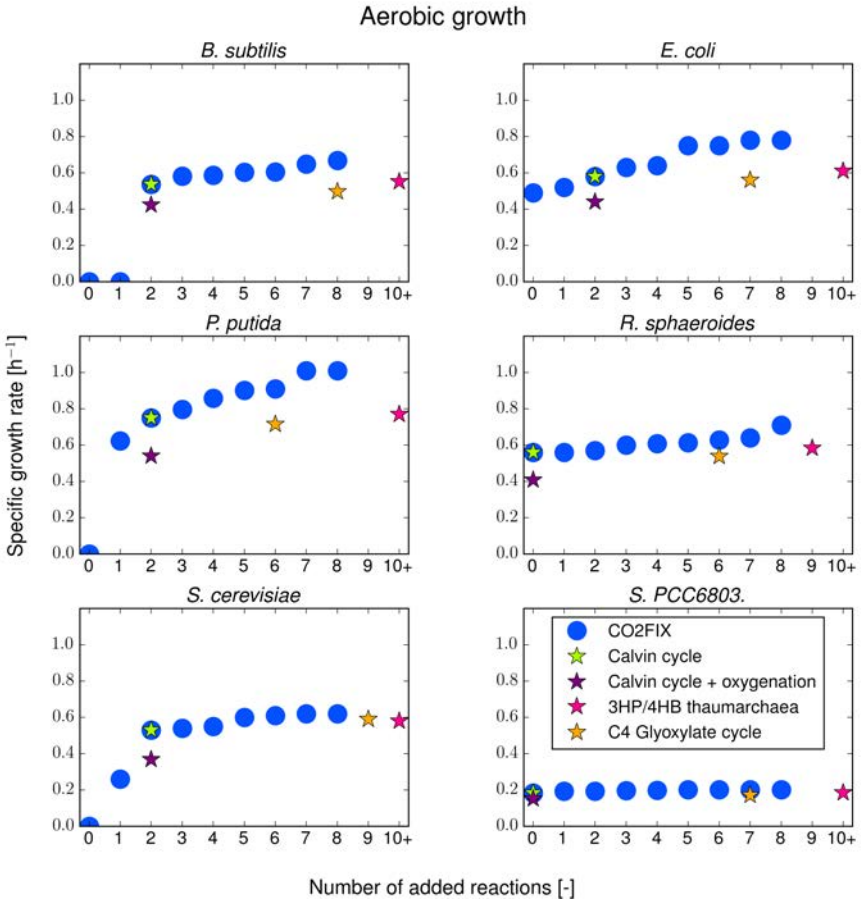


FIGURE 6.2: Growth rate predictions aerobic conditions.

CO2FIX is employed to design, for those organisms, CO_2 fixation pathways requiring few non-native reactions. Features of these newly identified pathways are compared to a set of reference pathways, including all known natural CO_2 fixation pathways and earlier proposed promising synthetic CO_2 fixation pathways [23] (see figures 6.2 and 6.3).

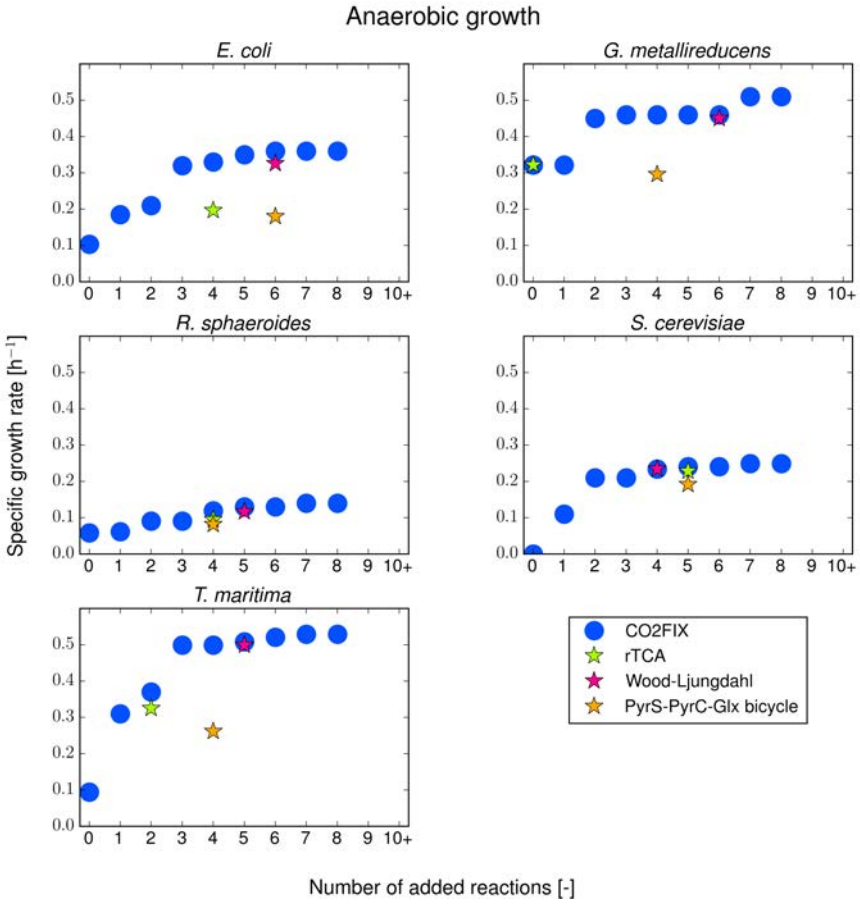


FIGURE 6.3: Growth rate predictions anaerobic conditions.

The heterotrophs require remarkably few reaction additions for CO_2 fixation, irrespective of aerobic (see figure 6.2) or anaerobic (see figure 6.3) conditions. Only two reaction additions are required for *B. subtilis*, and *P. putida* and *S. cerevisiae* require only a single reaction addition each. The other organisms are already predicted to be capable of CO_2 fixation. Surprisingly, this includes *E coli* and *T. maritima*. These organisms are not capable of CO_2 fixation, but candidate CO_2 fixation pathways are nonetheless present in their respective GSMs.

The autotrophs present a more varied profile with regards to reaction additions enhancing their growth rate on CO_2, and the dependence thereof on oxygen. In anaerobic conditions, a considerable improvement for *G. metallireducens* and *R. sphaeroides* can be obtained with two reaction additions. *R. sphaeroides* has a low anaerobic grow rate, four reaction additions double the predicted growth rate. In contrast, the predicted aerobic growth rate of *R. sphaeroides* shows only a modest improvement with each added reaction, and the predicted aerobic growth rate of *Synechocystis* barely improves after the first reaction addition.

Generally, more reaction additions are required to introduce the reference pathways than that are required to obtain an equal growth rate via CO2FIX. The only exceptions are: (i) the introduction of the rTCA cycle in *R. sphaeroides*; the four added reactions enable a slightly higher growth rate than the CO2FIX solution for three reactions, and (ii) the Calvin cycle (assuming no oxygenation) being or coinciding with the optimal reaction set in all aerobic scenarios. In addition, regardless of organism and oxygen availability, a higher predicted growth rate is obtained with five reaction additions than with any reference pathway.

Candidate pathways

We elucidated the candidate CO_2 fixation pathways consisting of both native and non-native reactions for each organism for each number of added reactions. In total, this led to the elucidation of 165 pathways differing by at least one reaction. Surprisingly, within the exactly identical pathways, only eight pathways were found for two organisms, two pathways for three organisms, and only a single pathway was found for four organisms. In addition, there is a large variation in the number of pathways identified for different organisms, and for aerobic *versus* anaerobic conditions. For instance, we find 49 and 14 pathways for *E. coli* under aerobic and anaerobic conditions respectively whereas only 6 under aerobic conditions were found for *Synechocystis* and 5 under anaerobic conditions for *G. metallireducens* (see table 6.1).

Candidate CO_2 fixation pathways have been selected for their efficient stoichiometric conversion of CO_2 to biomass. This has mostly led to the identification of ATP-efficient pathways. Additional characteristics for CO2FIX pathways are thermodynamic favorability and fast kinetics. The thermodynamic favorability of the candidate pathways has been assessed using MDF [305] for both atmospheric CO_2 concentrations (355 ppm) and those attainable in industrial gas streams (10,000 ppm). The pathway kinetics have been determined using PSA [23] based on available experimental measurements of enzyme specific activities in BRENDA [336].

TABLE 6.1: Number of identified candidate CO_2 fixation pathways per organism and number of added reactions. Values equal to 0 are not shown. *The CO_2 fixation pathway in the *T. maritima* GSM does not convert 3 CO_2 to 1 pyruvate. *[2]The *B. subtilis* GSM does not support anaerobic growth.

		Number of added reactions								
	Organism	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Aerobic	*Synechocystis*	1			2	1	2			
	R. sphaeroides	2	1	1	1	2	4	1	1	2
	E. coli	2	7	8	7	13	5	1	3	3
	P. putida		2	12	3	2	1		2	
	S. cerevisiae			2	5	1	4	1	1	1
	B. subtilis			4	4	4	1	3		
Anaerobic*[2]	*G. metallireducens*	1	1	1	2					
	R. sphaeroides	1	1	3	4	1		1		
	E. coli	1	3	7	1		1	1		
	S. cerevisiae		4	5	1	2	1			
	T. maritima	*	6	5			1	2		

Scoring the pathways relative to one another is complicated due to the interdependence of ATP-efficiency, MDF and PSA. In short, a pathway that 'wastes' a lot of ATP is thermodynamically favorable and typically faster, and in contrast an ATP-efficient pathway typically has thermodynamic bottlenecks. Nonetheless, some CO2FIX pathways compare favorably to the reference pathways on these metrics (see table 6.2) and are discussed further below.

Highlighted newly identified synthetic pathways

From the large pool of identified pathways several, often species-specific, pathways were identified. Those thermodynamically feasible pathways enable *in silico* growth, are ATP-efficient, and have attractive kinetics. Some promising novel identified pathways for CO_2 fixation are discussed here in more detail.

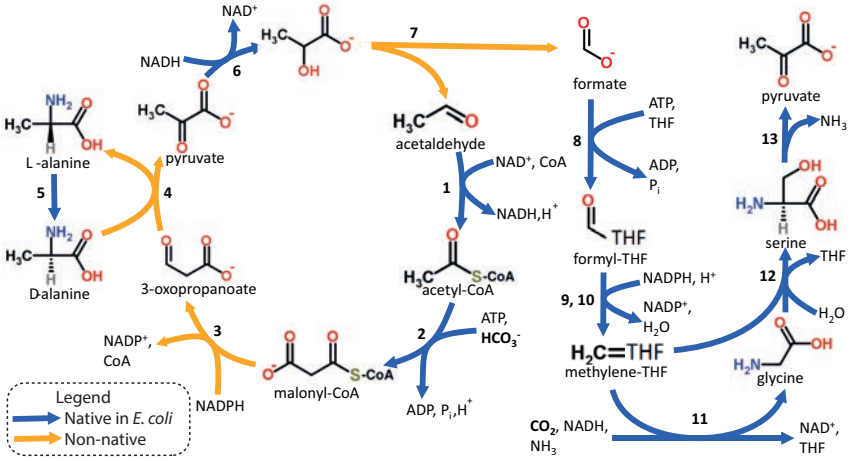


FIGURE 6.4: **The lactate aldolase pathway.** The enzymes are 1: acetaldehyde dehydrogenase, 2: acetyl-CoA carboxylase, 3: malonyl-CoA reductase, 4: beta-alanine-pyruvate transaminase, 5: alanine racemase, 6: lactate dehydrogenase, 7: lactate aldolase, 8: formate-THF ligase, 9: methenyl-THF cyclohydrolase, 10: methylene-THF dehydrogenase, 11: glycine cleavage system, 12: glycine hydroxymethyltransferase, and 13: serine deaminase.

The novel lactate aldolase pathway was identified for several organisms: *E. coli*, *B. subtilis*, *P. putida*, *Synechocystis* and *S. cerevisiae*. This relatively ATP-efficient and O_2-insensitive pathway only required 3-5 reactions in the prokaryotic models and 7 for *S. cerevisiae*. The key non-native enzyme in this pathway, lactate aldolase, splits lactate into formate and acetaldehyde. The lactate aldolase enzyme is only described in one literature report from 1986 [161], however, the enzymatic mechanism is theoretically sound and it can probably be engineered if it is not naturally available [22]. The cycle product formate can be efficiently assimilated to pyruvate via a linear reductive glycine pathway.

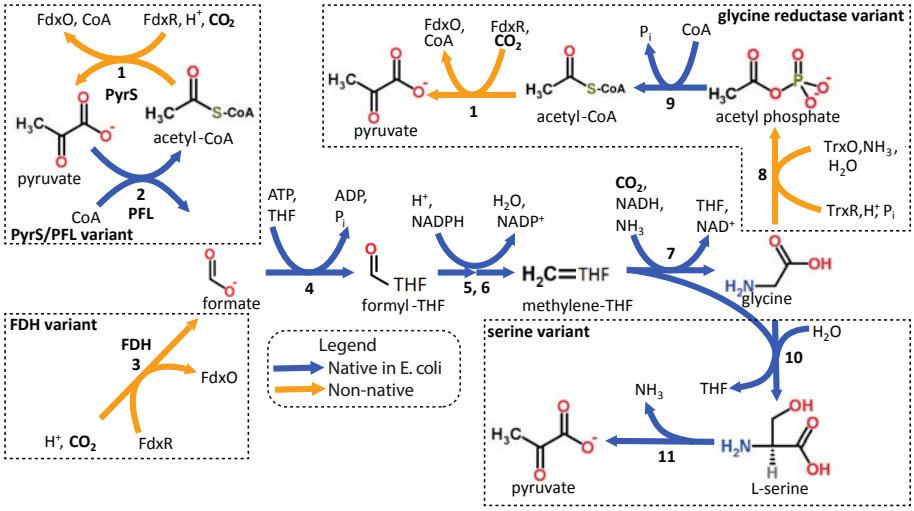


FIGURE 6.5: **Reductive glycine pathway variants.** The enzymes are 1: pyruvate synthase, 2: pyruvate formate lyase, 3: formate dehydrogenase, 4: formate-tetrahydrofolate ligase, 5: methenyl-THF cyclohydrolase, 6:methylene-THF dehydrogenase, 7: glycine cleavage system, 8: glycine reductase, 9: phosphotransacetylase, 10: serine hydroxymethyltransferase, and 11: serine deaminase. TrxO/TrxR: Oxidized and reduced thioredoxin, FdxO/FdxR: Oxidized and reduced ferredoxin.

Variants of the O_2-sensitive reductive glycine pathway are found for *E. coli*, *S. cerevisiae* and *T. maritima* (see figure 6.5). These variants require one to three non-native enzymes and have three distinct parts: (i) CO_2 is converted to formate by formate dehydrogenase, or by a cycle of pyruvate synthase (PyrS) and pyruvate formate lyase (PFL). This PyrS-PFL cycle was recently described as part of a non-autotrophic CO_2 fixation pathway in *Clostridium thermocellum* [455]. (ii) Formate is converted to glycine via a common pathway segment. The glycine cleavage system typically operates in the opposite direction, but both directions are possible [22]. (iii) Glycine is converted to pyruvate via the glycine reductase or the serine branch. The reductive glycine pathway has been previously mentioned for formate assimilation [22, 24], but not for CO_2 fixation.

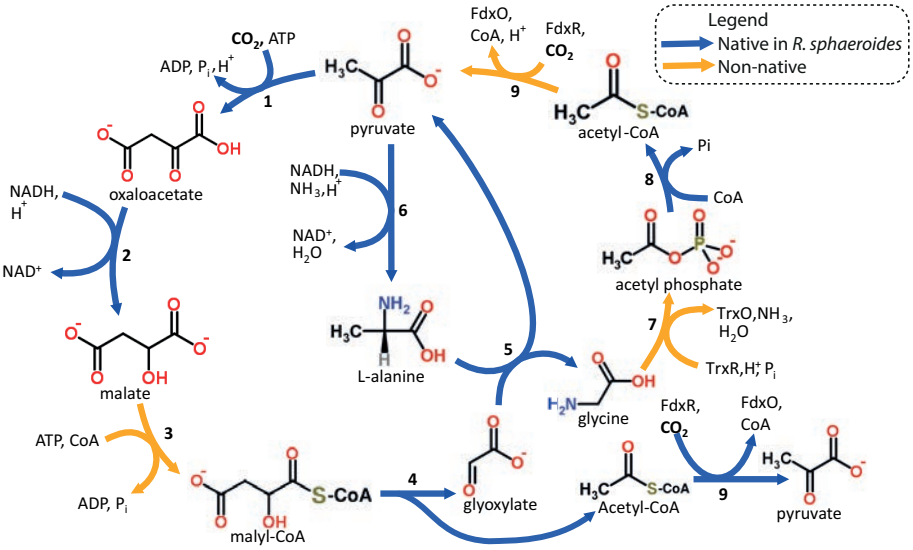


FIGURE 6.6: **The alanine-glyoxylate transaminase pathway.** The enzymes are 1: pyruvate carboxylase, 2: malate dehydrogenase, 3: malate thiokinase, 4: malyl-CoA lyase, 5: alanine-glyoxylate transaminase, 6: L-alanine dehydrogenase, 7: glycine reductase, 8: phosphotransacetylase, and 9: pyruvate synthase. TrxO/TrxR: Oxidized and reduced thioredoxin, FdxO/FdxR: Oxidized and reduced ferredoxin.

A promising CO_2 fixation pathway was identified solely for *R. sphaeroides*. This pathway, termed the alanine-glyoxylate transaminase pathway, is ATP-efficient, and has the highest MDF and PSA values of all identified and reference pathways (see table 6.2). In addition, it only requires three non-native reactions in *R. sphaeroides*.

In contrast, the aforementioned commonly identified lactate aldolase and reductive glycine pathways were not identified for *R. sphaeroides*. Although this may be related to different metabolic contexts of these organisms, it can not be ruled out that this distinction is due to a modelling artifact; the *R. sphaeroides* GSM is the only GSM not obtained from the BiGG model repository [217].

TABLE 6.2: Characteristics of selected CO2FIX and reference pathways. *Oxygen causes photorespiration. *[2] Ignoring energy-conserving electron bifurcation systems. *[3] One or more enzyme specific activities unavailable. *[4] A thioredoxin-involving reaction was lumped with a thioredoxin-regenerating transhydrogenase as formation energies for thioredoxin can not be accurately computed.

Pathway	O_2 Sensitive	MDF atm CO_2	MDF 1% CO_2	ATP	PSA
CO2FIX pathways					
reductive glycine (PyrS/PFL,serine)	yes	-0.2	2.4	2	0.75
reductive glycine (FDH,serine)	yes	2.9	5.4	2	0.82
reductive glycine (PyrS/PFL,glycine reductase)	yes	1.2*[4]	4.0*[4]	1	0.88
reductive glycine (FDH,glycine reductase)	yes	1.4*[4]	4.5*[4]	1	0.93
alanine-glyoxylate transaminase pathway	yes	2.7	6.8	2	2.18
lactate aldolase pathway	no	4.2*[4]	4.8*[4]	4	NA*[3]
Reference pathways					
Wood-Ljungdahl pathway	yes	1.6	5.1	1*[2]	NA*[3]
rTCA cycle	yes	-1.7	2.7	2	2.28
PyrS-PyrC-glx cycle	yes	3.9	6.5	3	1.42
PyrS-PEPC-glx cycle	yes	4.6	6.5	4	1.21
Calvin cycle (no oxygenation)	yes*	7.2	7.2	7	0,54
Calvin cycle (20% oxygenation)	no	4.5	6.5	10.9	0.37
3HP/4HB thaumarchaea	no	7.9	8.3	6	0.52
C4 glyoxylate cycle	no	5.6	5.6	8	0.90

Discussion

CO2FIX pathways are optimally tuned to the native metabolism of a chosen organism, as represented by a GSM. The designed pathways are weighed by the number of required non-native reactions, so that designs requiring fewer genetic modifications and optimizations for non-native enzymes, are favored. Moreover, the pathways are evaluated on their potential to sustain growth in the absence of other carbon sources; on their ATP efficiency; on their thermodynamic feasibility; and on their kinetic properties. These features impact the autotrophic growth rate of the organism as well as its potential productivity in an industrial setting. These characteristics have also been the main evaluation criteria in previous work on the design and selection of synthetic and natural CO_2 fixation pathways for biotechnological applications [23, 52, 128, 437].

The distinguishing feature of CO2FIX is the embedding of a CO_2 fixation pathway in the native metabolism of an organism. This has enabled the identification of CO_2 fixation pathways that require the metabolic network of an organism to be expanded by only a few reactions. Furthermore, CO2FIX has identified pathways for each considered organism that require at most five non-native reactions and enable higher predicted growth rates than all reference pathways (see figures 6.2 and 6.3). Although we find many pathway

variants across organisms (see figure 6.5), only eleven indistinguishable pathways were found in multiple organisms. The high fraction of identified pathways that differ between organisms strongly supports our species-specific approach.

However, this high fraction of species-specific pathways probably originates not only from biological differences which constitutes our prime interest, but also from modelling artifacts due to non-standardized naming conventions and inconsistent representations of biological processes in manually generated GSMs [171]. In principle, these modelling artifacts could be avoided through the use of GSMs that were automatically generated according to a well-defined procedure such as Model SEED GSMs [176]. However, in our experience the available automatically generated GSMs still require extensive manual curation for their practical application in this context [data not shown]. Nonetheless, the ongoing developments in the field of automatic GSM generation will ultimately enable comparisons devoid of GSM generation artifacts.

Similarly, future improvements to CO2FIX can be expected in the realms of thermodynamic and kinetic evaluations. To the best of our knowledge, MDF and PSA are the best methods currently available for theoretical determinations of thermodynamic feasibility and kinetics on a pathway level, but they rely on several assumptions that may or may not hold true. If improved or alternative methods become available, these can be easily incorporated in the modular framework of CO2FIX.

CO2FIX expansion possibilities

There are several opportunities for further expansions of the CO2FIX algorithm: (i) weights for reaction additions, (ii) direct inclusion of biotechnological production objectives, (iii) expansion of reference reaction database based on synthetic and promiscuous enzyme activities, and (iv) explicit exploitation of pathways in compartments.

CO2FIX identifies CO_2 fixation pathways that are embedded in native metabolism by adding a small predefined number of reactions to a species-specific GSM. Expansion of metabolism based on number of reactions is, however, somewhat oversimplified. Some reactions require large enzyme complexes and will require the expression of several non-native genes, whereas in other cases a single enzyme performs multiple steps in a pathway. In addition, the difficulty of expressing a non-native gene is not universal. For example, genes from closely related organisms are easier to functionally express. These pathway design considerations can, in principle, be easily incorporated into CO2FIX. In Phase 1, the MILP can be modified to not add a predefined number of reactions, but rather to spend a predefined number of 'points', where

each reaction in the reference reaction database is assigned a 'cost' depending on the expected difficulty in realizing it in the target organism.

Similarly, phase 1 of CO2FIX can be modified to not maximize the growth rate of an organism, but rather the yield of a biotechnologically interesting product. In our experience, this does not directly lead to major differences in the identified pathways as ATP-efficiency typically remains a crucial factor [data not shown]. However, the MILP in phase 1 could be replaced by a GSM-driven strain design algorithms such as OptStrain [334], which both adds reactions from a reference reaction database and deletes native reactions from a GSM in order to design a production strain.

The reference reaction database used for this work contains reactions from the BiGG [217] and Metacyc [66] reaction databases. These databases provide a comprehensive overview of characterised biochemical reactions. However, promiscuous enzyme activities - *i.e.*, side reactions - may not be extensively described in these databases, and there are many more theoretically possible reactions for which natural enzymes either do not exist or have not yet been found [121]. If a database were constructed encompassing 'plausible' or 'engineerable' enzyme activities, CO2FIX could pinpoint those activities that are promising for CO_2 fixation, which could in turn steer the rational design of enzymes.

Another way in which novel or improved CO_2 fixation pathways can be designed is through the explicit inclusion of eukaryotic compartments or bacterial microcompartments in the design. In the current implementation of CO2FIX, all reactions in the reference reaction database are cytoplasmic. In principle, the existence of compartments enables a cell to maintain distinct metabolite pools and even the co-existence of anaerobic and aerobic conditions. In other words, the pathway thermodynamics - and by extent kinetics - can benefit from the compartmentalization in the cell. This will require an explicit specification in the reference reaction database of the compartments in which reactions can be realized, and a modified thermodynamics evaluation method that handles compartmentalization and transport processes.

Ongoing work.

At the time of the writing of this thesis chapter, the development of CO2FIX is still ongoing. We are evaluating our current design choices and are in the process of fully automating the various phases of CO2FIX. In particular, the MDF calculations [305] have so far relied on manual uploading of specifically generated SBtab [256] files to the online version of Equilibrator [138] due to inconsistent results with the offline version. This manual step is part of a time-intensive iterative cycle of prediction, evaluation, and curation of the reference reaction database based on recurring thermodynamically infeasible

cycles. Ideally, all CO2FIX phases are fully automated such that any pathway not satisfying predefined requirements are directly filtered out. In contrast, for example, the optimal growth rates that are currently presented in figures 6.2 and 6.3 correspond to all candidate pathways, rather than only to the promising CO2FIX pathways.

The CO2FIX pathways contain a wealth of diversity in both truly distinct pathways, as well as variations on the same pathway. This diversity of CO2FIX pathways has not yet been fully explored, and even more promising pathways may remain unexposed. In addition, our current pathway identification procedure focuses on pathways that convert exactly 3 units of CO_2 to 1 unit of pyruvate, effectively ignoring pathways with obligate by-products. Although CO_2 fixation pathways with by-products are most likely not industrially relevant, there may be surprisingly interesting pathways omitted by the current procedure.

Conclusions

The herein developed method, CO2FIX, designs species-specific CO_2 fixation pathways that require a limited extension of native metabolism, while being ATP-efficient, thermodynamically feasible, and kinetically attractive. Our prime design directive, embedding in native metabolism, is in line with the attained, recent, *in vivo* experimental success on realizing a synthetic CO_2 fixation pathway in *E. coli* [13]. Therefore, we expect CO2FIX designs to contribute to ongoing efforts aimed at enabling CO_2 fixation in heterotrophs and enhancing CO_2 fixation in autotrophs. These efforts will ultimately culminate in biochemical production processes that are both green and efficient.

Materials and Methods

Genome-scale metabolic models. iJO1366 (*E. coli*)[319], iMM904 (*S. cerevisiae*) [282], iJN678 (*Synechocystis*) [301], iJN746 (*P. putida*) [298], iYO844 (*B. subtilis*) [312], iAF987 (*G. metallireducens*) [132], and iLK478 (*T. maritima*) [466] were downloaded from the BiGG model database [217]. iRsp1095 (*R. sphaeroides*) [187] was obtained from the supplementary files of the corresponding paper.

***In silico* media conditions.** The lower flux bound of all exchange reactions for carbon sources except for CO_2 were set to 0 [mmol/gdw/h]. The lower bound of the CO_2 exchange reaction was set to -1000 [$mmol\ gdw^{-1}\ h^{-1}$]. All other lower and upper bounds of exchange reactions were left to their default values.

Addition of anoxygenic photosystem. Autotrophic growth requires a mechanism to generate ATP from either light or an inorganic electron donor. Therefore, a reaction representing an anoxygenic photosystem ($H^+_{cytoplasm} \rightarrow H^+_{extracellular}$) was added to the GSMs iJO1366, iMM904, iJN746, iYO844, iLJ478. The upper bound of this reaction was set to 100 [$mmol\ gdw^{-1}\ h^{-1}$], which corresponds to a photon uptake rate of 50 [$mmol\ gdw^{-1}\ h^{-1}$] assuming two transported protons per absorbed photon. These values lie within the capabilities of the anoyxgenic photosystem of *R. sphaeroides* [188].

Addition of H_2 hydrogenases. Autotrophic growth requires a mechanism to generate reducing equivalents (NAD(P) and reduced ferredoxin) using an external electron donor. Therefore, three H_2 hydrogenase reactions were added to the GSMs that did not yet contain these:

- $2.0\ H_2 + NADP + oxidized\_ferredoxin \rightarrow 3.0\ H^+ + NADPH + reduced\_ferredoxin$
- $NADP + H_2 \rightarrow NADPH + H^+$
- $NAD + H_2 \rightarrow NADH + H^+$

In addition, transport and exchange reactions for H_2 were added to the GSMs not having these reactions. The lower bound of the exchange reaction was set to -100 [$mmol\ gdw^{-1}\ h^{-1}$].

CO2FIX. CO2FIX consists of four separate parts: (i) a Mixed-Integer Linear Programming (MILP) algorithm adds a pre-defined number of reactions to a GSM from a reference reaction database to maximize the biomass production using CO_2 as a carbon source. (ii) a Linear Programming (LP) algorithm is used to identify all newly added and native reactions that are part of the candidate CO_2 fixation pathway. (iii) The CO_2 candidate fixation pathway is subjected to a thermodynamic feasibility test using Max-min Driving Force (MDF) [305]. (iv) The candidate CO_2 fixation pathway is subjected to a kinetic feasibility test using Pathway Specific Activity (PSA) [23]. The separate steps are discussed in more detail below.

Construction of reference reaction database The construction of the reference reaction database consists of nine steps:

i All reactions from GSMs originating from BiGG [217]: iJO1366 [319], iMM904 [282], iJN678 [301], iJN746 [298], iYO844 [312], iAF987 [132], and iLK478 [466] are combined in a model M_{ref}.

ii M_{ref} is expanded by adding all MetaCyc [66] reactions as available from MetaNetX.org [150].

iii M_{ref} is further expanded by addition of all reactions corresponding to the reference CO_2 fixation pathways.

iv Reactions in M_{ref} were converted to the MNXref namespace [36] via direct mapping of metabolite identifiers (see Namespace conversion).

v Reactions from M_{ref} that contain metabolites that can not be converted into the KEGG namespace are deleted, as conversion of all metabolites to the KEGG namespace is crucial for the thermodynamics calculations via EQuillibrator [138].

vi Reactions representing biomass production, cross-membrane transport, and respiration, as well as elementally unbalanced reactions are removed from M_{ref}

vii The compartmentalization of all metabolites in all reactions in M_{ref} is set to cytoplasmic.

viii M_{ref} is curated by identifying sets of duplicate reactions and removing all but one.

ix Directionality of all non-BiGG reactions in M_{ref} is set according to their $\Delta_r G'^0$ values as calculated via EQuillibrator:

- $\Delta_r G'^0 \leq -20$: Reaction is irreversible; forward direction

- $-20 \leq \Delta_r G'^0 \leq 20$: Reaction is reversible
- $20 \leq \Delta_r G'^0$: Reaction is irreversible; backwards direction

Reactions from BiGG are specifically curated for their application in GSMs. Therefore, we assume their pre-set directionalities are appropriate.

CO2FIX phase 1 - Generating CO_2 fixation pathways. A MILP is used to add a predefined number of reactions from the reference reaction database to a GSM to maximize the biomass production using CO_2 as a carbon source:

$$\begin{array}{lll} \text{Maximize} & x_{biomass} & \\ \text{Subject to} & S\mathbf{x} = \mathbf{0} & \\ & lb_j \leq x_j \leq ub_j & \forall j \in Model \\ & lb_j y_j \leq x_j \leq ub_j y_j & \forall j \in Database \\ & y_j \in \{0,1\} & \forall j \in Database \\ & \sum y_j = N & \forall j \in Database \end{array}$$

where $x_{biomass}$ is the flux through the biomass synthesis reaction, **x** is the flux vector, S is the stoichiometric matrix covering both the model and database reactions, **0** is a null vector, **lb** and **ub** are the lower and upper bound vectors respectively, **y** is a boolean vector indicating the reactions added from the reference reaction database, and N indicates the number of reactions to be added from the reference reaction database. Integer cuts were added to the algorithm in order to find additional solutions. For each N we determined the five best solutions. Subsequently, we ensured that the best solutions were identified by the MILP by setting the lower bound of $x_{biomass}$ to the highest biomass production found for the selected N and repeating the procedure; occasionally better solutions were identified. All solutions corresponding to at least 40% of the maximally found growth rate for the selected N were stored. The implementation of CO2FIX phase 1 was based on the Growmatch [235] implementation as available as part of the CobraPy Toolbox [112].

CO2FIX phase 2 - Isolating CO_2 fixation pathways. The reactions that were added in CO2FIX phase 1 have combined with native reactions to form a CO_2 fixation pathway. In this step, all reactions that are part of the pathway are identified along with their relative stoichiometric coefficients, which are required for CO2FIX phases 3 and 4. We define a CO_2 fixation pathway as the set of reactions that converts 3 units of CO_2 to 1 unit of pyruvate, while excluding reactions related to transport, cofactor regeneration, and ATP generation. CO2FIX phase 2 constitutes of 4 steps:

i Five reactions are added to the GSM to provide an unlimited ability to generate ATP and to regenerate the cofactors NADPH, NADH, ubiquinol, and reduced ferredoxin. These reactions are:

- $ADP + Pi + H^+ \rightarrow ATP + H_2O$
- $NADP + H^+ \rightarrow NADPH$
- $NAD + H^+ \rightarrow NADH$
- $quinone + H^+ \rightarrow quinol$
- $oxidized\_ferredoxin \rightarrow reduced\_ferredoxin$

ii Each reaction in the GSM and the reference reaction database is modified so that flux is only carried in the forward direction. Backward reactions are reversed, and reversible reactions are split in two.

iii We use pFBA [249] to identify all reactions and corresponding fluxes that are required for the conversion of 3 [$mmol\ gdw^{-1}\ h^{-1}$] CO_2 into 1 [$mmol\ gdw^{-1}\ h^{-1}$] pyruvate:

$$\begin{array}{lll} \text{Minimize} & \sum_j x_j & \forall j \in Model \\ \text{Subject to} & S\mathbf{x} = \mathbf{0} & \\ & \mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub} & \\ & x_{pyruvate} = 1 & \\ & x_{CO_2} = -3 & \end{array}$$

where **x** is the flux vector, S is the stoichiometric matrix, **0** is a null vector, **lb** and **ub** are the lower and upper bound vectors respectively, and $x_{pyruvate}$ and x_{CO_2} are the fluxes through the pyruvate and CO_2 exchange reactions.

iv Reactions with positive flux values (positive components of **x**) are selected. Reactions corresponding to transport, cofactor regeneration, and ATP generation are filtered out. These reactions constitute the candidate CO_2 fixation pathway and the values in **x** are their stoichiometric pathway coefficients.

CO2FIX phase 3 - Thermodynamic feasibility. CO2FIX generates SBtab [256] pathway files which can be directly used for Max-min Driving Force (MDF) [305] calculations using the online implementation of EQuillibrator [138] as available at `http://equilibrator.weizmann.ac.il/pathway`. For several metabolites and metabolite pairs we have fixed the concentrations or concentration ratios: [Pi] = 10 mM, [PPi] = 1 mM, [ATP]/[ADP] = 10, [ADP]/[AMP] = 1, [NADPH]/[NADP+] = 10, [NADH]/[NAD+] = 0.1,

[ferredoxinred]/[ferredoxinox] = 10, based on [23, 437]. Different from [23, 437], we varied the CoA concentrations between 0.1-5 mM based on the uncertainty in CoA concentrations measured in *E. coli* [34, 162]. Concentrations for CO_2 and HCO_{3-} were fixed at 0.012 mM and 1.6 mM to simulate atmospheric (355 ppm) CO_2 concentrations, and at 0.34 mM and 4.5 mM to simulate an industrial gas sparged setting with 1% CO_2. The concentrations of all other metabolites were allowed to vary between 1 μM and 10 mM, representing a physiological range [305]. All MDF calculations were performed based on a pH of 7.5 and an ionic strength of 0.1 M.

CO2FIX phase 4 - Kinetic feasibility. The calculation of the Pathway Specific Activity (PSA) [23] for a pathway involves a 4-step process:

i EC numbers of all pathway reactions are obtained via the mapping available from MNXref [36] and BiGG [217] if possible. All remaining reactions are manually assigned the appropriate EC number.

ii Available experimentally measured specific activities for each EC number are obtained from BRENDA [336] via the SOAP API.

iii Enzyme specific activity is calculated for each enzyme in the pathway by discarding the lowest 50% and highest 10% of the specific activities for each EC number and averaging the remaining values, as in [23, 437].

iv PSA is calculated according to the formula:

$$PSA = 1/\sum_{i=1}^{m} \frac{w_i}{V_i}$$

where m is the number of enzymes in the pathway, V_i is the specific activity of enzyme i [$\mu mol\ min^{-1}\ mg^{-1}$], and W_i is the required flux through enzyme i to produce 1 μmol of pyruvate.

Namespace conversion Metabolites and reactions were converted to the MNXref [36] namespace from the BiGG [217], MetaCyc [66] and KEGG [201] namespaces, as well as from the MNXref namespace to the KEGG namespace. The MNXref namespace is a specifically designed reference namespace with direct links to other commonly used namespaces. We downloaded the MNXref namespace from http://www.metanetx.org/mnxdoc/mnxref.html [150] on 16 September 2016. We converted the metabolites from one namespace to another via the direct mapping between metabolite identifiers available for the MNXref namespace.

Manual curation of GSMs and reference reaction database. The GSMs and reference reaction database were manually curated to circumvent several recurring issues. These issues include, for example, unlimited energy generation, biologically inaccurate redox regeneration, and the inability of EQuilibrator to perform a ΔG estimation. These reactions were manually removed or made irreversible in the GSMs and reference reaction database.

Updating reaction directionalities in GSMs Reaction directionalities of all reactions in each GSMs were updated according to the directionalities in the reference reaction database. Specifically, if a GSM reaction was present in the reference reaction database, but with a different directionality, the directionality in the GSM was changed to that of the reference reaction database.

Aerobic simulations: Blocking oxygen-sensitive reactions. Several cellular processes are oxygen-sensitive and must thus be blocked during aerobic simulations. We manually looked up reported oxygen sensitivity/tolerance of all enzymes involving the oxygen-sensitive metabolites ferredoxin, flavodoxin, thioredoxin, and 'corronoid iron sulfur protein'. Unless such an enzyme is specifically reported to be oxygen-tolerant, the corresponding reactions were blocked for aerobic simulations. In addition, reactions corresponding to pyruvate formate lyase and the proton-translocating Rnf complex were also blocked.

Implementation and simulation. All computational simulations and analyses were performed in Python. The GSMs were loaded and analysed using the CobraPy Toolbox [112]. The IBM CPLEX solver [92] was used for all linear and mixed-integer linear programming problems. The online implementation of EQuillibrator [138] at `http://equilibrator.weizmann.ac.il/pathway` was used for the MDF calculations.

Acknowledgements

We thank Niels Zondervan for his help with obtaining enzyme specific activities from BRENDA via the SOAP API, Elad Noor for his advice and help with EQuillibrator, and Nhung Pham for her comments on the manuscript.

Author contributions

Conceived and designed the study: RGAvH NC.
Performed the experiments: HvK.
Analyzed the data: Hvk NJC RGAvH MSD.
Wrote the manuscript: RGAvH NJC MSD.

Funding

We gratefully acknowledge financial support from the Wageningen university IPOP project, the European Union Horizon2020 project EmPowerPutida (Project reference: 635536), the Netherlands Organization for Scientific Research via the SIAM Gravitation Grant to Willem M. de Vos (Project reference: 024.002.002). The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

Chapter 7

Metabolic modeling for algae biotechnology

Adapted from:

Maarten J. M. F. Reijnders, **Ruben G. A. van Heck**, Carolyn M. C. Lam, Mark A. Scaife, Vitor A. P. Martins dos Santos, Alison G. Smith, and Peter J. Schaap. "Green genes: bioinformatics and systems-biology innovations drive algal biotechnology". In: Trends in Biotechnology, 32(12) 2014.

Abstract

Many species of microalgae produce hydrocarbons, polysaccharides, and other valuable products in significant amounts. However, large-scale production of algal products is not yet competitive against non-renewable alternatives from fossil fuel. Metabolic engineering approaches will help to improve productivity, but the exact metabolic pathways and the identities of the majority of the genes involved remain unknown. Recent advances in bioinformatics and systems-biology modeling coupled with increasing numbers of algal genome-sequencing projects are providing the means to address this. A multidisciplinary integration of methods will provide synergy for a systems-level understanding of microalgae, and thereby accelerate the improvement of industrially valuable strains. In this review we highlight recent advances and challenges to microalgal research and discuss future potential.

Diversity of microalgae and their biotechnological potential

Microalgae are simple photosynthetic eukaryotes that are among the most diverse of all organisms. Microalgae inhabit all aquatic ecosystems, from oceans, lakes, and rivers to even snow and glaciers, as well as terrestrial systems including rocks and other hard surfaces. Microalgae exhibit significant variation in physiology and metabolism, a reflection of the high level of genetic diversity that exists between different phyla owing to multiple endosymbiotic events, horizontal gene transfer, and subsequent evolutionary processes, producing a polyphyletic collection of organisms [105, 274]. Given this diversity, mining the genomes of these organisms provides a great opportunity to identify novel pathways of biotechnological importance. In particular, microalgae are of considerable interest for the synthesis of a range of industrially useful products, such as hydrocarbons and polysaccharides [50, 381], owing to rapid growth rates, amenability to large-scale fermentation, and the potential for sustainable process development [450].

Algae as a source of biofuel molecules, such as triacylglycerides (TAGs), the precursor for biodiesel [275], have been a focus in recent years, with potential yields an order of magnitude greater than competing agricultural processes [258]. Evaluations of current technologies demonstrate that microalgae are commercially feasible for biofuel production, but are not yet cost-competitive with petroleum products [196, 330], the metric upon which commercial success ultimately lies. For example, the net energy input versus output for large-scale algae biodiesel production was estimated to be 1.37, compared to 0.18 for conventional/low-sulfur diesel [330]. Currently, for microalgae to synthesize TAG it is necessary to expose them to stress conditions such as nutrient limitation, which reduces growth and increases energy dissipation. The trade-off between biosynthesis of TAG and cell growth is therefore a severely limiting factor [222]. If a better understanding of the metabolic and regulatory networks were available, they could be rewired for increased TAG synthesis, with fewer drawbacks than for existing algal cells.

The production of other interesting algal products will also benefit from a better understanding of microalgae at a systems level. For example, polysaccharides such as starch and cell wall materials can be used for biotechnological applications [61]. These carbohydrates can be degraded to fermentable sugars for bioethanol production [181], or serve as chemical building blocks for renewable materials, but the composition and proportions of the different sugar components require optimization. Similarly, various valuable secondary metabolites produced by microalgae are of interest in the food, nutrition, and cosmetics industries [50], but often they are produced in trace

amounts, or only under conditions that are not amenable to industrial cultivation.

Over 30 microalgal genomes have been sequenced, and numerous transcriptomics, proteomics, and other systems-biology studies have been performed. Nevertheless, our understanding of metabolic pathways within these microalgae remains limited [180]. Significant knowledge gaps need to be filled between omics data, the annotation thereof, and our systems-level understanding. This will allow the conversion of these resources into usable Genome-Scale Metabolic models (GSM) and provide the basis for effective metabolic engineering, synthetic biology and biotechnology. We consider here the potential application of advanced methods to improve the functional annotation of algal omics data, to increase the resolution of GSMs, and ways to integrate available computational methods for effective exploitation of microalgae in biotechnology.

Annotation challenges for microalgae

The nuclear genome of the green alga *Chlamydomonas reinhardtii*, sequenced in 2007 [274], is approximately 120 Mb and comprises some 15 000 genes. Although *C. reinhardtii* is commonly used as a reference for the annotation of other microalgae, only a subset of $\sim$50 proteins have an experimentally validated function according to the UniProt database (http://www.uniprot.org), compared to 6800 proteins for the model plant *Arabidopsis thaliana*. Consequently, most *C. reinhardtii* genes have been computationally annotated by inferred homology with *A. thaliana*, and other plant species and microbes [274], using BLAST (Basic Local Alignment Search Tool) or family-wise alignment methods such as HMMER and InterProScan (table 7.1). BLAST-based methods often use the principle of one-to-one recognition, meaning that annotation of a query gene is based on the annotation of a single known gene. This limits the success rate for recognition and correct functional annotation of the more distantly related *C. reinhardtii* genes, but becomes even more problematic when the in silico-derived functional annotation of *C. reinhardtii* is subsequently used for annotation of other algal species. This is because, owing to a lack of common ancestry, two algal species can be more diverse than, for example, any two plant species. Therefore, these methods, which are highly suitable for high-throughput analysis because of their simplicity, are less appropriate for accurate in-depth annotation of algal genomes. In the CAFA (Critical Assessment of protein Function Annotation) experiment [347], the accuracy of more advanced functional annotation algorithms was assessed. The CAFA concluded that 33 of 54 tested functional annotation algorithms

outperformed the standard BLAST-based method (table 7.1). The substantial improvement can be explained by the fact that these second-generation methods do not apply the one-to-one recognition principle but, to increase their success rate, use instead a one-to-many recognition strategy and/or include context-aware principles for annotation. An example is Argot2 (Box 1) [124], which applies the one-to-many recognition strategy by calculating the statistical significance of all candidate homologous genes found by BLAST [7] and HMMER [136], combined with an assessment of semantic similarities of associated GO terms. In a context-aware multilevel approach, annotation is not merely based on sequence similarity, but other factors such as protein–protein interactions [228], transcript expression patterns [228], phylogenetic trees [118], compartmentalization information [91], and literature [452] are also taken into account. FFPred2 from UCL–Jones [91] is the prime example of such a homology-independent functional annotation algorithm.

TABLE 7.1: Features of commonly used functional annotation tools

Methods	Success rate*	Speed	Availability	Additional notes	Refs
BLAST	Limited	Fast	Online/offline	Dependent on global sequence similarity for success. Suitable for high-throughput analysis	[7]
HMMER	Moderate	Fast	Online/offline	Family-wise alignment method. Suitable for high-throughput analysis	[136]
InterProScan	Moderate	Slow	Online/offline	Family-wise alignment method. Uses precomputed protein domains	[194]
FFPred2	High	Slow	Limited online/offline	Algorithms currently trained on non-algal datasets. Not suitable for high-throughput analysis	[57, 91]
Argot2	High	Moderate	Limited online	Initial selection is dependent on BLAST and HMMER output. Additionally predicts compartmentalization. User-friendly interface	[124]

*For distantly related sequences.

Advanced multilevel annotation methods effectively increase the recall of function prediction while maintaining an acceptable precision. The challenge in genomic annotation for microalgae lies in the small number of experimentally validated algal genes and the lack of algae-specific contextual data such as protein interaction and compartmentalization data. This results in a relatively low number of genes that are predicted to have a specific biological function. To overcome this, multiple annotation methods and data sources should be combined. The combined result increases the number of annotated genes, while a consensus prediction among the different methods improves the accuracy of the annotation [360]. Owing to their simplicity and speed, first-generation methods can be used for initial high-throughput analysis of a large set of genes. Second-generation methods can then be used for a refined analysis of these genes. However, to utilize these advanced methods fully, a significant amount of experimentally determined contextual data is required.

Although increasing amounts of gene expression data are being generated, little structural and protein interaction data are being generated for algae. In the absence of such experimental facts it is still possible to generate this contextual information by in silico prediction methods [57, 367], but whilst studies have shown that this is a feasible option [144], caution is necessary because there is a high risk of error propagation.

Apart from functional annotation it is also important to establish the cellular location of a protein. For this there are several tools available, including Argot2 (Box 1) [124], TargetP [116], SignalP [332], PSORTb [289], and PredAlgo [408]. The last is a tailor-made multi-subcellular localization prediction tool dedicated to three compartments of green algae: the mitochondrion, the chloroplast, and the secretory pathway. However, owing to the limited number of algal proteins with a known cellular localization, which can be seen for example from the quantitative subcellular localization of roughly 80 proteins [448], or the collection of roughly 1000 chloroplast-localized proteins from *C. reinhardtii* [412], the algorithm is trained with a relatively small *C. reinhardtii* dataset [408]. This raises questions regarding reliability for other algal species because the polyphyletic nature of different microalgae means some algal species are distantly or not related, and this can result in a different subcellular localization of homologs. Therefore it is advisable to use PredAlgo in combination with non-algal-specific tools in a similar way as for functional annotation.

To support large-scale annotation of algal sequence data, up-to-date databases and readily available supporting tools are required. Online databases provide the means to share data easily such that the scientific community can profit as a whole. Supporting tools can assist in annotating genes, pathways, and performing statistical analysis. While genomic data for various algae are available in NCBI and UniProt, the amount of public data is lagging behind in comparison to plant and bacterial species. In addition, tools and databases that do more than storing the available sequencing data are needed. A small number of tools are available, although these are often limited to *C. reinhardtii*. One such tool is ChlamyCyc [270], a *C. reinhardtii*-specific pathway/genome database of the MetaCyc [66] facility for metabolic pathway analysis. A peptide database, ProMEX, is available that contains over 2000 *C. reinhardtii* peptides which are usable for proteomics analysis [449]. In addition, the Augustus tool, which is commonly used for prediction of eukaryotic genes [206], has a tailor-made section for *C. reinhardtii*. Finally, the Algal Functional Annotation Tool [254] incorporates annotation data for a few microalgal species from several pathway databases, ontologies, and protein families. Broadening the scope of these annotation tools for a range of microalgae would allow comparative analysis, which is useful for easy mapping of various differences between microalgae. In this context, a useful tool which has

been applied to plant research is Phytozome (http://www.phytozome.net) [156], a comparative hub for analysis of plant genomes and gene families. It acts as a reference for the key data of many plant species, and provides click-to-go features such as BLAST and summaries key data. Phytozome has grown to be a major asset to the plant science community. Although it contains data from a few green algae, an expanded web-portal focused on algal systems-bioinformatics research could be of immense benefit to the field, particularly for those studying the more industrially relevant diatoms and heterokont species (table 7.2). Such a web-portal would provide access to new and existing tools specifically useful for algal species and facilitate exposure to a broad audience. In addition, it could act as a hosting platform for small but useful tools such as a refined algal literature research algorithm and tools that suggest genes to fill gaps in metabolic or regulatory pathways for microalgae. Adopting an algal web-portal would provide a good overview of all available data and tools, and help to reduce the redundancy that is often seen in biology and bioinformatics.

Box 1. Argot2
One of the top performers in the CAFA experiment is Argot2 (annotation retrieval of gene ontology terms) [124]. It stands out in terms of simplicity, as well as by incorporation of BLAST and HMMER. Argot2 combines an easy interface with multilayer analysis, making it a perfect starting point for biologists wishing to annotate their data.
Argot2 requires a nucleotide or protein sequence as input. It queries the UniProt and Pfam databases using BLAST and HMMER respectively, providing an initial high-throughput sequence analysis. A weighting scheme and clustering algorithm are then applied to the results to select the most accurate gene ontology (GO) terms for each query sequence. The user can choose to perform this entire process online at the Argot2 webserver, limited to one hundred sequences per query. Alternatively, if the BLAST and HMMER steps are performed locally and provided to the webserver, over 1000 sequences can be submitted per query. After the analysis is completed, which can take several hours depending on the amount of input data, the user is provided with the prediction results as well as the intermediate BLAST and HMMER files. These predictions include molecular function, biological processes, and cellular component GO terms for each query. Predicted GO terms are ranked by a score based on statistical significance and specificity. Optionally, the user can choose to compute protein clusters based on functional similarity.

TABLE 7.2: List of selected industrially useful microalgae

Species	Genome size[a] (Mb)	Available proteins[b]	Reported industrially relevant characteristics[a]	Refs
Chlamydomonas reinhardtii	120	15 144	Model system for unicellular green algae	
Monoraphidium neglectum	68	16 761	Up to 21% dry weight neutral lipid under nitrogen starvation	[46]
Nannochloropsis gaditana *Nannochloropsis oceanica*	34 28	15361 242	Can produce high amounts of ω-3 long-chain polyunsaturated fatty acids Up to 50% dry weight oil content	[381]
Phaeodactylum tricornutum	27	10 673	Can produce antibacterial fatty acids (9Z)-hexadecenoic acid (palmitoleic acid; C16:1 n-7) and (6Z, 9Z, 12Z)-hexadecatrienoic acid (HTA; C16:3 n-4)	[104]
Chlorella variabilis	46	9831	The first sequenced *Chlorella* genome. A model genome for understanding other *Chlorella* species	[368]
Ostreococcus tauri	12.6	9050	Smallest sequenced microalgal genome with simple cellular structure	
Chlorella protothecoides	22.9	7039	Up to 55% dry weight lipid content in heterotrophic growth. Highest published biomass yield, average 3.37 gDW L^{-1} h^{-1} in heterotrophic growth	[107, 151, 381]
Chlorella vulgaris	N.a.[c]	292	Up to 42% lipid content in photobioreactor with artificial waste water. Up to 26% total lipid in dry weight in heterotrophic growth	[134, 368]
Dunaliella salina	N.a.	238	Up to 10% carotenoids in dry weight; 90% β-carotene in carotenoids	[340]
Haematococcus pluvialis	N.a.	60	Highest reported yield of antioxidant astaxanthin (3.8% dry weight)	[8]
Botryococcus braunii	~166–211	30	Up to 57% total lipids in dry weight. Contains exopolysaccharides	[405, 442, 443]
Neochloris oleoabundans	N.a.	0	Up to 56% total fatty acids in dry weight under nitrogen-deprivation	[222]

(a) Genome size or characteristics are according to NCBI unless otherwise specified. **(b)** Estimated protein numbers are according to UniProt unless otherwise specified. **(c)** N.a., not available.

Understanding algal metabolism at a systems level

The sheer number of genes for metabolic enzymes, combined with the complexity of cellular metabolism, means that it is not straightforward to establish metabolic capability, even for well-annotated species. This limitation has led to the development of metabolic models which represent a snapshot of metabolism of an organism in a network format. Once an annotated algal genome or transcriptome is available, a corresponding Genome-Scale Metabolic model (GSM) can be reconstructed and the topology of the metabolic network of the algal species can be analyzed. An initial draft model can be generated directly from the genomic annotation, and is then adjusted and expanded based on experimental data, literature, and gap-filling procedures. The final GSM then includes all reactions the alga is known to perform as well as the associated genes and constraints, for example, reaction directionalities and rate limits. Owing to their comprehensive representation of metabolism, GSMs form the basis for a large and diverse set of mathematical methods for predicting metabolic behavior. These methods include the

widely employed flux balance analysis (FBA) [318] and flux variability analysis (FVA) [261], but also methods integrating fluxomic, transcriptomic, or proteomic data (Box 2) [318]. For an extensive overview of mathematical methods using metabolic models we refer to Zomorrodi *et al.* [472]. We focus here on recent developments in the modeling of microalgae specifically.

GSMs of microalgae reflect the modeling counterpart of their current annotation; therefore, inconsistencies between GSM predictions and experimental findings indicate missing and/or poor annotations. For example, experimentally identified metabolites were compared to metabolites that could be produced in metabolic reconstructions of *C. reinhardtii* [80, 269] (figure 7.1). Metabolites found experimentally but not in the models initiated pathway elucidation and identification of the corresponding genes, and thereby led to an improved genomic annotation [269]. This procedure was automated by Christian *et al.* who designed a gap-filling method to identify reactions allowing production in a model of experimentally detected metabolites [80]. These updated reactions and annotations [80, 269] were subsequently stored in ChlamyCyc [270], allowing continuous expansion of the database. Concurrently, a separate *C. reinhardtii* GSM, iAM303, was created in which the included open reading frames were experimentally validated. This led both to improved structural genomic annotation and to additional support for the reactions included in the model [263]. This GSM was greatly expanded in iRC1080 in 2011 and additional ORFs were validated [71]. The predictive power of the latter GSM was tested for 30 environmental conditions and 14 gene knockouts. In addition, iRC1080 predicted essential genes (lethal phenotype upon knockout) under different experimental conditions, although these predictions remain to be validated [71]. Recently GSMs for *Ostreococcus tauri* and *Ostreococcus lucimarinus* have been constructed [231] (figure 7.1), demonstrating expansion in the field. The initial GSMs, based on the available gene annotations, revealed that these could not account for the production of many biomass constituents [231]. The gap-filling method designed in [80] was subsequently employed to find suitable reactions for the production of these metabolites [231].

Box 2. Flux analysis in microalgae part 1
Flux Balance Analysis (FBA) [318] is the most commonly applied method to simulate metabolism in GSMs. It identifies a theoretically optimal use of metabolic capabilities for a selected metabolic objective in a specific environment. Because some microalgae can grow autotrophically in chemically defined medium, the boundary conditions for consumption of all medium components are well specified in those cases. This is advantageous for in silico metabolic flux analysis using GSMs to address, for example, how a microalga can achieve maximal growth under defined illumination. In addition, disabling the metabolic capabilities associated with a gene allows simulation of mutant strains. FBA can thus assess the potential of different strains and different environmental conditions. To run FBA, all reactions are organized in a stoichiometric matrix S. Each column in S represents a different reaction, and each row a different metabolite. A nonzero value at position [i,j] thus indicates the stoichiometric coefficient of metabolite i in reaction j. FBA then employs two different constraints. (i) Metabolism is assumed to be in steady-state; production/degradation of intermediate compounds is not possible, and (ii) thermodynamics (reversibility) and substrate availability both dictate lower and upper flux bounds for individual reactions. Finally, one or more reactions are selected to represent the metabolic objective, for example, algal biomass production. Together, the S matrix, the constraints, and the objective function form a linear programming problem:

$$\begin{array}{ll} \text{Maximize} & \mathbf{c'x} \\ \text{Subject to} & S\mathbf{x} = \mathbf{0} \\ & \mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub} \end{array}$$

where **x** is the flux vector, **c** is the objective vector, **0** is a null vector ensuring steady-state, and **lb**/**ub** are the lower/upper bounds for each reaction. The vector **x** represents a flux distribution with the theoretically maximal value for the metabolic objective. However, because of the presence of alternative/cyclic pathways, there are often alternative flux distributions with equally high values for the objective function.

Box 2. Flux analysis in microalgae part 2
Flux Variability Analysis (FVA) [261] explores for each reaction to what extent the flux can vary while permitting only a small reduction in the obtained objective value. In addition, experimental data can be used to provide additional constraints. For example, 13C-labeling experiments provide experimentally measured fluxes as inputs for the model simulations [444, 447]. Several FBA-based methods also facilitate the integration of transcriptomic, proteomic, and metabolomic data with metabolic models to constrain reactions based on measured RNA or protein levels [190, 462]. Thereby, flux distributions are identified which are most consistent with the expression data [42]. Because of the greater number of quantitative genome-wide transcriptomic studies compared to those analyzing the proteome, applications using transcriptomic data have been relatively more abundant. However, the methods generally do not distinguish between these two types of data, and metabolic models can therefore be integrated with, and their predictions compared to, experimental data yielding new insights into metabolic functioning.

It is well recognized that the exact choice of growth conditions is highly important in attaining desired metabolic activities. GSMs can explore how different growth conditions affect metabolism and can identify theoretically optimal conditions for a given metabolic objective. For example, multiple GSMs of *C. reinhardtii* were used to simulate metabolism under autotrophic, heterotrophic, and mixotrophic conditions to verify model predictions [315], to investigate how metabolite production is influenced [51, 315], and to contrast mutant strains [71]. *C. reinhardtii* GSMs were also used to determine how the quantity of light [71, 86, 220] and its spectral composition [71] affect metabolism. Of particular interest is the possibility to predict an optimal light spectrum for a given metabolic goal [71]. In contrast to these successful GSMs of *C. reinhardtii*, the metabolism of other algae is only poorly understood. For example, some industrially relevant algae can currently not be grown efficiently without bacterial presence [365]. Potentially, these algae and associated bacteria can be modeled simultaneously to deduce their relationship, as has been done for other microbial communities [470, 471].

The most comprehensive algal GSMs to date are iRC1080 [71] and AlgaGEM [315], which account for various cellular compartments. However, they vary in degree of compartmentalization (figure 7.1). In iRC1080, half (865/1730) of the non-transport reactions occur in cellular compartments other than the cytosol. By contrast, this is only about 12% (201/1617) for AlgaGEM. This reflects the fact that independently generated GSMs for the

same organism can differ significantly in their representation of metabolism because different sources of information are included. By combining the information from all currently available *C. reinhardtii* GSMs, as well as from improved annotation methods, a single and more-comprehensive GSM may be obtained. This consensus *C. reinhardtii* GSM would be an important starting point for the generation of GSMs for other interesting microalgae, with the *proviso* mentioned earlier that it might not be applicable to distantly related microalgae. Alternatively, *ab initio* models can be made using genomic data for the alga in question, but employing the strategies and tools developed for *C. reinhardtii*, as has been done for *Ostreococcus* [231]. Ultimately, GSMs of various microalgae will be valuable for designing strategies that increase the production of compounds of interest [425, 472]. This, combined with the design of novel synthetic pathways, such as the species-independent prediction demonstrated for novel isobutanol, 3-hydroxypropionate, and butyryl-CoA biosynthesis [78], will pave the way for model-driven engineering of algal species.

Integrating bioinformatics and modeling for algal biotechnology

The GSMs provide a basis for both computational and laboratory-driven experiments, assisting in the discovery of biotechnology-driven solutions for genetic bottlenecks in algae. For example, to enable microalgae to become a viable industrial biosynthesis platform, their photosynthetic efficiency, product yield, and their growth rates under conditions for product synthesis will need to be addressed. Photosynthetic efficiency, with an estimated maximum of 8–9% in wild type algae [77, 244], sets a limit to both product synthesis and growth rate. Because of efficient light-harvesting antenna, algal cells can absorb much more light than they are able to use for photosynthesis [244], with the excess being lost as heat or fluorescence. In dense algal cultures, such as might be found in industrial cultivation systems, this reduces light penetration, placing a limit on the depth of the culture, increasing the surface area to volume ratio required for maximum productivity. Truncated light-harvesting chlorophyll antenna size (tla) mutants of *C. reinhardtii* with reduced antenna size have been shown to have improved solar energy conversion efficiency and photosynthetic productivity in mass culture and bright light [218]. Another study has modeled different pathways for the process of carbon fixation [23] as a means to overcome the low oxygenase activity of Rubisco [446]. Bar-Even et al. [23] computationally identified alternative carbon fixation pathways by using approximately 5000 known metabolic enzymes, hoping to find

carbon fixation pathways with superior kinetics, energy efficiency, and topology. Some of their proposed pathways were estimated to be up to two- to threefold more efficient than the conventional Calvin–Benson cycle. Using an algal GSM to study these pathways would help in understanding how these predictions may affect biomass and product synthesis in microalgae.

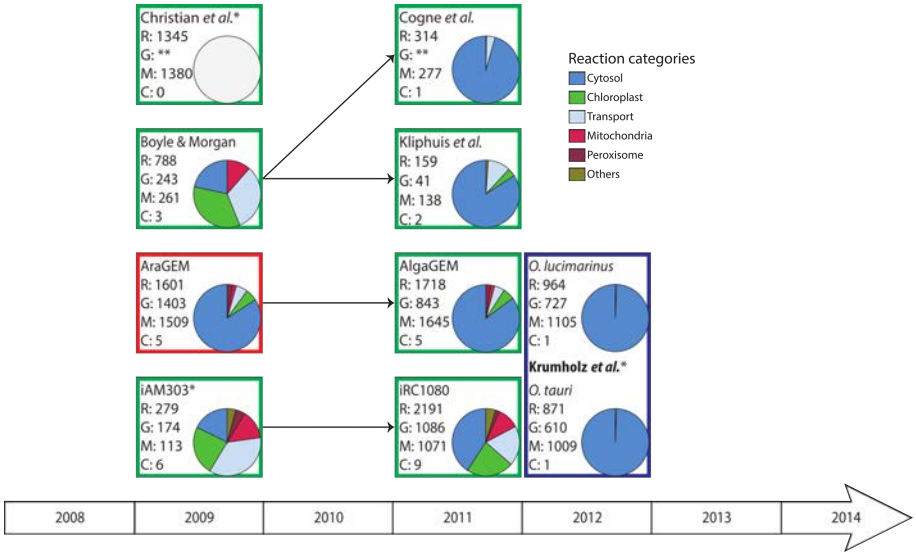


FIGURE 7.1: Overview of metabolic models of microalgae. Green boxes represent *C. reinhardtii* GSMs, the red box represents an *A. thaliana* GSM associated with one microalgae GSM, and the blue box represents two *Ostreococcus* GSMs. A connection between two GSMs indicates that the former was used in the reconstruction of the latter. The boxes are annotated with the model names if available and otherwise with the author name(s). The numbers in each box indicate the total number of reactions (R), total number of genes (G), unique decompartmentalized metabolites (M), and biological cellular compartments (C) found in the model files. The pie charts depict the distribution of biochemical reactions among different compartments as well as compartment-spanning reactions (transport). The meaning of the different colors is shown in the legend. The group 'Others' contains the compartments: flagellum, Golgi apparatus, thylakoid lumen, nucleus, and eyespot. **(*)** Additional information obtained from authors. **(**)** Gene information not available from model files.

As explained earlier, nitrogen limitation is a necessary stimulus for TAG accumulation by microalgae [222]. This also triggers a reduction in photosynthetic membrane lipids and cessation of cell growth. The link between accumulation of lipid (including TAG) and macronutrient stress has been investigated using a systems approach, such as in a proteomic analysis of *C. vulgaris*,

which led to identification of new transcription factors associated with lipid accumulation, offering the prospect of TAG overproduction independently of nutrient limitation [160]. In another approach, in the diatom, *Thalassiosira pseudonana*, TAG production was increased not by targeting the biosynthesis of lipids, or the production of competing energy sinks, but instead by RNAi knockdown of lipases involved in glycerolipid catabolism [428]. The integration of knowledge gained from GSMs and similar metabolic engineering offers scope for improved efficiency based on rational design. For example, farnesyl pyrophosphate is a precursor of terpenoids, steroids, and carotenoids, and the metabolite itself is also a product of interest in algae. Bacterial promoters responsive to the toxic accumulation of farnesyl pyrophosphate have been identified and used to regulate the expression of the precursor biosynthesis operon. This increased the yield of amorphadiene twofold over chemically inducible and constitutive gene expression [94]. Such an approach in microalgae would be foreseeable in the future, when promoters in various algal species are better understood, through model-driven design that incorporates systems data.

Alongside genomic sequence information, a key requirement is the ability to carry out genetic transformation, and while this is routine for *C. reinhardtii*, and a few other species such as the diatom *P. tricornutum*, in the past few years there has been a rapid increase in published methods for the transformation of several species of industrial interest including *Nannochloropsis sp.* [211]. Moreover, the ability to engineer the chloroplast genome offers considerable opportunities for metabolic engineering, given the focus of this organelle on biosynthesis [343]. Nevertheless, for predictive metabolic engineering there is an urgent need to expand the toolbox, particularly for the regulation of transgene expression. In this context, there are several well-established systems for inducible gene expression in *C. reinhardtii*, most notably promoters that are regulated in response to nitrate (NIT1 or NIA1) [313] or copper (CYC6) [346]. More recently, vitamin-responsive cis elements have been identified, namely a cobalamin (vitamin B12)-responsive promoter [174] as well as a thiamine (vitamin B1)-responsive riboswitch [355], and these have been demonstrated to be useful regulatory tools. Vitamins have the advantages of being benign, cheap, and effective at low concentrations. However, the majority of these elements have been discovered by coincidence rather than by design, and a more rational approach will come from use of transcriptomic data to provide promoters responsive to particular regulators, for example in response to CO_2 levels [126]. Further facilitation of transgene expression comes from the use of 2A peptides [356] which cause self-cleavage to release individual domains from a fusion protein. They thus provide the capacity for operon-like transgene expression within the nucleus. Marker recycling methods for chloroplast

engineering have also been developed for *C. reinhardtii* [98, 343]. However, despite these developments, progress remains parallel in nature and heavily focused upon the development of *C. reinhardtii*. Information from algal genomes will be key to increasing the molecular tools available.

Box 3. Integrative and systematic understanding of algae
Improvements in algal annotations will need to interact closely with systems modeling of the metabolic and regulatory networks to refine our understanding of the capabilities of a specific alga and to provide a basis for applications in biotechnology. Figure 7.2 shows the connection between the various stages in bioinformatics and systems-biology modeling. New algal genomic, transcriptomic, and proteomic data are collected (step 1), allowing the identification of genes and proteins (step 2). After first-generation high-throughput functional annotation (step 3), a refinement step using second-generation functional annotation algorithms (step 4) is applied. The bioinformatics annotation itself is an iterative process for genes and proteins until they are deemed sufficient (step 5). These annotations (step 6), as well as data available from public databases and the literature (step 7), are then used by systems-biology modeling to reverse-engineer a GSM (step 8) to study metabolic interactions in different circumstances in detail. After attaining a GSM, experimental validation of the metabolic model (step 9) should be performed to validate model predictions or pinpoint inaccuracies and knowledge gaps. Depending on these results, additional omics data or refinements of annotation are required. Owing to the low number of experimentally validated algal proteins, the feedback loop from algal modeling back to genes/proteins function prediction plays a significant role in strengthening the knowledge foundation, and this will ultimately underpin efficient engineering of algal genomes for industrial product synthesis. Once an algal GSM is constructed it should be made available in a common public database and literature.

Nonetheless, for microalgae to be developed as a commercially viable biotechnology platform, rational design to address the current shortcomings must be achieved through the development of fit-for-purpose metabolic engineering or synthetic-biology resources. The diversity of algae provides considerable biotechnological potential but also presents a serious challenge to establishing common tools and approaches. The relative immaturity of the field, combined with the enticing potential of integrating predictive design of microalgae with the bioinformatics and systems-biology modeling framework (figure 7.2), offers new perspectives for future improvements in algal biotechnology. By adapting cutting-edge developments in functional annotation for

microalgae, and using these for the modeling of their metabolic and regulatory pathways, it will be easier to establish common features of algal genomes, and at the same time identify novel pathways for exploitation. A more accurate and elaborate functional annotation of omics data by combining first- and second-generation methods will allow reverse-engineering based on algal GSMs. These can then be used to inform hypothesis-driven metabolic engineering experiments in microalgae. Such an integrated approach is currently missing, but will provide the knowledge necessary for predictive modifications of algal industrial biotechnology platforms in the future.

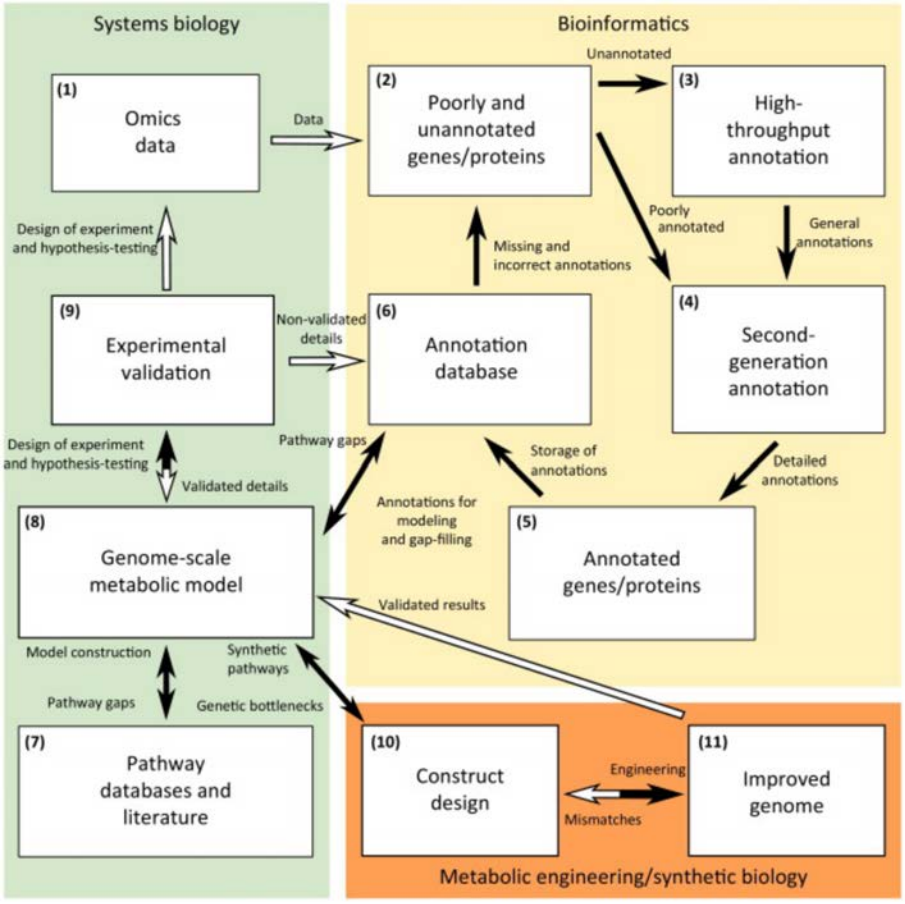


FIGURE 7.2: A multidisciplinary workflow integrating bioinformatics, systems biology, and metabolic engineering/synthetic biology of microalgae. Black arrow, in silico data or predictions; white arrow, experimental (wet-lab) data.

Concluding remarks

The significant gap of unknown and non-validated gene and protein functions in algae remains one of the top challenges faced by scientists wanting to tap further into the potential of these organisms for sustainable biosynthesis. Predictive design of metabolic engineering strategies for microalgae still has a long journey ahead. An improved understanding of the metabolism, regulation, and growth of algae, together with their interactions with coexisting bacteria, is a crucial first step. Extending bioinformatics approaches for function prediction through incorporation of new methodology, integrated and flexible databases, in combination with metabolic modeling and model-driven design of experiments at the systems-biology level, will underpin this process and enable the future era of algal industrial biotechnology.

Funding

We acknowledge support from the European Commission 7th Framework Programme (FP7) project SPLASH (Sustainable PoLymers from Algae Sugars and Hydrocarbons), grant agreement number 311956.

Chapter 8

Metabolic modeling for gut research

Adapted from:
Kees C. H. van der Ark*, **Ruben G. A. van Heck***, Vitor A. P. Martins Dos Santos, Clara Belzer, Willem M. de Vos. "More than just a gut feeling: Constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes". Submitted for publication.

*Equal contributions

Abstract

The human gut is colonized with a myriad of microbes, with substantial interpersonal variation. This complex ecosystem is an integral part of the gastrointestinal tract and plays a major role in the maintenance of body homeostasis. Its dysfunction has been correlated to a wide array of diseases, but the understanding of causal mechanisms is hampered by the limited amount of cultured microbes, poor understanding of phenotypes, and the limited knowledge about interspecies interactions. Genome-Scale Metabolic models (GSMs) have been used in many different fields, ranging from metabolic engineering to the prediction of interspecies interactions. We provide showcase examples for the application of these GSMs and focus on (i) the prediction of minimal, synthetic or defined media, (ii) the prediction of possible functions and phenotypes, and (iii) the prediction of interspecies interactions. All three applications are key in understanding the role of individual species in the gut ecosystem as well as the role of the microbiota as a whole. Using GSMs in the proposed fashions has led to designs of minimal growth media, an increased understanding of microbial phenotypes and their influence on the host immune system, and in dietary interventions to improve human health. Ultimately, an increased understanding of the gut ecosystem will enable targeted interventions in gut microbial composition to restore homeostasis and appropriate host-microbe crosstalk.

Understanding the gut microbiome

The human gut is colonized since birth with complex microbial communities, mainly consisting of bacteria with millions of unique genes that show substantial interpersonal variation in adult life [344]. This complex ecosystem – the gut microbiome– is an integral part of the gastrointestinal tract (GIT) and is intrinsically involved in the maintenance of body homeostasis. Aberrations in the microbial composition have been correlated to a wide array of diseases, ranging from obesity to diabetes, and from inflammatory bowel diseases to autism [139, 467]. These correlations have spawned interest in developing strategies to improve human health by rationally steering this composition and thereby the function of the gut microbiome [114, 207]. This approach has been greatly stimulated by the success of transplantations of faecal microbiota, which showed that 'bugs-can-beat-drugs' in fighting recurrent *Clostridium difficile* infections [302]. However, rationally steering microbiome composition and function requires a thorough understanding of the causal mechanisms underpinning these correlations. Thus far, this understanding has been hampered by (i) the limited amount of cultured and sequenced gut bacteria, (ii) the poor phenotypic characterization of the majority of gut microbes, and (iii) the limited understanding of the interactions of microbes with each other as well as their host. As in other areas of research, the deployment of descriptive and predictive mathematical models has the potential to provide insights that ultimately enable to overcome these limitations. In this review we will discuss the use of genome-scale constraint-based metabolic models for an increased understanding of the gut microbiome and its role in gut homeostasis and (dys-)function.

Genome-Scale Metabolic models (GSMs) in gut microbiota research

GSMs are mathematical representations of the knowledge on an organism's metabolic capacity and have been previously applied in bacterial systems for a variety of purposes, including the design of cultivation media, phenotypic characterizations, and study of interspecies interactions (Table 1).

Strong developments in both GSMs and gut microbiome research are bound to facilitate moving from correlation studies to gaining mechanistic insights. GSMs can integrate knowledge on the metabolism of one or more gut microbes and predict how this metabolic system functions and responds to a constantly changing environment. The gut environment includes nutrient gradients both along the length of the GIT, as well as along the mucosal gradient and villi, and have strong effects on the microbial function [123, 362].

GSMs provide a valuable framework for the integrated study of gut function as they enable the generation of testable hypotheses that can lead to novel insights into causal relationships between the gut microbiome and human health. Considerable progress in these relations has been obtained with the short chain fatty acids (SCFAs) that are produced as main bacterial metabolites in the colon, as illustrated for butyrate, an established functional compound [165, 390]. The impact of SCFAs on metabolic health has been reviewed recently [342]. In a model system it was found that acetate is secreted by *Bifidobacterium adolescentis* L2-32 and taken up by *Faecalibacterium prausnitzii* A2-165 that in turn produces butyrate. This enabled the prediction of *F. prausnitzii* acetate requirements for butyrate production and how this relates to its low abundance in cases of Crohn's disease [4], showing how an observed correlation can possibly be explained mechanistically using GSMs. However, recent research indicated the potential anti-inflammatory activity of a small protein produced by *F. prausnitzii* [345].

In the remainder of this review we will discuss the use of GSMs in gut microbiota research and how GSMs can advance gut research towards the understanding of gut homeostasis and (dys)function. We will focus on the metabolic reactions of the microbes in the gut, on their growth, on their interactions and on the metabolites produced. These are either primary products of microbial metabolism or breakdown products of our diets or host compounds, having a plethora of functions, ranging from SCFAs that fuel enterocytes and have specific signalling and immune functions, to vitamins and other host growth-promoting compounds [469]. Most of these metabolites cannot be easily detected in the human GIT as these are taken up by the host and processed in the liver. Since GSMs stochiometrically represent all metabolic reactions in a microbe or microbial community, such models enable to estimate the production of these transient metabolites, estimate their distributions within the global metabolic network and provide hypotheses for the metabolic interactions among gut microbes and of those with their host. Moreover, GSMs are instrumental in optimizing growth of GIT microbes in laboratory conditions and hence are relevant for the production of biomolecules that are involved in host signalling, such as TLR ligands or specific functional proteins [322, 345]. First, we briefly describe the process of genome-scale metabolic reconstruction and its implications for network modeling. Secondly, we describe applications of GSMs for gut microbiome research that enable: (i) selecting minimal and defined growth media for previously cultured as well as not yet cultured gut microbes, (ii) predicting growth and phenotypes of gut microbes and their influence on health and disease, and (iii) modeling co-cultures and multispecies interactions of gut microbes and the human host (figure 8.1).

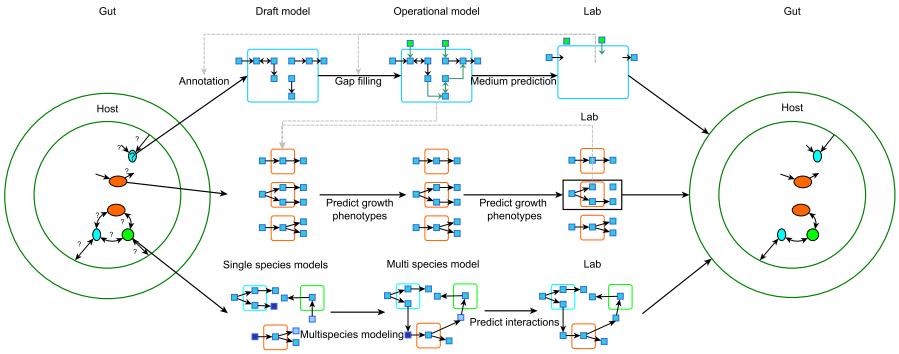


FIGURE 8.1: Simplified overview of the use of GSM to increase understanding of the metabolic interactions in the gut microbiome. Individual species require metabolites (squares) to grow. These metabolites can be predicted by GSMs, which results in medium and growth (rate) prediction (i, top). The possible solution the bacteria use to metabolize these metabolites can change under different conditions (ii, middle), which leads to altered interactions between bacteria (iii, bottom)

Genome-scale metabolic reconstruction and network modeling

The basis of GSM construction is the genome annotation of the microbe of interest since this predicts the enzymes a microbe encodes, and thereby provides a list of chemical reactions the microbe can perform. This list of chemical reactions forms the draft metabolic model, which is often far from complete [418]. Typically, there are missing reactions due to incorrect, missing, or low-quality annotations, even for well-studied organisms [317]. Moreover, our knowledge of the biochemical pathways is often insufficient, with unknown conversions still being discovered [58]. These missing reactions – also called gaps – severely limit the possibilities for GSM analyses, as parts of the metabolic network are not connected. Therefore, gap-filling algorithms are used to predict the presence of additional reactions that can be obtained from reaction databases such as KEGG [200] or Metacyc [66] and used to connect disconnected parts of the network [316, 418]. Thereby, these algorithms provide hypotheses on enzymes that were missed in the genome annotation. In some cases, a corresponding gene, not initially annotated as such, is identified and the genome annotation is improved. In the remaining cases, the reactions become 'orphan reactions', *i.e.*, reactions that are thought to occur in the microbe based on existing pathways of other microbes but that have not been linked to any genes. The addition of orphan reactions might lead to erroneous model

predictions, but is often essential to obtain a functioning GSM and facilitates targeted gene identification [316, 418]. Model construction and gap-filling algorithms have been extensively described elsewhere [316, 324, 418].

After gap-filling, GSMs are expected to be able to sustain *in silico* growth of the modeled organism. Growth is modeled as the formation of biomass in a complex reaction involving a large number of biomass precursors such as DNA, RNA, proteins, lipids, ATP, NADPH, and various small molecules. If all of these precursors can be formed in the right ratios the GSM predicts that growth is possible. The most common way to predict growth phenotypes is through Flux Balance Analysis (FBA), reviewed in [318]. FBA determines an optimal flux distribution for the production of biomass components while adhering to several types of constraints: (i) mass-balance constraints; the production and consumption of intracellular metabolites cancels out, (ii) thermodynamic feasibility constraints; reactions can only operate in thermodynamically feasible directions, and (iii) capacity constraints; fluxes through reactions are bounded to biologically feasible ranges. Capacity constraints are also used to define the medium conditions by directly defining which metabolites can be imported. Thereby, GSMs can be easily modified to simulate growth phenotypes in a wide range of different experimental conditions.

GSMs are typically evaluated by comparing predicted growth phenotypes for both wild type and mutant strains to the available experimental data. This experimental data usually consists of growth measurements for a large number of media containing different carbon, nitrogen, phosphorus and sulphur sources. For the comparison, both the experimental data and the GSM predictions are discretized to the two states 'growth' and 'no growth'. This binary discretization leads to two different types of inconsistencies: (i) growth predicted by the GSM but not experimentally found, and (ii) growth that is experimentally validated but not predicted by the GSM. In the first case, the GSM overestimates the microbe's abilities, suggesting it may include reactions that the microbe cannot perform. In contrast, the other case suggests that the GSM is missing reactions. This comparison can thus be used to evaluate both the annotation and the gap-filling process that underlie the GSM construction. For example, if the removal of a single reaction from the GSM results in a large improvement of GSM predictions, this suggests that this reaction was erroneously added and should be considered for removal. This process of using experimental data to find incorrect GSM predictions and subsequently making changes to the GSM has also been combined into algorithms, such as GrowMatch [235], that will make a minimal number of changes to a GSM while maximizing its coherence to experimental data.

The established manual GSM reconstruction process ultimately results in high-quality GSMs, but is extremely time-consuming [316]. The advent of

high throughput sequencing and concurrent rapid increase in available biological data warrants a faster approach, which is provided by the RAVEN toolbox [3] and the Model SEED approach [176]. In both cases, the process of genome annotation, draft GSM construction and gap-filling has been fully automated, although some level of manual curation is recommended to sustain a high quality [3, 176].

The need for manual curation in automatically generated GSMs is particularly relevant for poorly or not yet characterized microbes, which is the case for many gut-inhabiting microbes. In the annotation step many gene annotations may be missed due to poor homology to known sequences, or due to the reactions being completely new. Especially in the latter case, there is a large risk of reactions being erroneously added in the subsequent gap-filling step. In particular, it is important to be aware of the underlying assumptions in a gap-filling process. For example, the gap-filling process used by the Model SEED requires the composition of a medium in which the microbe grows [176]. The established GSM generation process is thus not directly suitable for the generation of GSMs for microbes with complex dependencies and syntrophic relations with a host or with other microbes.

Using GSMs to design defined culture media

The basis of classic microbiology is the ability to culture bacteria in a pure culture on a well-defined medium. Pure cultures have been successfully obtained for over 1000 different gut microbes [350]. However, as it has been predicted that there are at least two to three times more different gut species, the majority of gut microbes remain uncultured and inaccessible for study in isolation [364], although recent studies have increased the number of cultured gut bacteria [56]. A major issue in the culturing of these microbes is the lack of suitable growth media. Growth media are often based on the ecosystem a microbe naturally occurs in, but the gut is extremely complex with many different nutrients, highly variable nutrient levels, and many interspecies interactions. Here we describe how GSMs have previously been used to design minimal or defined media, and how a similar approach can be used to culture not-yet cultured bacteria.

GSMs can be used to reduce a complex growth medium to a minimal growth medium, as has been shown for the lactic acid bacterium *Lactobacillus plantarum* WCFS1 [414], and illustrated in figure 8.2. Lactic acid bacteria are important in many industrial food processes and some are marketed as probiotics [415]. Therefore, the GSMs of lactic acid bacteria are used to study their metabolic capabilities and behavior in fermentation processes [416, 441],

as well as their probiotic functions [371, 375]. The GSM of *Lactobacillus plantarum* WCFS1 was automatically constructed based on its genome sequence and subsequently extensively manually curated [219, 414]. The GSM was then used to predict the essentiality of 36 compounds in a chemically defined growth medium. The GSM predictions were correct for 29/36 (81%) of the compounds, but were incorrect for the vitamins folate, thiamine, and vitamin B6, as well as for the amino acids arginine, glutamate, isoleucine, and tryptophan. The incorrect predictions pinpointed errors in both the GSM construction process and in the experimental procedures, and also pinpointed distinct metabolic features of *L. plantarum* WCFS1, for example: (i) The incomplete folate biosynthesis pathway in the GSM was in part due to a missing EC number for a correctly annotated gene, as well as no reactions in Metacyc for another EC number. (ii) The GSM lacked a complete isoleucine biosynthesis pathway, but growth was observed in the isoleucine omission experiment. This turned out to be a result of isoleucine contamination in the other amino acids. (iii) A missing reaction for thiamine biosynthesis was assigned to a gene involved in molybdopterin biosynthesis. In *Enterobacteria* these reactions are carried out by two paralogs, but it appears that both reactions are carried out by a single enzyme in *L. plantarum* [414]. These results clearly illustrate how a GSM-driven systematic evaluation of medium compositions can increase the understanding of a microbe's metabolism.

Similarly, a GSM was used to remove all non-essential metabolites from a rich medium in order to design a minimal medium for the bacterium *Lactococcus lactis* IL1403 [314]. The GSM predicted that arginine, methionine and valine are essential for growth, and that either glutamate or glutamine is required additionally. However, recent single amino acid omission experiments have led to the conclusion that arginine, asparagine, histidine, methionine, serine, isoleucine, leucine, and valine are essential medium components for *L. lactis*, and that glutamate and glutamine are not [6]. At first glance this might incorrectly seem like poor performance by the GSM. The agreements and disagreements between predictions and experiments can be summarized in three points: (i) they agree on the essentiality of arginine, methionine, valine and the non-essentiality of the ten amino acids not previously mentioned, (ii) they do not evaluate glutamate and glutamine in the same manner - the GSM predicts that one of them is required, whereas the experiment indicates that either one can be omitted, but that glutamine cannot be omitted if the concentration of glutamate is additionally reduced to 10% of the normal concentration - and (iii) they disagree on the essentiality of asparagine, histidine, isoleucine, leucine, and serine, but also disagree on the meaning of 'essential'. In the *L. lactis* IL1403 GSM a compound was essential if its omission reduced the specific growth rate below 0.01/h. In the omission experiment a compound was essential if the growth rate dropped below 40% of the growth

rate in the rich medium. This introduces a certain level of ambiguity and, for example, if the experimental threshold would instead be at 20%, asparagine and serine would not have been considered essential.

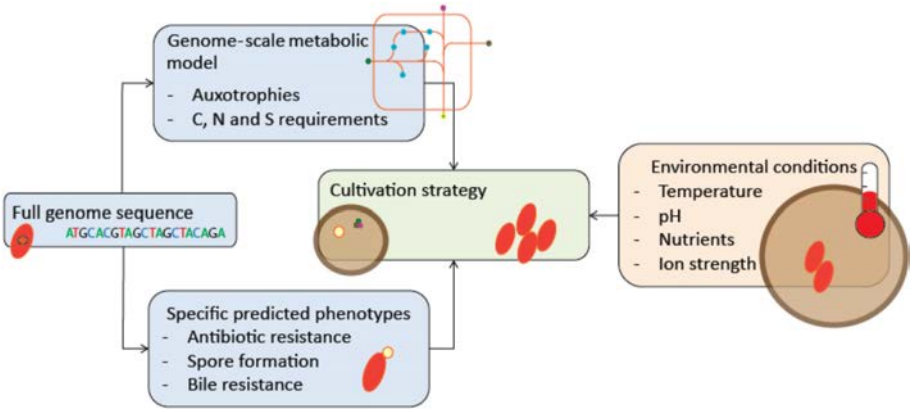


FIGURE 8.2: Suggested cultivation strategy. The initial cultivation strategy of a microbe can be optimized by thorough analysis of its genome and isolation conditions. The genome contains information on metabolic pathways, as represented in GSMs, that inform on auxotrophies and suitable carbon, nitrogen, and sulfur sources. In addition, the genome annotation can reveal additional considerations such as antibiotic or bile resistance, or the ability to form spores. The isolation condition of a microbe, for example the human gut, provides information on suitable environmental conditions such as temperature, pH, and ion strength.

The ability to culture pathogens and probiotics is important to study them in isolation and to determine their role in the gut microbiome. Therefore, a GSM was used to design a minimal growth medium for *Staphylococcus aureus* N315, a pathogen that frequently infects hospitalized patients [30]. The GSM predicted that several amino acids were essential, but in vivo experiments indicated otherwise. Later on, an updated GSM predicted that *S. aureus* N315 has no intrinsic auxotrophies for amino acids, but that some particular isolates do require some amino acids [172]. This discrepancy between the updated GSM and the experimental results for the isolates was explained by the repression of amino acid synthesizing genes. The repression could be relieved by progressively eliminating the amino acids from the medium, supporting the GSM prediction that *S. aureus* can indeed synthesize these amino acids. This study showed how a GSM can aid in omitting nutrients from a known defined medium.

These three case studies show that GSMs are a good starting point for designing minimal media. In fact, the ability of GSMs to design growth media was recently emphasized by the development of the Minimal ENvironmental TOol (MENTO) [464]. MENTO predicts the minimal medium requirements for an organism based on its GSM, and was used to study broad nutritional trends in over 2500 automatically generated SEED [176] models. For 5 well-characterized organisms, the predictions based on the SEED models were also compared to the predictions based on manually curated models. The comparison indicated that although the SEED models underpredicted growth abilities of the modeled organisms, they are not actually worse predictors than the manually curated models. Nonetheless, the authors indicate that while the SEED models are suitable for studying broad nutritional trends, one should be careful in interpreting results for any specific organism. A SEED model may thus require some manual curation before using it to predict suitable minimal growth media.

Such a manually curated SEED model was recently used for minimal medium design for *Faecalibacterium prausnitzii*, a prevalent and beneficial gut microbe that is commonly grown on the chemically undefined YCFAG medium [173]. The automatically generated SEED model was manually curated such that it correctly captured the known biochemistry and physiology of *F. prausnitzii*. This curation involved changing the biomass reaction, updating reaction directionalities, adding species-specific pathways, and filling gaps. The curated GSM was then used to predict a chemically defined growth medium called CDM1. CDM1 did, however, not facilitate in vitro growth and was subsequently supplemented with additional nutrients to form an extended medium CDM2, which did facilitate in vitro growth. The researchers then used LC-MS to identify what metabolites in CDM2 are net consumed, and what metabolites are net produced. The metabolite consumption and production data was then used to improve the GSM and the corresponding genome annotation. Ultimately, the researchers were able to design a refined and chemically defined medium CDM3 that facilitated both in silico and in vitro growth, albeit that growth was still rather poor and unreliable [173].

Metagenomic studies [291] and single-cell genomics [226, 239] of gut bacteria have already yielded genomes that could be used to create draft GSMs. However, the available biochemical information to turn draft GSMs into functional GSMs for uncultured bacteria is limited. To gain more insight in secreted metabolites and available nutrients in the gut, imaging mass spectrometry can be applied [357]. These uptake and secretion patterns can be incorporated into GSMs. We propose to use GSMs to predict minimal or defined media on which the microbes of interest can be cultured. Combined with additional ecological and genomic markers, such as temperature, antibiotic resistance and spore formation, it should be possible to culture more bacterial

species (figure 8.2). The next steps are in predicting how varying environments result in different phenotypes.

Phenotype prediction

Most microbes have versatile and complex metabolic pathways. Often, many alternative pathways are available for the conversion of the available substrate to all biomass components. GSMs can be used to explore all possible phenotypes for a wild type or mutant strain in a given environment. In addition, GSMs can be used to interpret experimental data that is difficult to directly connect to metabolic rates, such as transcriptomics and proteomics data. GSMs, which are ultimately based on genotypes, are thus a means to explore possible phenotypes in a wide range of different experimental conditions. The GSM-driven methods for exploring the genotype-phenotype relationship have recently been extensively reviewed [248]. The ability to predict how different microbial phenotypes result from different environments can ultimately have consequences for human health. For example, GSMs may be able to identify the conditions under which conditional pathogens become pathogenic [310], or, in contrast, when therapeutic bacteria or probiotics may convey their beneficial properties [375, 435].

GSM-driven exploration of the metabolic capacities of pathogens has been explanatory for pathogenic phenotypes. For example, a GSM was used to predict virulence of *Salmonella* in a mouse model system. The GSM describes a very versatile metabolism that enables *Salmonella* to utilize 31 host nutrients, allowing it to grow fast within the host cell. The GSM predicted the pathogenicity of phenotypes and was accurate in 92% of the cases [397]. In addition, it was found that the metabolic capabilities of *Salmonella* show similarities in host dependency for growth substrates and biosynthesis to other pathogens. Like *Salmonella*, other pathogens are also capable of degrading purine nucleosides, pyrimidine nucleosides, fatty acids, glycerol, arginine, N-acetylglucosamine, glucose and gluconate. Similarly, it was hypothesized that comparisons of metabolic patterns between *Pseudomonas aeruginosa* and non-pathogenic relatives could yield insight into opportunistic pathogenic phenotypes of this species [310], as has later been done successfully for *Burkholderia* species [27]. The metabolic model for the pathogenic *P. aeruginosa* also showed a versatile metabolic pattern and accounted for virulence inducing pathways, such as exopolysaccharide alginate synthesis [354].

Transcriptomics and proteomics experiments aim to discover what an organism is doing, but the data is often difficult to analyse because there are no one-to-one relationships between expression levels, protein quantities, and enzymatic activities. GSMs can aid in elucidating the metabolic activities

from these data by visualising the data on a metabolic map or by predicting metabolic fluxes. For example, transcriptomics data of two strains of *Lactobacillus reuteri*, with potentially opposite effects on the human immune system, were analysed by visualising the data on two GSMs. The analysis revealed that both strains produce vitamins, essential amino acids, and mucosal binding proteins, but that they differed in their production of potential inducers of tumour necrosis factor [375]. The prediction of metabolic fluxes from omics data relies on the concept that, on average, gene expression levels are a proxy for enzymatic activities. The GSM then predicts a flux distribution that matches the trends in the expression data, while accounting for mass balance, thermodynamics, and capacity constraints. Several such methods have been developed in the last few years, and have been extensively summarised and evaluated recently [259].

A different approach to find out what an organism is doing, rather than what it can do, is by combining GSMs with other models, such as regulatory networks [68, 127, 212]. The regulatory networks of well-studied species such as *E. coli*, *M. tuberculosis* and *M. genitalium* have been elucidated and incorporated in metabolic models [65, 203, 215]. Based on these model organisms, attempts have been made to automate the incorporation of regulatory networks into GSMs [308], also especially aiming at less well-characterized species [69]. These models incorporate the influence of environmental factors on the behavior of the modeled organism, which may be extremely relevant for microbes residing in a dynamic environment such as the human gut.

These examples show how GSMs can be used to explore possible phenotypes, and to predict actual phenotypes based on omics data or regulatory models. We propose to use GSMs in combination with transcriptomics data to predict the phenotypes of gut bacteria. In this way, the role of bacteria can be predicted under different gastrointestinal conditions, on which also other microbial species have a big influence.

GSM predicted co-cultures and interspecies interactions

Within the gut microbiome there are numerous microbial interactions and networks. Three types of simple multispecies interactions have been described and modeled before: mutualism, commensalism and neutralism [221, 271]. GIT-colonizing microbial species often depend on each other for growth signals and substrates or compete for the metabolites, thus this ecosystem is ideal for the modelling of interspecies interactions and using interspecies interactions predictions to gain a mechanistic insight into this ecosystem [48, 191]. Interactions between microbes have been modeled on different phylogenetic

levels, ranging from strains [430] to species [370, 403] and ecosystem communities [246]. The different modes of modeling are described below and summarized in figure 8.3.

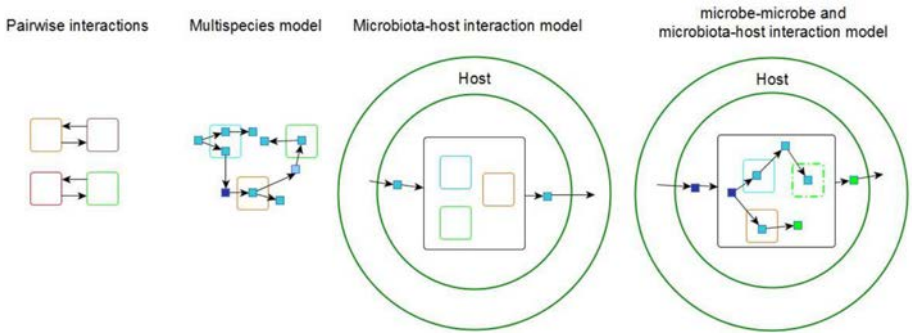


FIGURE 8.3: Modes of interspecies interactions as modeled before. Pairwise interactions only account for two species to share metabolites. Multispecies models allow sharing of metabolites between more than two species. Microbiota-host interaction models lump all the microbial species into one meta-model and model the interaction with the host. Microbe-microbe and microbiota-host interactions are multilevel models that take into account microbial interactions and interactions with the host.

The most straightforward way to create multispecies GSMs is by combining multiple GSMs into a single model, where the individual GSM networks share the extracellular environment. This approach has been taken to model an ecologically relevant mutualism between the bacterium *Desulfovibrio vulgaris* and the archaeon *Methanococcus maripaludis* S2 [401]. In this syntrophic relationship, *D. vulgaris* ferments lactate, and *M. maripaludis* consumes the fermentation products formate, dihydrogen and acetate. This work addressed several challenges in the use of multispecies GSMs that originate from the presence of two independent types of biomass. Normally, for single-species models, all fluxes in a GSM are scaled to the amount of biomass [mmol gdw^{-1} h^{-1}], and FBA is used to predict a flux distribution that leads to maximal biomass production. In the case of a multispecies GSM both of these characteristics become problematic. The first issue is problematic because there are two different types of biomass, but was solved by re-scaling, and expressing all fluxes in [μmol h^{-1}]. The second issue is problematic because although the assumption that the metabolic goal of a single microbe is to produce biomass is reasonable, it is not reasonable to assume that multiple different microbes strive to produce as much biomass as possible, irrespective of which species it belongs to. In this study, the second issue was evaluated by adding different weights to the different biomass reactions. The predicted biomass production

for *D. vulgaris* was practically independent of the relative weights, whereas the *M. maripaludis* biomass production increased if it received higher weights. This is due to the sequential nature of the interaction between these bacteria, where *D. vulgaris* effectively 'feeds' *M. maripaludis*. Such a direct approach to multispecies GSMs does not work in case of cross-feeding or substrate competition.

A GSM containing seven well-known species was constructed to predict pairwise modes of interaction, but not for all species simultaneously [221]. This number was rapidly expanded to 118 species coupled in 6.903 pairs driven by automated curation of over a hundred GSMs [146]. The competition between the bacteria was measured by pairwise simulated growth on media that stimulated competition by using overlapping resources. Similarly, cooperation was simulated by using a medium that contained the minimal resources for both bacteria to grow. Competition was generally won by species that grow fast on versatile media, such as *E. coli*. Cooperation was more evident in *Clostridia* species that are able to degrade lignin and cellulose, which releases free sugars to other bacteria. This type of macromolecule degradation is highly important in degradation of host dietary compounds and thus directly relates to gut health.

Instead of looking into the details of the interactions between a few species, GSMs have also been used to elucidate general properties of the co-occurrence of microbes. Specifically, there are two main mechanisms driving species co-occurrence: (i) habitat filtering: microbes occupy a similar nutritional niche and compete, and (ii) species assortment: microbes have complementary metabolisms and cooperate. A recent study aimed to identify which of these two mechanisms is the driving force behind the co-occurrence of microbes in the human gut [246]. Therefore, they automatically generated 154 GSMs based on KEGG [133, 200] for gut microbes whose co-occurrences were determined based on a gut metagenome dataset containing measurements from 124 individuals. These GSMs were used to determine metabolic competition and complementarity indices between each pair of species. As the species co-occurrence was best explained via the metabolic competition index, the authors concluded that habitat filtering is the main driving force behind species co-occurrence in the human gut. In an other recent study, GSMs were used to study species co-occurrence based on 261 microbial species in 1297 communities from diverse habitats [465]. The GSMs were used to calculate both the resource competition and interaction potential within these communities. Resource competition was significantly higher in the 1297 communities versus random assemblies, indicating that habitat filtering was again identified as the main driving force behind community composition. However, there were also 7221 subcommunities of up to 4 co-occurring species within the larger

communities. Within these subcommunities, the interaction potential – defined as the difference in minimal number of metabolites required for growth between a non-interacting and a cooperating community - was significantly higher than in full communities and random assemblies.

In order to understand how gut communities form and change, it is important to consider spatial and temporal effects. The novel modelling framework COMETS [167] - COmputation of Microbial Ecosystems in Time and Space – simulates multiple GSMs on a lattice over time using dynamic FBA [167] COMETS does not require any prior information on how the modeled microbes interact, but nonetheless captures interesting and non-intuitive spatiotemporal dynamics of multispecies interactions. For example, it correctly predicted that the slowest-growing microbe of a three-species ecosystem would also ultimately be the most-prevalent one, and that the growth rate of a colony with a mutualistic partner can be improved by placing a competing colony in between them. COMETS has also been used to study how robust competing and mutualistic interactions are to genetic perturbations. Specifically, it has been possible to predict the effects of gene knockouts on a synthetic community of *Escherichia coli* and *Salmonella enterica* [166] on competition-inducing and mutualism-inducing growth media [81]. Interestingly, the community was more robust to genetic perturbations in *E. coli* under cooperative conditions, but more robust to genetic perturbations in *S. enterica* under competing conditions [81]. These results highlight that GSMs can mechanistically explain the intriguing interactions of multispecies interactions.

A multispecies interaction of particular interest is the interaction between gut microbes and their host. The host is the most important environmental factor for gut microbes, but is also metabolically active itself. GSMs have been created for hosts of particular interest, such as mouse [385] and human [421], and have even been trimmed down to tissue-specific GSMs, including a GSM for colon-derived tissue [56]. The mouse GSM [226] was recently used to study how different diets and the presence of the gut microbe *Bacteroides thetaiotaomicron* affect its metabolism [173]. A *B. thetaiotaomicron* model was constructed using SEED [176] and, after manual curation, was linked to the mouse GSM via a shared lumen compartment. Although a single microbe is not directly representative of the gut community, the combined GSM mechanistically explained how both organisms benefit from the mutualism, correctly predicted how the interaction affects biofluid metabolome composition, and even described how gut microbes can rescue hosts with lethal gene deletions [173].

Host-microbe interactions have also been modeled using a single 'supra-organism model' [48] to represent all gut microbes. These GSMs don't focus on individual microbes or their interactions, but rather on the interaction of the community with the environment or host. Such a GSM was used together

with metagenomics data to study how host-microbe interactions differ in case of obesity or inflammatory bowel disease (IBD) [158]. This revealed a differential expression of enzyme groups expressed by the complete microbiota between diseased and healthy people, without investigating the roles of individual species or their interactions. The differences were found in the upregulation of membrane transport and downregulation of vitamin metabolism, nucleotide metabolism and transcription. This study suggests that the differences in enzyme expression originate from an altered interaction between the microbes and their environment. They are the result of a change in the environment of the bacteria and do not come from a change in core metabolic processes. By combining previous approaches of modelling interspecies interactions and considering the whole microbiota as one entity, a predictive tool for dietary interventions was created [384]. The tool, CASINO - Community And Systems-level Interactive and Optimization - predicts dietary interventions based on interactions between the host, the microbiota and the applied diet. CASINO was used to model the interactions of four microbes in two synthetic communities that differed by a single microbe. It correctly predicted the produced metabolites, including essential amino acids, and the contribution of each species to the production of each metabolite. CASINO was then used to predict the impact of a dietary intervention in 44 individuals, based on relative abundances of the most prevalent microbes in each individual before and after the intervention. The predicted production of SCFAs and amino acids mostly matched the in vivo measurements. Finally, CASINO was used to design a beneficial diet for subjects with a poor microbiota composition [384].

We propose to use predictions for multispecies interactions to study the influence of perturbations in environmental factors and communities. In this way it can be predicted how individual species contribute to healthy and diseased conditions. Moreover, this approach was instrumental in the prediction of diets to improve the metabolic function of gut microbiota [384]. Ultimately, this will lead to increased understanding of the interactions of the gut microbiota and its host, on its role in gut homeostasis and (dys-)function and should pave to way to improve human health by the use of specific gut microbes or dietary interventions.

Conclusion and perspectives

After a few decades of characterizing gut microbiota composition many gut microbes have been sequenced [333, 344]. Over 200 of these genome sequences have been used to generate GSMs, in most cases by automated tools [66, 176]. These GSMs have been used to predict growth phenotypes of single microbes and communities in laboratory and in vivo settings.

Here, we reviewed three ways in which GSMs contribute in elucidating gut microbiome function. We described how GSMs are used to: (i) culture bacteria, (ii) predict bacterial phenotypes under changing conditions and (iii) study the interactions both among the bacterial species and with their host.

We have shown that recent advances in automated generation of GSMs [176], single-cell genomics [41], metagenomics [154, 344] and metatranscriptomics [18, 19, 268] can increase the availability and accuracy of GSMs. Metagenomic as well as single-cell genomics yield more full genomes sequences of microbes that can be used for generating GSMs. These GSMs will contribute in understanding how both uncultured and cultured bacteria live and behave in complex ecosystems [191]. In vivo or in vitro validation of GSM predictions and subsequent GSM updates remain key in improving GSM quality and ultimately understanding the complex gut ecosystem.

GSMs allow understanding why species are present and what they do, instead of who they are, as was the focus in the last decades. We expect that GSMs will contribute to elucidate the mechanisms behind known probiotics, as well as in identifying new probiotics, and understanding the role of different bacteria in complex ecosystems. Ultimately, GSMs can contribute to the design of controlled interventions that steer gut composition and activity to improve human health.

Author contributions

KCHvdA, WMdV and CB defined the topic and scope of the review. KCHvdA and RGAvH developed and wrote the manuscript. WMdV and CB guided and assisted in writing the manuscript. VMdS, WMdV and CB critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

The work was supported by the Netherlands Organization for Scientific Research (Spinoza Award and SIAM Gravity Grant 024.002.002) granted to WMdV, the project IPOP on Systems Biology of the Wageningen University (RvH) and the H2020 Projects SysmedIBD (contract n° 305564) and Empowerputida (reference nr 635536) granted to VMdS. The funding organizations had no role in design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Chapter 9

General discussion

The goals of this thesis were: To increase the understanding of microbial metabolism and to functionally redesign microbial systems using metabolic models. The metabolically versatile bacterium *P. putida* was used as a case study.

P. putida: Model-driven discovery and design

A defining feature of *P. putida* is its ability to degrade an immense variety of compounds [84, 195, 424], including hard-to-degrade aromatics [290, 327, 372]. Nonetheless, many degradation pathways of *P. putida* have not been fully elucidated. This is highlighted by the inability of the most comprehensive *P. putida* Genome-Scale Metabolic model (GSM) iJP962 [311] to degrade 120 compounds *in silico*, whereas *P. putida* can degrade these compounds *in vitro* (**chapter 2**). This GSM is based on extensive literature research as well as the *P. putida* KT2440 genome annotation from 2002 [290].

Therefore, we structurally and functionally reannotated the *P. putida* KT2440 genome (**chapter 2**). The structural reannotation identified 311 CoDing Sequences (CDS) that were not identified previously, and determined that 102 previously identified CDS were false positives. The total number of identified CDS in *P. putida* is now 5592 *versus* 5350 in 2002 [290]. The functional reannotation associated 1250 CDS to EC numbers *versus* 463 in 2002 [290]. In total, 902 unique EC numbers are assigned in the updated annotation versus the 360 previously, suggesting that the updated genome annotation provides a substantially expanded view on *P. putida* metabolism.

A crucial part of the functional reannotation was its contextualization and assessment using the GSM iJP962. In an iterative approach, the GSM was (i) expanded based on reactions corresponding to newly annotated genes, (ii) used to evaluate the contribution of these new reactions in resolving knowledge gaps, and (iii) used to generate hypotheses on missing functional annotations for manual inspection. This iterative approach ultimately enabled the elucidation of proposed degradation pathways for 86/120 compounds for which these were not present in iJP962; a substantial increase in coverage of *P. putida* metabolism.

The combination of reannotation and GSM expansion pinpointed future research directions to increase and consolidate the understanding of *P. putida* metabolism: (i) Verification of annotation and proposed degradation pathways. In particular, the degradation pathways for several compounds consist of reactions that were absent from biochemical databases or rely on enzyme

promiscuity to extend substrate specificity. These newly proposed degradation pathways warrant experimental verification. (ii) Identification of transport systems. The lack of identified transporters and corresponding mechanisms played a major role in the inability of the GSM to successfully degrade compounds. (iii) Update of the *P. putida* GSM. In this study the GSM was extended to evaluate whether newly identified genes and corresponding reactions closed the knowledge gap on degradation pathways in the context of the rest of metabolism. However, the created GSM should, in my opinion, not be regarded as an updated GSM to use henceforth. In particular, the GSM was only modified to decrease false negative growth predictions, but not false positive growth predictions, and inconsistencies between the GSM and the updated annotation were not addressed. A substantial effort remains to thoroughly update a *P. putida* GSM based on recent experimental literature and the updated annotation.

To create an updated *P. putida* GSM I would start with PpuQY1140 [463]. PpuQY1140 is a recently published consensus *P. putida* GSM that was generated using a pathway-based systematic comparison of iJN746 [298], iJP962 [311], and PpuMBEL1071 [395]. This pathway-based comparison was focused on differences in predictions between the GSMs, which is complementary to the topological differences we identified between iJN746 and iJP962 using COMMGEN in **chapter 3**. As PpuQY1140 is mostly based on iJP962 [463], and an updated version of iJN746 is expected 'soon', I suggest using PpuQY1140, the updated iJN746, and the reaction list corresponding to the updated annotation (**chapter 2**) as an input for COMMGEN to create a *P. putida* consensus GSM. This GSM will comprehensively represent the available knowledge on the metabolism of *P. putida*.

P. putida is commonly found in soil, especially in polluted soils, while other members of the *Pseudomonas* genus inhabit very different growth environments. For example, *P. syringae* is a well-known plant pathogen that is also commonly found in clouds at several kilometers altitude and has been hypothesised to cause precipitation using an ice-nucleation protein [287, 339], and *P. aeruginosa* is an oppurtunistic and antibiotic-resistant human pathogen that thrives in hospitals. These distinct lifestyles in the *Pseudomonas* genus suggest that there is an immense functional genetic diversity among its members, but also that the core genome is enriched for universally essential features. In **chapter 4** we evaluated this hypothesis by linking the persistence of protein domains, *i.e.,* the fraction of strains having a particular protein domain, with the essentiality of the corresponding genes. As gene essentiality data is limited in both represented organisms and experimental conditions, we complemented the available data with simulated data obtained using several previously published *Pseudomonas* GSMs [49, 298, 310, 311, 341] exposed to thousands of different growth media. This large dataset of gene essentiality

combined with an analysis of 432 published *Pseudomonas* genomes demonstrated a clear link between persistence of protein domains and gene essentiality.

The aforementioned GSM applications describe the use of GSMs to increase understanding of genetics and metabolism in a purely *in silico* manner. If, however, the description of microbial metabolism in GSMs is sufficiently accurate, they may also have the potential for the functional redesign of microbes. Based on the work presented in this thesis, I will discuss four different avenues of this model-driven redesign: (i) further extending the *P. putida* substrate spectrum, (ii) enabling anaerobic growth in *P. putida*, (iii) enabling CO_2 fixation in *P. putida* and other organisms, and (iv) metabolic engineering.

There are several compounds that *P. putida* does not degrade *in vitro*, while complete degradation pathways appear to be genetically encoded (**chapter 2**). The only missing functionality appears to be the lack of transport proteins for these compounds. This suggests that the substrate spectrum of *P. putida* may be further enhanced by the heterologous expression of transporters for these compounds. Irrespective of the successful degradation of these compounds following the expression of a transporter, these experiments will either consolidate or increase knowledge on *P. putida* metabolism; successful degradation strongly supports the presence of the pathways, whereas a lack thereof pinpoints errors in the current understanding of *P. putida* metabolism.

Enabling an anaerobic lifestyle in *P. putida* has been a goal of several prior experimental studies [238, 292, 377, 395, 398] that have so far consistently resulted in increased anaerobic survival, but not in growth. In **chapter 5** we present an alternative systems biology approach to the design of an anaerobic *P. putida*. This approach is based on complementary *in silico* analyses that have been used to identify potential bottlenecks to anaerobic growth in *P. putida*. The identification of these bottlenecks has enabled the design of *P. putida* strains that are capable of anaerobic growth according to the *P. putida* GSMs. One of these designed strains is currently being evaluated experimentally with very promising preliminary results.

Microbial CO_2 fixation has the potential to drive the biobased economy by eliminating the need for cultivating, harvesting, and chemically pretreating the plant biomass. The organisms naturally fixing CO_2 and the employed pathways are, however, not necessarily the most optimal for industrial exploitation. Therefore, in **chapter 6**, we used species-specific GSMs as a basis to design CO_2 fixation pathways for eight industrially relevant organisms including *P. putida*. These pathways were primarily selected for high growth rate predictions in the GSM, but also for their ATP-efficiency, thermodynamic feasibility and attractive kinetics. For all eight organisms the introduction of efficient CO_2 fixation pathways requires surprisingly few non-native reactions.

Metabolic engineering refers both to the theoretical model-driven strain design for the production of a compound, as well as to the experimental realization thereof. Although none of the chapters of this thesis delves into metabolic engineering, a big part of my PhD work has been devoted to this topic nonetheless. Currently, we are creating and evaluating pyruvate production strains that were designed using a combination of the iJP962 GSM [311] and a modification of the OptKnock [59] strain design method that uses objective tilting [129], which is conceptually similar to RobustKnock [411] but faster and easier to implement. Unfortunately, at the time of this writing, I cannot yet conclude whether the *in silico* predictions accurately describe the phenotypes of the designed strains.

GSMs have been a cornerstone of the research presented in this thesis. They have been pivotal in increasing the understanding of *P. putida* metabolism, and have led to interesting strain designs. However, during this work I have also been confronted with many misconceptions and issues surrounding the use of GSMs. In the remainder of this chapter I will focus on these topics and provide my outlook on what can be improved. In particular, I will focus on: (i) Non-biological artifacts in GSMs stemming from the generation procedure, (ii) the role of GSMs as comprehensive knowledge base, (iii) the relation between GSM predictions and experimental data, (iv) the evaluation of GSM quality, and (v) the role of GSMs in metabolic engineering.

GSM generation artifacts

GSM generation involves many choices on the followed procedure. For example: (i) The used gene caller, (ii) the used gene annotation method, (iii) the reaction database to use, (iv) the reaction directionality determination method, (v) the gap-filling method, (vi) the biomass reaction, and (vii) the process for manual curation. Each of these choices ultimately affects the exact contents of a GSM. GSM contents thus not only reflect the biology of the modeled organism, but also contain artifacts stemming from the generation procedure. Hence, when multiple GSMs are independently made for the same organism they may end up being practically incomparable (**chapter 3**). The problem surrounding GSM generation artifacts only becomes truly apparent when trying to use GSMs to contrast two or more organisms. Specifically, any observed differences can be true biological differences, but may also just reflect the different GSM generation procedures (**chapter 6**).

To avoid GSM generation artefacts from impacting a comparison between GSMs, the GSM have to be generated according to the exact same procedure. Therefore, I think that the focus of future GSM generation should be on automated procedures such as the Model SEED [176] and the RAVEN toolbox [3].

Currently the automatically generated GSMs still require extensive manual curation, and it should, in my opinion, be a priority of the metabolic modeling community to decrease this requirement as much as possible.

The GSM generation artefact that is the most problematic, in my experience, is the lack of a common chemical naming system (namespace); a recurring problem in my work (**chapters 2,3,6,7**). The namespace in a GSM is typically based on reaction databases such as KEGG [201], metacyc [66], and rhea [286], or on a specifically designed GSM namespace such as BiGG [217], or SEED [176], or on a reference namespace such as MNXref [36], or what also happens, unfortunately, is that the GSM creators design their own novel namespace, or that multiple namespaces are used within a single GSM. The lack of a commonly used namespace is a substantial hurdle when not using a GSM in complete isolation, for example: (i) Gap-filling methods require the same namespace between GSM and reference database, (ii) Integrating metabolomics data requires a matching namespace, (iii) Determining reaction directionalities via Equillibrator [138] requires the KEGG [201] namespace, (iv) Combining reactions from different sources requires a common namespace, (v) Determining reaction directionalities via network patterns [149] requires reference GSMs in the same namespace, (vi) Using multiple GSMs requires a common namespace or separate scripts for each GSM.

The conversion of one namespace to another is not straightforward, although somewhat simplified through reference namespaces such as MNXref [36]. MNXref is a specifically designed namespace that connects to other namespaces commonly used for metabolic modeling. Still, however, a reference namespace can not fully solve the namespace problem between GSMs as namespaces also operate on different granularities, for example, one GSM may contain 'glucose' whereas another contains 'α-D-glucose'. In **chapter 3** we demonstrate that network topology can be used to further match the namespace of GSMs and reaction databases that operate on different granularities. As it is unlikely that a single namespace will be adopted by the GSM community, there is an urgent need for an easy and suitable method to convert one namespace to another.

GSMs as comprehensive knowledge base

GSMs are regarded as comprehensive mathematical representations of the current knowledge on the metabolism of an organism. Their generation is, as aforementioned, a rather ambiguous process involving many choices. Typically the process and choices are not fully transparent to the end-user as the ambiguities and exact choices are not recorded. For example, the provenance of why reactions are included is often missing. Reactions that have gene associations are not exempt from this requirement of provenance. Specifically, the

GSM creators should record on what basis the corresponding gene was annotated, which tools were used, which versions of these tools, which reference databases, and ultimately how certain it is that this gene-protein-reaction association is actually correct (**chapter 4**). Without this information, the end-user will simply have to accept the GSM as 'the truth', or will have to manually assess the validity of all its contents. In addition, GSMs are already outdated by the time they are publically available. The time between the initial phase of genome annotation and draft GSM generation and the final publication easily extends over more than a year. From a somewhat pessimistic perspective, GSMs are thus rather ambiguous and outdated representations of the knowledge on the metabolism of an organism.

Eventually an updated GSM is published, typically a few years later by the original creators, and the rest of the field gains access to a more up-to-date representation of the knowledge on the organism's metabolism. This time-discrete publishing of GSMs does not reflect the continuous developments in annotation tools, reference databases, reaction databases, experimental characterization, and in the GSM itself. The scientific community as a whole would be better served via continuous development and sharing rather than the occasional publication describing a major GSM update.

Continually updated GSMs can be obtained via platforms for automatic GSM generation such as the Model SEED [176] and the RAVEN toolbox [3]. The automatic generation of GSMs has the benefit of a clear procedure without human intervention implying that, in principle, all contents of the GSM can be accounted for; there can be full automated provenance. For example, for each reaction in an automatically generated GSM the origin of its reversibility can be pinpointed, whereas this is not possible if undocumented manual choices were made. Nonetheless, there remains tremendous value in the extensive manual curation that is typically applied during GSM generation, as highlighted by the remaining need to curate automatically generated GSMs [3, 173, 176]. This need for manual curation does, however, result in a similar situation as for manually generated GSMs where the scientific community does not have continuous access to an up-to-date GSM.

Currently, the scientific community thus has to choose for continuous GSM updates via automatic GSM generation, or for high quality via manual GSM curation. In order to improve the access to curated and up-to-date information a new framework is required. To me, a wiki-based framework for GSM construction, curation, expansion and analysis seems ideal. A wiki has several advantages over the current system:

- Provenance. All (suggested) modifications can be stored such that complete provenance is kept. This enables to track why each and every component of the GSM was included.

- Dynamic links to external resources. I would not suggest to automatically update the GSM based on an update in an external resource, but rather to flag the relevant part of the GSM for inspection. Examples of relevant resources are (i) the genome annotation, (ii) a reaction database, (iii) a thermodynamics calculator such as Equillibrator [138], and (iv) GSMs of closely related organisms.

- Ongoing development. Currently, if a GSM user finds an error there is no system to broadly inform other users. In the best-case scenario the error will be reported to people working on a GSM update, and in a few months to years other users will also benefit from the fix. In contrast, a wiki encourages all users to provide their expert insights from which other users directly benefit.

- Consensus representation. As aforementioned, the correct representation of biological information in GSMs is ambiguous. This ambiguity can become evident if alternative representations of a process can be suggested and discussed by domain experts. Ultimately, a community-consensus can be used to select the final representation, rather than the opinion of the small group of scientists publishing a GSM.

- Version control. It is not uncommon that multiple versions of a GSM can be found online with little to no description on how they relate to the version accompanying the corresponding article. I have even seen several examples where the version of a GSM described in the paper does not match the accompanying SBML file in terms of number of reactions, genes, metabolites, or predictions. An online resource that tracks changes can provide a date or version stamp whenever an analysis is run or the GSM is exported, enabling other users to use the same version to repeat an analysis.

- Ease of access. Arguably, a git repository would share many of the same benefits as a wiki. A wiki is, however, easy to use for people with limited programming experience. Given that GSMs ideally incorporate expert knowledge from various non-computational disciplines, their inspection and curation should be facilitated.

- Standardized testing. As discussed further below, I have found GSM evaluation to be lacking both in extent and standardization. The extent can be increased in a community-framework as any user can submit new testing criteria and data. I expect standardization within a community to follow as scientists will discuss the advantages and disadvantages of testing approaches.

Model predictions and experimental data

Models are simplifications of real systems that approximate the real system's behaviour in their predictions. If the model predictions do not match the behaviour of the real system, it is logically concluded that the model is wrong. However, in the case of GSMs, these conclusions should not be so readily draw as there is a fundamental mismatch between what GSMs describe and what is typically available as experimental data. GSMs describe the theoretical potential of an organism, whereas data represent what the organism did in a particular instance. In addition, some commonly used data types are but proxies for what they represent, for example BIOLOG [43] data.

BIOLOG experiments are commonly performed to evaluate whether or not an organism can use a wide variety of different carbon (C), nitrogen (N), phosphor (P) and sulfur (S) sources [43]. They are typically carried out using several 96-well plates where each well contains a default medium lacking a C, N, P, or S source, corresponding to a supplemented metabolite that potentially serves as a source for that element. In addition, each well contains a tetrazolium dye. This tetrazolium dye is a redox indicator, and will color purple with ongoing on redox activity. Redox activity in a particular well implies that the organism can process the corresponding compound. Subsequently, the capability of processing is taken as a proxy for growth and used for the validation of GSMs [130, 176, 312] (**chapters 2,3**). However, correct predictions regarding the ability of an organism to grow in a particular medium, may thus be deemed incorrect based on the BIOLOG data.

The BIOLOG data published as part of [294] demonstrate another implicit inconsistency between GSM predictions and experimental data. GSMs played no part in this particular work, but GSMs could never describe the accompanying data. In this work, *P. putida* knockout strains were created that either missed the soluble transhydrogenase *sthA* or the membrane-bound transhydrogenase *pntAB* or both. These strains were subsequently subjected to BIOLOG experiments, and several 'gain-of-function' phenotypes are reported. From the functional standpoint of a GSM, this is impossible; a reduction in available enzymatic activities can never lead to a new flux mode. Possibly, the 'gain-of-function' phenotype is a result of the knockout strain having a perturbed regulatory network. This can not be captured by GSMs analyses as these assume that regulatory systems have evolved to enable optimal metabolic activities, which is not the case for genetically modified strains or strains that are subjected to new growth media. These strains first have to undergo experimental evolution.

Experimental evolution has succesfully pushed strains to phenotypes that were priorly predicted by a GSM. This holds for both the adaptation of wild-type strains to a particular substrate [186, 249], as well as for the adaptation

of genetically modified strains to their new genotype [140, 141]. Experimental evolution is, in my opinion, an integral part of any GSM-driven strain design. I do not think that this should be considered as a downside, but rather as an opportunity to have Nature's optimization algorithms work for us, solving problems we may not even be aware of, and tuning the system to optimality in a way we can currently not realize via rational engineering. We will use experimental evolution for the practical implementation of the work presented in **chapters 5** and **6**. Note that the main difficulty in using experimental evolution is the identification of the conditions under which the desired phenotype is evolutionary favorable. In the examples mentioned in **chapters 5** and **6** this is trivial, as both aim to enable growth and increase growth rates.

For any experimental evolution set-up, it is important to consider the distinction between growth rates and growth yield. GSM predictions are often interpreted as growth rates [129, 176], but what they predict is the growth yield [249, 417]. In practical terms, this means that experimental evolution for GSM-based predictions requires a system where cells are not competing for resources, and selection is based on final biomass production. Alternatively, GSMs can be modified such that growth rate is predicted instead of growth yield. The reason GSMs predict growth yield is the unlimited flux for most reactions, and the lack of a link between how much protein is required to maintain a particular flux. Thermodynamically efficient reactions typically have poor kinetics, and thus require a lot of protein to maintain a sufficient flux. As GSMs do not explicitly model proteins, there can effectively be an infinite amount of proteins catalyzing these reactions with poor kinetics. If a GSM is to predict growth yield, it will thus have to be expanded to include molecular crowding, and to describe the relationship between protein abundance and attainable flux for each reaction.

GSM evaluation

GSMs — like most mathematical models — are commonly published along with an evaluation of their quality. If the quality of the GSM is deemed high, this suggests that its predictions are trustworthy and that the GSM can be used for a large variety of applications. There is, however, no clear standard protocol to determine the quality of a GSM, which has both benefits and drawbacks. A clear benefit is that GSM evaluation can be tailored to the modeled organism for optimal relevance. A drawback is that the lack of a common standard complicates the comparisons between GSMs, and enables the cherry-picking of evaluation criteria for a given GSM.

The lack of a common standard for GSM evaluation is not only restricted to which methods are being used, but also how they are used. For example, consider the most commonly used method for GSM evaluation: phenotype

predictions. Conceptually, this method evaluates the ability of the GSM to predict whether or not a wildtype or mutant strain of the modeled organism can grow in a variety of environments. In principle, the analysis seems straightforward: Use FBA [318] to predict whether growth is possible in various media conditions for which experimental data is available, and compare the predictions and experimental measurements. However, there are various ways in which this analysis can be approached. In short, there are differences in: (i) The used reference data. For example, in [235] phenotype prediction accuracies of multiple GSMs are presented while these are generated based on different datasets. (ii) The used summary statistic. In many cases the accuracy is reported, but also the geometric mean accuracy [232], and the Matthews correlation coefficient [266] are used [170]. (iii) The definition of growth. There is no universal threshold growth rate above which *in silico* 'growth' occurs and below which not. For example, *in silico* gene essentiality has been defined as resulting in growth rates ==0 [319], <0.001 [39], <0.01 [314], <5% of wildtype growth rate [363], or < a cut-off growth rate depending on other mutant strains [111]. (iv) The implication of non-modeled information. If a gene is not included in a GSM, this can either be interpreted as the GSM predicting it is not essential, or that the GSM can not make a prediction regarding the gene's essentiality. (v) The implication of dependent data. Suppose a gene deletion leads to an auxotrophy *in vitro*, but not in the GSM. The GSM is thus incorrect in the prediction of the gene essentiality, but will also correctly predict that the gene is not essential if the medium is supplemented with the compound for which the deletion strain is auxotrophic. In my opinion, the latter prediction is irrelevant given the inaccuracy of the former.

The suitability of phenotype predictions as GSM evaluation method is also somewhat dubious considering the role these predictions play during model construction. Growth phenotype data is typically used during model construction to match the *in silico* and *in vitro* phenotypes [418]. In particular, algorithms such as GrowMatch [234] will modify a GSM to maximize the consistency between its predictions and experimental data. Other types of modeling often have distinguished 'training data' and 'test data'. The training data is used during model construction, and the test data is used to evaluate how the 'trained' model performs. For GSMs this concept is not directly suitable as the ability to use a specific substrate likely depends on some substrate-specific reactions that would not be added if that substrate was not part of the training data. Nonetheless, it is not surprising that GSMs can accurately predict growth phenotypes if they were specifically altered to predict these growth phenotypes.

Growth phenotype predictions also do not capture all relevant GSM characteristics. For example, a GSM that has thermodynamically infeasible cycles resulting in unrealistically efficient ATP generation and in CO_2 fixation can

outcompete other GSMs for the same organism in terms of phenotype predictions (**chapter 3**).

The second most commonly used metric for GSM evaluation is 'scope'. Conceptually, the scope of a GSM reflects how much biological knowledge is represented in the GSM. A rough indication of this is the sheer number of reactions, metabolites, and genes that are included in the GSM. A more precise estimate also considers the fraction of reactions that were added during gap-filling and are not based on the genome annotation. Although the scope thus represents comprehensiveness and biological relevance of the GSM, it does not provide any information on the quality of the GSM predictions.

Phenotype predictions and scope both provide insight into the quality of a GSM, but are not suitable as sole evaluation criteria. The GSM evaluation criteria should assess GSMs on those aspects that are relevant for their intended applications. The applications and validation methods hardly overlap, however. For example, Oberhardt *et al.* described six different validation methods and thirteen applications, but only the prediction of essential genes falls in both categories [309]. Here, I will discuss a variety of potential evaluation criteria that are intended to probe different aspects of GSM quality.

The first set of criteria relate to the individual contents that make up a GSM. Each individual reaction should be elementally balanced, its direction should be thermodynamically supported, and there needs to be full provenance on why the reaction was included in the GSM. Each metabolite should be clearly linked to at least one major chemical database to avoid ambiguous metabolites (**chapter 3**). Without this information, the GSM validity of the GSM contents can not be independently confirmed.

The second set of criteria is that that there can be no cycles that arbitrarily spend or gain redox equivalents or energy-containing molecules such as ATP. These are basic requirements for the practical application of GSMs that may be missed if not explicitly assessed.

The next set of criteria relate to the qualitative GSM growth phenotype predictions. Besides the aforementioned wildtype and mutant growth phenotype predictions in different media, also broader phenotypic traits should be explicitly tested. For example: (i) Does the obligate aerobe need oxygen? (ii) Can the facultative anaerobe do without oxygen? (iii) Can the autotroph fix CO_2? (iv) Is there no CO_2 fixation if it should not be there? (v) Can the photoautotroph use light to generate energy? (vi) Can the chemolithotroph generate reducing equivalents from an inorganic substrate? These are species-specific tests that relate to their basic and defining characteristics.

Quantitative flux predictions are relevant for many applications, but are also substantially more difficult to assess. A commonly reported quantitative prediction for GSM evaluation is the growth rate. The growth rate is, however, not a suitable readout for quantitative flux predictions as (i) the maintenance

values are typically fit such that the GSM predicts a reasonable growth rate, and (ii) the growth rate is but a single summary value, and (iii) GSMs predict growth yields, not rates. The flux distributions that a GSM predicts are more wholly evaluated using fluxomics data. Fluxomics data are reaction rates deduced from experimental measurements of the metabolic processing of a ^{13}C-enriched substrate [12]. These reaction rates may appear straightforward to compare to FBA prediction, which are also reaction rates, but the reality is more complicated. The main complication is that FBA predictions correspond to an arbitrary point in a usually rather large optimal solution space (see Figure 1.2), and the real reaction rates may lie in a different point of this optimal solution space, or at any 'distance' outside of it in any direction. Therefore, the optimal solution space should either be shrunk using additional assumptions via, for example parsimonious FBA [249], or the comparison should focus on whether FBA predictions can, rather than do, match the experimentally measured rates. In addition, a scoring function will be required to summarize the extent of the match, but it is unclear what needs to be considered. For example: (i) If a predicted flux range is large, the inclusion of the 'real' value is more likely. Should it therefore be scored lower? (ii) If the predicted flux range does not include the real value, how should the score be penalized? Options are, for example, via absolute distance; distance on a log scale; or relative distance to the predicted flux range. (iii) Are distances measured to the mean of the predicted flux range or to its closest point? A suitable scoring function will have to be devised in future work by comparing the performance of scoring functions for various datasets, organisms, and GSMs.

Application-driven predictions are both the most relevant and the most difficult to assess. If a GSM is not evaluated for a specific application, it is an immense asumption that it can be used for it. On the other hand, it is not trivial to assess the ability of a GSM to be used for, for example, metabolic engineering. Despite the many GSM-based preoduction strain design methods [262], there is no straightforward process to evaluate GSMs based on previous experimental successes in metabolic engineering. This can partially be explained as experimentally common practices such as overexpression, downregulation and enzyme engineering are not directly compatible with GSMs as regulation and enzyme characteristics are not explicitly modeled. Nonetheless, GSMs should be able to predict the experimentally observed phenotype, similar to the flux predictions discussed above. If, however, a specific phenotype was obtained via the overexpression of a particular gene, it is reasonable that the GSM predictions should also steer towards that phenotype if a higher flux is forced through the corresponding reaction.

Evaluations based on application-driven GSM predictions are not going to be straightforward. There are no established commonly accepted protocols. There are no clear metrics. Nonetheless, the more straightforward and easier evaluation methods simply do not suffice to determine the quality of GSMs given the immense range of described applications. For each intended application a thorough review on common practices is warranted with the goal to define standard protocols for evaluation of GSMs based on the specifics of the intended application. A framework in which this could be organised is via an open challenge or crowdsourcing, as has been succesful for other systems biology disciplines [277].

GSMs in metabolic engineering

One of the most-described applications of GSMs is production strain design as part of a metabolic engineering effort. There are many GSM-based strain design methods that are covered in even more review papers that mostly suggest that GSM-driven strain design works like a charm. However, the number of actual applications appears to lag behind. A recent assessment of GSM-driven strain design methods found that only six out of the thirty-four considered methods have some experimental validation [262]. In total, this article refers to 14 cases of experimental validation of strains designed using GSMs. In other words, there are more than twice as many published methods as there are examples of any single of them working as intended. The underlying reason for the lack of success in GSM-driven strain design is not clear.

An obvious explanation could be that the methods simply do not work or that the currently available GSMs are not of sufficient quality. It is hard to argue against this explanation because, as aforementioned, GSMs are hardly evaluated on these criteria. For the field to advance it will be necessary to have a substantial number of testcases. Preferably these testcases regard situations in which different methods and GSMs propose alternative strategies to obtain a production strain for a specific compound. In particular, also cases where the strain design methods did not work or where the compound is not industrially interesting should be published in some form. Incorrect predictions carry information on the limitations and false assumptions underlying the used method and GSM. This information is crucial to determine the conditions under which GSM-driven strain design performs well, and how it can be further improved.

Another possibility is the lack of the systems biology 'design-build-test-improve' cycle in GSM-driven strain design. The GSM-based methods are used to design a strain that is expected to have a certain phenotype following experimental evolution. If the strain does not have the expected phenotype, there is no clear procedure on how to re-evaluate the designs using GSMs.

This ties in directly with the previous discussion on the evaluation of GSMs for metabolic engineering. If there is no method to redesign the system based on the experimental results, there is no real engineering cycle.

A third explanation, which is rather undocumented, is the lacking availability of many published methods. An overview of the availability of published strain design methods indicates that only fifteen out of thirty-two considered tools are accessible [262]. In my own experience, accessible does not imply proper functioning. This lack in available tools severely hampers their independent assessment and application. I strongly advocate for the need for computational articles to be accommodated by fully functional code that directly replicates the reported results. The tendency to not share code in systems biology results in a huge waste of scientific resources as any scientist who wants to use — or even try — a method will have to start from scratch with coding, testing, and optimizing.

It is interesting to consider this issue in the context of the publication-based incentive system in science. The creators of the method have published and thus have received their 'reward'. However, a potential user has no access and may not have the time or skill to recreate the method. I have on several occasions contacted the creators of a published method only to hear that it was either 'not yet available' or that 'we could collaborate'. In my opinion, the intent to only provide access to a method when rewarded with a co-authorship is detestable; and not in any way related to the colloborative pursuit of knowledge that science should represent. If, however, the potential user does have the time and skill to recreate the method, there is no incentive to share this with the scientific community as the merits of the publication and corresponding academic acknowledgement have already been awarded elsewhere. I advocate for an incentive to recreate and share versions of published yet publically unavailable tools. Such an incentive could be financially via funding, or perhaps recognition-based similar to recent efforts to acknowledge scientists who devote their time to the inherently not rewarding process of peer review [152].

Concluding remarks

GSMs have become commonplace for the study of metabolism (see Figure 1.1). The generation of GSMs has been simplified by the emergence of fully automatic tools, and GSMs have been successfully used for a large range of applications. GSM-based research may well become one of the most impactful fields in the next few decades with applications in industrial biotechnology, agriculture, fundamental biological research, and personalized medicine. Some modest progress towards these prospects have been made as part of this thesis, and I have outlined the major challenges I think systems biologists have to overcome for GSMs to fully enable the desired *à la carte* design of biological systems.

Summary

Ruben G. A. van Heck

The goals of this thesis are to increase the understanding of microbial metabolism and to functionally (re-)design microbial systems using Genome-Scale Metabolic models (GSMs). GSMs are species-specific knowledge repositories that can be used to predict metabolic activities for wildtype and genetically modified organisms. **Chapter 1** describes the assumptions associated with GSMs, the GSM generation process, common GSM analysis methods, and GSM-driven strain design methods. Thereby, **Chapter 1** provides a background for all other chapters. In this work, there is a focus on the metabolically versatile bacterium *Pseudomonas putida* **chapters 2,3,4,5,6**), but also other model microbes and biotechnologically or societally relevant microbes are considered (**chapters 3,4,6,7,8**).

GSMs are reflections of the genome annotation of the corresponding organism. For *P. putida*, the genome annotation that GSMs have been built on is more than ten years old. In **chapter 2**, this genome annotation was updated both on a structural and functional level using state-of-the-art annotation tools. A crucial part of the functional annotation relied on the most comprehensive *P. putida* GSM to date. This GSM was used to identify knowledge gaps in *P. putida* metabolism by determining the inconsistencies between its growth predictions and experimental measurements. Inconsistencies were found for 120 compounds that could be degraded by *P. putida in vitro* but not *in silico*. These compounds formed the basis for a targeted manual annotation process. Ultimately, suitable degradation pathways were identified for 86/120 as part of the functional reannotation of the *P. putida* genome.

For *P. putida* there are 3 independently generated GSMs, which is not uncommon for model organisms. These GSMs differ in generation procedure and represent different and complementary subsets of the knowledge on the metabolism of the organism. However, the differing generation procedures also makes it extremely cumbersome to compare their contents, let alone to combine them into a single consensus GSM. **Chapter 3** addresses this issue through the introduction of a computational tool for COnsensus Metabolic Model GENeration (COMMGEN). COMMGEN automatically identifies inconsistencies between independently generated GSMs and semi-automatically resolves them. Thereby, it greatly facilitates a detailed comparison of independently generated GSMs as well as the construction of consensus GSMs that more comprehensively describe the knowledge on the modeled organism.

GSMs can predict whether or not the corresponding organism and derived mutants can grow in a large variety of different growth conditions. In comparison, experimental data is extremely limited. For example, BIOLOG data describes growth phenotypes for one strain in a few hundred different media, and genome-wide gene essentially data is typically limited to a single growth medium. In **chapter 4** GSMs of multiple *Pseudomonas* species were used to

predict growth phenotypes for all possible single-gene-deletion mutants in all possible minimal growth media to determine conditionally and unconditionally essential genes. This simulated data was integrated with genomic data on 432 sequenced *Pseudomonas* species which revealed a clear link between the essentiality of a gene function and the persistence of the gene within the *Pseudomonas* genus.

Chapters 5 and **6** describe the use of GSMs to (re-)design microbial systems. *P. putida* is, despite its acknowledged versatile metabolism, an obligate aerobe. As the oxygen-requirement limits the potential applications of *P. putida*, there have been several experimental attempts to enable it to grow anaerobically, which have so far not succeeded. **Chapter 5** describes an *in silico* effort to determine why *P. putida* can not grow anaerobically using a combination of GSM analyses and comparative genomics. These analyses resulted in a shortlist of several essential and oxygen-dependent processes in *P. putida*. The identification of these processes has enabled the design of *P. putida* strains that can grow anaerobically based on the current understanding of *P. putida* metabolism as represented in GSMs.

Efficient microbial CO_2 fixation is a requirement for the biobased community, but the natural CO_2 fixation pathways are rather inefficient, while the synthetic CO_2 fixation pathways have been designed without considering the metabolic context of a target organism. **Chapter 6** introduces a computational tool, CO2FIX, that designs species-specific CO_2 fixation pathways based on GSMs and biochemical reaction databases. The designed pathways are evaluated for their ATP efficiency, thermodynamic feasibility, and kinetic rates. CO2FIX is applied to eight different organisms, which has led to the identification of both species-specific and general CO_2 fixation pathways that have promising features while requiring surprisingly few non-native reactions. Three of these pathways are described in detail.

In all previous chapters GSMs of relatively well-understood microbes have been used to gain further insight into their metabolism and to functionally (re-)design them. For complex microbial systems, such as algae (**chapter 7**) and gut microbial communities (**chapter 8**), GSMs are similarly useful, but substantially more difficult to create and analyze. Algae are widely considered as potential centerpieces of a biobased economy. **Chapter 7** reviews the current challenges in algal genome annotation, modeling and synthetic biology. The gut microbiota is an incredibly complex microbial system that is crucial to our well-being. **Chapter 8** reviews the ongoing developments in the modeling of both single gut microbes and gut microbial communities, and discusses how these developments will enable the move from studying correlation to causation, and ultimately the rational steering of gut microbial activity.

Chapter 9 discusses how the previous chapters contribute to the research goals of this thesis. In addition, it provides an extensive discussion on current GSM practices, the issues associated therewith, and how these issues can be tackled. In particular, the discussion focuses on issues related to: (i) The inability to distinguish between biological difference and GSM generation artifacts when using multiple GSMs, (ii) The lack of continuous GSM updates, (iii) The mismatch between what GSM predictions and experimental data represent, (iv) The need for standardization in GSM evaluation, and (v) The lack of experimental validation of GSM-driven strain design for metabolic engineering.

Bibliography

[1] Acevedo-Rocha, Carlos G et al. "From essential to persistent genes: a functional approach to constructing synthetic life". In: *Trends in Genetics* 29.5 (2013), pp. 273–279.

[2] Achaz, Guillaume et al. "Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin". In: *Molecular Biology and Evolution* 17.8 (2000), pp. 1268–1275.

[3] Agren, Rasmus et al. "The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum". In: *PLoS Comput Biol* 9.3 (2013), e1002980.

[4] Akashi, Hiroshi and Gojobori, Takashi. "Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis". In: *Proceedings of the National Academy of Sciences* 99.6 (2002), pp. 3695–3700.

[5] Alako, Blaise TF et al. "TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure". In: *Nucleic acids research* 34.suppl 2 (2006), W104–W109.

[6] Aller, Kadri et al. "Nutritional requirements and media development for Lactococcus lactis IL1403". In: *Applied microbiology and biotechnology* 98.13 (2014), pp. 5871–5881.

[7] Altschull, Stephen F and Madden Thomas L, Schaffer Alejandro A Zhang Jinghui Zhang Zheng Miller Webb; Lipman David J. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic Acids Research* 25.17 (1997), pp. 3389–3402.

[8] Ambati, Ranga Rao et al. "Astaxanthin: Sources, extraction, stability, biological activities and its commercial applications—A review". In: *Marine drugs* 12.1 (2014), pp. 128–152.

[9] Anders, Simon and Huber, Wolfgang. "Differential expression analysis for sequence count data". In: *Genome biology* 11.10 (2010), p. 1.

[10] Andersen, Mikael Rørdam, Nielsen, Michael Lynge, and Nielsen, Jens. "Metabolic model integration of the bibliome, genome, metabolome and reactome of Aspergillus niger". In: *Molecular Systems Biology* 4.1 (2008), p. 178.

[11] Antonia Molina-Henares, M. et al. "Identification of conditionally essential genes for growth of Pseudomonas putida KT2440 on minimal medium through the screening of a genome-wide mutant library". In: *Environmental Microbiology* 12.6 (2010), pp. 1468–1485.

[12] Antoniewicz, Maciek R. "Methods and advances in metabolic flux analysis: a mini-review". In: *Journal of industrial microbiology & biotechnology* 42.3 (2015), pp. 317–325.

[13] Antonovsky, Niv et al. “Sugar synthesis from CO 2 in Escherichia coli”. In: *Cell* 166.1 (2016), pp. 115–125.

[14] Arias, Sagrario et al. “Genetic analyses and molecular characterization of the pathways involved in the conversion of 2-phenylethylamine and 2-phenylethanol into phenylacetic acid in Pseudomonas putida U”. In: *Environmental microbiology* 10.2 (2008), pp. 413–432.

[15] Aung, Hnin W, Henry, Susan A, and Walker, Larry P. “Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism”. In: *Industrial Biotechnology* 9.4 (2013), pp. 215–228.

[16] Aziz, Ramy K et al. “The RAST Server: rapid annotations using subsystems technology”. In: *BMC genomics* 9.1 (2008), p. 1.

[17] Babaei, Parizad, Ghasemi-Kahrizsangi, Tahereh, and Marashi, Sayed-Amir. “Modeling the differences in biochemical capabilities of pseudomonas species by flux balance analysis: how good are genome-scale metabolic networks at predicting the differences?” In: *The Scientific World Journal* 2014 (2014).

[18] Bailly, J. et al. “Soil eukaryotic functional diversity, a metatranscriptomic approach”. In: *Isme Journal* 1.7 (2007), pp. 632–642.

[19] Baldrian, P. and Lopez-Mondejar, R. “Microbial genomics, transcriptomics and proteomics: new discoveries in decomposition research using complementary methods”. In: *Applied Microbiology and Biotechnology* 98.4 (2014), pp. 1531–1537.

[20] Ballal, Anand, Basu, Bhakti, and Apte, Shree Kumar. “The Kdp-ATPase system and its regulation”. In: *Journal of biosciences* 32.3 (2007), pp. 559–568.

[21] Baltrus, David A et al. “Dynamic evolution of pathogenicity revealed by sequencing and comparative genomics of 19 Pseudomonas syringae isolates”. In: *PLoS pathog* 7.7 (2011), e1002132.

[22] Bar-Even, Arren. “Formate assimilation: the metabolic architecture of natural and synthetic pathways”. In: *Biochemistry* 55.28 (2016), pp. 3851–3863.

[23] Bar-Even, Arren et al. “Design and analysis of synthetic carbon fixation pathways”. In: *Proceedings of the National Academy of Sciences* 107.19 (2010), pp. 8889–8894.

[24] Bar-Even, Arren et al. “Design and analysis of metabolic pathways supporting formatotrophic growth for electricity-dependent cultivation of microbes”. In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1827.8 (2013), pp. 1039–1047.

[25] Barrett, Tanya et al. “NCBI GEO: archive for functional genomics data sets—update”. In: *Nucleic acids research* 41.D1 (2013), pp. D991–D995.

[26] Barrick, Jeffrey E et al. “Genome evolution and adaptation in a long-term experiment with Escherichia coli”. In: *Nature* 461.7268 (2009), pp. 1243–1247.

[27] Bartell, Jennifer A. et al. “Comparative Metabolic Systems Analysis of Pathogenic Burkholderia”. In: *Journal of Bacteriology* 196.2 (2014), pp. 210–226.

[28] Bastard, Karine et al. “Revealing the hidden functional diversity of an enzyme family”. In: *Nature chemical biology* 10.1 (2014), pp. 42–49.

[29] Bay, Denice C and Turner, Raymond J. “Small multidrug resistance protein EmrE reduces host pH and osmotic tolerance to metabolic quaternary cation osmoprotectants”. In: *Journal of bacteriology* 194.21 (2012), pp. 5941–5948.

[30] Becker, Scott A and Palsson, Bernhard Ø. "Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation". In: *BMC microbiology* 5.1 (2005), p. 1.

[31] Becker, Scott A et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox". In: *Nature protocols* 2.3 (2007), pp. 727–738.

[32] Beckett, David, Berners-Lee, Tim, and Prud'hommeaux, Eric. "Turtle-terse RDF triple language". In: *W3C Team Submission* 14 (2008), p. 7.

[33] Belda, Eugeni et al. "The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis". In: *Environmental microbiology* (2016).

[34] Bennett, Bryson D et al. "Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli". In: *Nature chemical biology* 5.8 (2009), pp. 593–599.

[35] Berg, Ivan A. "Ecological aspects of the distribution of different autotrophic CO2 fixation pathways". In: *Applied and environmental microbiology* 77.6 (2011), pp. 1925–1936.

[36] Bernard, Thomas et al. "Reconciliation of metabolites and biochemical reactions for metabolic networks". In: *Briefings in bioinformatics* (2012), bbs058.

[37] Bertels, Frederic et al. "Automated reconstruction of whole-genome phylogenies from short-sequence reads". In: *Molecular biology and evolution* 31.5 (2014), pp. 1077–1088.

[38] Bertin, Yolande et al. "The gluconeogenesis pathway is involved in maintenance of enterohaemorrhagic Escherichia coli O157: H7 in bovine intestinal content". In: *PloS one* 9.6 (2014), e98367.

[39] Beste, Dany J. V. et al. "GSMN-TB: a web-based genome scale network model of Mycobacterium tuberculosis metabolism". In: *Genome Biology* 8.5 (2007).

[40] Bettenbrock, Katja et al. "Towards a systems level understanding of the oxygen response of Escherichia coli". In: *Adv Microb Physiol* 64 (2014), pp. 65–114.

[41] Blainey, P. C. "The future is now: single-cell genomics of bacteria and archaea". In: *Fems Microbiology Reviews* 37.3 (2013), pp. 407–427.

[42] Blazier, Anna S and Papin, Jason A. "Integration of expression data in genome-scale metabolic network reconstructions". In: *Frontiers in physiology* 3 (2012), p. 299.

[43] Bochner, Barry R, Gadzinski, Peter, and Panomitros, Eugenia. "Phenotype microarrays for high-throughput phenotypic testing and assay of gene function". In: *Genome research* 11.7 (2001), pp. 1246–1255.

[44] Bochner, BR. "Sleuthing out bacterial identities." In: *Nature* 339.6220 (1989), pp. 157–158.

[45] Bocs, Stephanie et al. "AMIGene: annotation of microbial genes". In: *Nucleic Acids Research* 31.13 (2003), pp. 3723–3726.

[46] Bogen, Christian et al. "Reconstruction of the lipid metabolism for the microalga Monoraphidium neglectum from its genome sequence reveals characteristics suitable for biofuel production". In: *BMC genomics* 14.1 (2013), p. 1.

[47] Bordbar, Aarash et al. "Constraint-based models predict metabolic and associated cellular functions". In: *Nature Reviews Genetics* 15.2 (2014), pp. 107–120.

[48] Borenstein, Elhanan. "Computational systems biology and in silico modeling of the human microbiome". In: *Briefings in bioinformatics* 13.6 (2012), pp. 769–780.

[49] Borgos, Sven EF et al. "Mapping global effects of the anti-sigma factor MucA in Pseudomonas fluorescens SBW25 through genome-scale metabolic modeling". In: *BMC systems biology* 7.1 (2013), p. 1.

[50] Borowitzka, Michael A. "High-value products from microalgae—their development and commercialisation". In: *Journal of Applied Phycology* 25.3 (2013), pp. 743–756.

[51] Boyle, Nanette R and Morgan, John A. "Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii". In: *BMC systems biology* 3.1 (2009), p. 1.

[52] Boyle, Nanette R and Morgan, John A. "Computation of metabolic fluxes and efficiencies for biological carbon dioxide fixation". In: *Metabolic engineering* 13.2 (2011), pp. 150–158.

[53] Boyle, Patrick M and Silver, Pamela A. "Parts plus pipes: synthetic biology approaches to metabolic engineering". In: *Metabolic engineering* 14.3 (2012), pp. 223–232.

[54] Boynton, Tye O et al. "Discovery of a gene involved in a third bacterial protoporphyrinogen oxidase activity through comparative genomic analysis and functional complementation". In: *Applied and environmental microbiology* 77.14 (2011), pp. 4795–4801.

[55] Brett, Christopher L, Donowitz, Mark, and Rao, Rajini. "Evolutionary origins of eukaryotic sodium/proton exchangers". In: *American Journal of Physiology-Cell Physiology* 288.2 (2005), pp. C223–C239.

[56] Browne, H. P. et al. "Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation". In: *Nature* 533.7604 (2016), pp. 543–+.

[57] Buchan, D. W. A. et al. "Protein annotation and modelling servers at University, College London". In: *Nucleic Acids Research* 38 (2010), W563–W568.

[58] Bui, T. P. N. et al. "Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal". In: *Nature Communications* 6 (2015).

[59] Burgard, Anthony P, Pharkya, Priti, and Maranas, Costas D. "Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization". In: *Biotechnology and bioengineering* 84.6 (2003), pp. 647–657.

[60] Burge, Sarah W et al. "Rfam 11.0: 10 years of RNA families". In: *Nucleic acids research* (2012), gks1005.

[61] Busi, María V et al. "Starch metabolism in green algae". In: *Starch-Stärke* 66.1-2 (2014), pp. 28–40.

[62] Campodonico, Miguel A et al. "Generation of an atlas for commodity chemical production in Escherichia coli and a novel pathway prediction algorithm, GEM-Path". In: *Metabolic engineering* 25 (2014), pp. 140–158.

[63] Carbonell, Pablo et al. "Retropath: automated pipeline for embedded metabolic circuits". In: *ACS synthetic biology* 3.8 (2013), pp. 565–577.

[64] Carmel, O et al. "The Na+-specific interaction between the LysR-type regulator, NhaR, and the nhaA gene encoding the Na+/H+ antiporter of Escherichia coli". In: *The EMBO journal* 16.19 (1997), pp. 5922–5929.

[65] Carrera, J. et al. "An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of Escherichia coli". In: *Molecular Systems Biology* 10.7 (2014).

[66] Caspi, Ron et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases". In: *Nucleic Acids Research* 42.D1 (2014), pp. D459–D471.

[67] Chandra, Govind, Chater, Keith F, and Bornemann, Stephen. "Unexpected and widespread connections between bacterial glycogen and trehalose metabolism". In: *Microbiology* 157.6 (2011), pp. 1565–1572.

[68] Chandrasekaran, S. and Price, N. D. "Metabolic Constraint-Based Refinement of Transcriptional Regulatory Networks". In: *Plos Computational Biology* 9.12 (2013).

[69] Chandrasekaran, Sriram and Price, Nathan D. "Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.41 (2010), pp. 17845–17850.

[70] Chang, Antje et al. "BRENDA in 2015: exciting developments in its 25th year of existence". In: *Nucleic acids research* (2014), gku1068.

[71] Chang, Roger L et al. "Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism". In: *Molecular systems biology* 7.1 (2011), p. 518.

[72] Chavarría, Max et al. "The Entner–Doudoroff pathway empowers Pseudomonas putida KT2440 with a high tolerance to oxidative stress". In: *Environmental microbiology* 15.6 (2013), pp. 1772–1785.

[73] Chen, Chiliang and Beattie, Gwyn A. "Pseudomonas syringae BetT is a low-affinity choline transporter that is responsible for superior osmoprotection by choline over glycine betaine". In: *Journal of bacteriology* 190.8 (2008), pp. 2717–2725.

[74] Chen, Chiliang et al. "The ATP-binding cassette transporter Cbc (choline/betaine/carnitine) recruits multiple substrate-binding proteins with strong specificity for distinct quaternary ammonium compounds". In: *Molecular microbiology* 75.1 (2010), pp. 29–45.

[75] Chien, Hungchien Roger et al. "Identification of the open reading frame for the Pseudomonas putidad-hydantoinase gene and expression of the gene in Escherichia coli". In: *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* 1395.1 (1998), pp. 68–77.

[76] Chindelevitch, Leonid et al. "MetaMerge: scaling up genome-scale metabolic reconstructions with application to Mycobacterium tuberculosis". In: *Genome Biology* 13.1 (2012).

[77] Chisti, Yusuf. "Constraints to commercialization of algal fuels". In: *Journal of biotechnology* 167.3 (2013), pp. 201–214.

[78] Cho, Ayoun et al. "Prediction of novel synthetic pathways for the production of desired chemicals". In: *BMC Systems Biology* 4.1 (2010), p. 1.

[79] Choby, Jacob E and Skaar, Eric P. "Heme Synthesis and Acquisition in Bacterial Pathogens". In: *Journal of molecular biology* (2016).
[80] Christian, Nils et al. "An integrative approach towards completing genome-scale metabolic networks". In: *Molecular bioSystems* 5.12 (2009), pp. 1889–1903.
[81] Chubiz, L. M. et al. "Species interactions differ in their genetic robustness". In: *Frontiers in Microbiology* 6 (2015).
[82] Claassens, Nico J. "A warm welcome for alternative CO2 fixation pathways in microbial biotechnology". In: *Microbial Biotechnology* (2016).
[83] Claassens, Nico J et al. "Harnessing the power of microbial autotrophy". In: *Nature Reviews Microbiology* (2016).
[84] Clarke, Patricia H. "The metabolic versatility of pseudomonads". In: *Antonie Van Leeuwenhoek* 48.2 (1982), pp. 105–130.
[85] Claudel-Renard, Clotilde et al. "Enzyme-specific profiles for genome annotation: PRIAM". In: *Nucleic acids research* 31.22 (2003), pp. 6633–6639.
[86] Cogne, Guillaume et al. "A model-based method for investigating bioenergetic processes in autotrophically growing eukaryotic microalgae: Application to the green algae Chlamydomonas reinhardtii". In: *Biotechnology progress* 27.3 (2011), pp. 631–640.
[87] Colijn, Caroline et al. "Interpreting Expression Data with Metabolic Flux Models: Predicting Mycobacterium tuberculosis Mycolic Acid Production". In: *Plos Computational Biology* 5.8 (2009).
[88] Consortium, UniProt et al. "UniProt: a hub for protein information". In: *Nucleic acids research* (2014), gku989.
[89] Consortium, UniProt et al. "UniProt: the universal protein knowledgebase". In: *Nucleic acids research* 45.D1 (2017), pp. D158–D169.
[90] Cook, Helen and Ussery, David W. "Sigma factors in a thousand E. coli genomes". In: *Environmental microbiology* 15.12 (2013), pp. 3121–3129.
[91] Cozzetto, Domenico et al. "Protein function prediction by massive integration of evolutionary analyses and multiple data sources". In: *BMC bioinformatics* 14.Suppl 3 (2013), S1.
[92] CPLEX, IBM ILOG CPLEX Optimization Studio. *12.6. 2*.
[93] Crespo, Anna et al. "Pseudomonas aeruginosa Exhibits Deficient Biofilm Formation in the Absence of Class II and III Ribonucleotide Reductases Due to Hindered Anaerobic Growth". In: *Frontiers in microbiology* 7 (2016).
[94] Dahl, Robert H et al. "Engineering dynamic pathway regulation using stress-response promoters". In: *Nature biotechnology* 31.11 (2013), pp. 1039–1046.
[95] Dailey, Tamara A et al. "Discovery and characterization of HemQ an essential heme biosynthetic pathway component". In: *Journal of Biological Chemistry* 285.34 (2010), pp. 25978–25986.
[96] Dam, Jesse CJ van et al. "RDF2Graph a tool to recover, understand and validate the ontology of an RDF resource". In: *Journal of biomedical semantics* 6.1 (2015), p. 1.
[97] Danchin, Antoine and Sekowska, Agnieszka. "The logic of metabolism and its fuzzy consequences". In: *Environmental microbiology* 16.1 (2014), pp. 19–28.

[98] Day, Anil and Goldschmidt-Clermont, Michel. "The chloroplast transformation toolbox: selectable markers and marker removal". In: *Plant biotechnology journal* 9.5 (2011), pp. 540–553.

[99] De Lorenzo, Víctor. "It's the metabolism, stupid!" In: *Environmental microbiology reports* 7.1 (2015), pp. 18–19.

[100] De Smet, Koen AL et al. "Three pathways for trehalose biosynthesis in mycobacteria". In: *Microbiology* 146.1 (2000), pp. 199–208.

[101] Demerec, M et al. "A proposal for a uniform nomenclature in bacterial genetics". In: *Genetics* 54.1 (1966), p. 61.

[102] Deng, Jingyuan. "An Integrated Machine-Learning Model to Predict Prokaryotic Essential Genes". In: *Gene Essentiality: Methods and Protocols* (2015), pp. 137–151.

[103] Deng, Jingyuan et al. "Investigating the predictability of essential genes across distantly related organisms using an integrative approach". In: *Nucleic acids research* 39.3 (2011), pp. 795–807.

[104] Desbois, Andrew P et al. "Isolation and structural characterisation of two antibacterial free fatty acids from the marine diatom, Phaeodactylum tricornutum". In: *Applied microbiology and biotechnology* 81.4 (2008), pp. 755–764.

[105] Dorrell, Richard G and Smith, Alison G. "Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates". In: *Eukaryotic cell* 10.7 (2011), pp. 856–868.

[106] Dötsch, Andreas et al. "The Pseudomonas aeruginosa transcriptional landscape is shaped by environmental heterogeneity and genetic variation". In: *Mbio* 6.4 (2015), e00749–15.

[107] Doucha, Jiři and Lívanskỳ, K. "Production of high-density Chlorella culture grown in fermenters". In: *Journal of applied phycology* 24.1 (2012), pp. 35–43.

[108] Dover, Nir and Padan, Etana. "Transcription of nhaA, the main Na+/H+ antiporter of Escherichia coli, is regulated by Na+ and growth phase". In: *Journal of bacteriology* 183.2 (2001), pp. 644–653.

[109] Dreyfuss, Jonathan M. et al. "Reconstruction and Validation of a Genome-Scale Metabolic Model for the Filamentous Fungus Neurospora crassa Using FARM". In: *Plos Computational Biology* 9.7 (2013).

[110] Duan, Jin et al. "The complete genome sequence of the plant growth-promoting bacterium Pseudomonas sp. UW4". In: *PloS one* 8.3 (2013), e58640.

[111] Duarte, Natalie C., Herrgard, Markus J., and Palsson, Bernhard O. "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model". In: *Genome Research* 14.7 (2004), pp. 1298–1309.

[112] Ebrahim, Ali et al. "COBRApy: constraints-based reconstruction and analysis for python". In: *BMC systems biology* 7.1 (2013), p. 74.

[113] Ekseth, Ole Kristian, Kuiper, Martin, and Mironov, Vladimir. "orthAgogue: an agile tool for the rapid prediction of orthology relations". In: *Bioinformatics* (2013), btt582.

[114] El-Semman, I. E. et al. "Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction". In: *BMC Syst Biol* 8 (2014), p. 41.

[115] Elbein, Alan D et al. "Last step in the conversion of trehalose to glycogen a mycobacterial enzyme that transfers maltose from maltose 1-phosphate to glycogen". In: *Journal of Biological Chemistry* 285.13 (2010), pp. 9803–9812.

[116] Emanuelsson, olof et al. "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence". In: *J. Mol. Biol.* 300.4 (2000), pp. 1005–1016.

[117] Engelen, Stefan et al. "Distinct co-evolution patterns of genes associated to DNA polymerase III DnaE and PolC". In: *BMC genomics* 13.1 (2012), p. 1.

[118] Engelhardt, Barbara E et al. "Genome-scale phylogenetic function annotation of large and diverse protein families". In: *Genome research* 21.11 (2011), pp. 1969–1980.

[119] Enright, Anton J, Van Dongen, Stijn, and Ouzounis, Christos A. "An efficient algorithm for large-scale detection of protein families". In: *Nucleic acids research* 30.7 (2002), pp. 1575–1584.

[120] Erb, Tobias J. "Carboxylases in natural and synthetic microbial pathways". In: *Applied and environmental microbiology* 77.24 (2011), pp. 8466–8477.

[121] Erb, Tobias J, Jones, Patrik R, and Bar-Even, Arren. "Synthetic metabolism: metabolic engineering meets enzyme design". In: *Current Opinion in Chemical Biology* 37 (2017), pp. 56–62.

[122] Eschbach, Martin et al. "Long-term anaerobic survival of the opportunistic pathogen Pseudomonas aeruginosa via pyruvate fermentation". In: *Journal of bacteriology* 186.14 (2004), pp. 4596–4604.

[123] Espey, M. G. "Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota". In: *Free Radical Biology and Medicine* 55 (2013), pp. 130–140.

[124] Falda, Marco et al. "Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms". In: *BMC bioinformatics* 13.4 (2012), p. 1.

[125] Fang, Gang, Rocha, Eduardo, and Danchin, Antoine. "How essential are nonessential genes?" In: *Molecular biology and evolution* 22.11 (2005), pp. 2147–2156.

[126] Fang, Wei et al. "Transcriptome-wide changes in Chlamydomonas reinhardtii gene expression regulated by carbon dioxide and the CO2-concentrating mechanism regulator CIA5/CCM1". In: *The Plant Cell* 24.5 (2012), pp. 1876–1893.

[127] Faria, J. P. et al. "Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models". In: *Briefings in Bioinformatics* 15.4 (2014), pp. 592–611.

[128] Fast, Alan G and Papoutsakis, Eleftherios T. "Stoichiometric and energetic analyses of non-photosynthetic CO 2-fixation pathways to support synthetic biology strategies for production of fuels and chemicals". In: *Current Opinion in Chemical Engineering* 1.4 (2012), pp. 380–395.

[129] Feist, Adam M and Palsson, Bernhard O. "The biomass objective function". In: *Current opinion in microbiology* 13.3 (2010), pp. 344–349.

[130] Feist, Adam M. et al. "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information". In: *Molecular Systems Biology* 3 (2007).

[131] Feist, Adam M et al. "Reconstruction of biochemical networks in microorganisms". In: *Nature Reviews Microbiology* 7.2 (2009), pp. 129–143.

[132] Feist, Adam M et al. "Constraint-based modeling of carbon fixation and the energetics of electron transfer in Geobacter metallireducens". In: *PLoS Comput Biol* 10.4 (2014), e1003575.

[133] Feng, Xueyang et al. "MicrobesFlux: a web platform for drafting metabolic models from the KEGG database". In: *BMC systems biology* 6.1 (2012), p. 94.

[134] Feng, Yujie, Li, Chao, and Zhang, Dawei. "Lipid production of Chlorella vulgaris cultured in artificial wastewater medium". In: *Bioresource Technology* 102.1 (2011), pp. 101–105.

[135] Filiatrault, Melanie J et al. "Identification of Pseudomonas aeruginosa genes involved in virulence and anaerobic growth". In: *Infection and immunity* 74.7 (2006), pp. 4237–4245.

[136] Finn, Robert D, Clements, Jody, and Eddy, Sean R. "HMMER web server: interactive sequence similarity searching". In: *Nucleic acids research* (2011), gkr367.

[137] Fishov, Itzhak, Zaritsky, Arieh, and Grover, NB. "On microbial states of growth". In: *Molecular microbiology* 15.5 (1995), pp. 789–794.

[138] Flamholz, Avi et al. "eQuilibrator—the biochemical thermodynamics calculator". In: *Nucleic acids research* (2011), gkr874.

[139] Flint, H. J. et al. "The role of the gut microbiota in nutrition and health". In: *Nat Rev Gastroenterol Hepatol* 9.10 (2012), pp. 577–89.

[140] Fong, Stephen S and Palsson, Bernhard Ø. "Metabolic gene–deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes". In: *Nature genetics* 36.10 (2004), pp. 1056–1058.

[141] Fong, Stephen S et al. "In silico design and adaptive evolution of Escherichia coli for production of lactic acid". In: *Biotechnology and bioengineering* 91.5 (2005), pp. 643–648.

[142] Foster, John W. "When protons attack: microbial strategies of acid adaptation". In: *Current opinion in microbiology* 2.2 (1999), pp. 170–174.

[143] Foster, John W. "Escherichia coli acid resistance: tales of an amateur acidophile". In: *Nature Reviews Microbiology* 2.11 (2004), pp. 898–907.

[144] Franceschini, Andrea et al. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration". In: *Nucleic acids research* 41.D1 (2013), pp. D808–D815.

[145] Frank, Sarah et al. "Pseudomonas putida KT2440 genome update by cDNA sequencing and microarray transcriptomics". In: *Environmental microbiology* 13.5 (2011), pp. 1309–1326.

[146] Freilich, S. et al. "Competitive and cooperative metabolic interactions in bacterial communities". In: *Nat Commun* 2 (2011), p. 589.

[147] Frunzke, Kurt and Meyer, Ortwin. "Nitrate respiration, denitrification, and utilization of nitrogen sources by aerobic carbon monoxide-oxidizing bacteria". In: *Archives of Microbiology* 154.2 (1990), pp. 168–174.

[148] Galperin, Michael Y et al. "Expanded microbial genome coverage and improved protein family annotation in the COG database". In: *Nucleic acids research* (2014), gku1223.

[149] Ganter, Mathias, Kaltenbach, Hans-Michael, and Stelling, Jörg. "Predicting network functions with nested patterns". In: *Nature communications* 5 (2014).
[150] Ganter, Mathias et al. "MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks". In: *Bioinformatics* 29.6 (2013), pp. 815–816.
[151] Gao, Chunfang et al. "Oil accumulation mechanisms of the oleaginous microalga Chlorella protothecoides revealed through its genome, transcriptomes, and proteomes". In: *BMC genomics* 15.1 (2014), p. 1.
[152] Gasparyan, Armen Yuri et al. "Rewarding peer reviewers: maintaining the integrity of science communication". In: *Journal of Korean medical science* 30.4 (2015), pp. 360–364.
[153] Gaßel, Michael et al. "The KdpF subunit is part of the K+-translocating Kdp complex of Escherichia coli and is responsible for stabilization of the complex in vitro". In: *Journal of Biological Chemistry* 274.53 (1999), pp. 37901–37907.
[154] Gill, S. R. et al. "Metagenomic analysis of the human distal gut microbiome". In: *Science* 312.5778 (2006), pp. 1355–1359.
[155] Golovan, Serguei et al. "Characterization and overproduction of the Escherichia coli appA encoded bifunctional enzyme that exhibits both phytase and acid phosphatase activities". In: *Canadian journal of microbiology* 46.1 (1999), pp. 59–71.
[156] Goodstein, David M et al. "Phytozome: a comparative platform for green plant genomics". In: *Nucleic acids research* 40.D1 (2012), pp. D1178–D1186.
[157] Gralla, Jay D and Vargas, David R. "Potassium glutamate as a transcriptional inhibitor during bacterial osmoregulation". In: *The EMBO journal* 25.7 (2006), pp. 1515–1521.
[158] Greenblum, S., Turnbaugh, P. J., and Borenstein, E. "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.2 (2012), pp. 594–599.
[159] Gross, Harald and Loper, Joyce E. "Genomics of secondary metabolite production by Pseudomonas spp." In: *Natural product reports* 26.11 (2009), pp. 1408–1446.
[160] Guarnieri, Michael T et al. "Proteomic analysis of Chlorella vulgaris: Potential targets for enhanced lipid accumulation". In: *Journal of proteomics* 93 (2013), pp. 245–253.
[161] Gulyui, MF and Silonova, NV. "Various metabolic reactions of formate in animal tissues". In: *Ukrainskii biokhimicheskii zhurnal (1978)* 59.4 (1986), pp. 29–35.
[162] Guo, An Chi et al. "ECMDB: the E. coli Metabolome Database". In: *Nucleic acids research* 41.D1 (2013), pp. D625–D630.
[163] Gurobi Optimization, Inc. *Gurobi Optimizer Reference Manual*. 2016.
[164] Hahn, Heinz P. "The type-4 pilus is the major virulence-associated adhesin of Pseudomonas aeruginosa–a review". In: *Gene* 192.1 (1997), pp. 99–108.
[165] Hamer, H. M. et al. "Review article: the role of butyrate on colonic function". In: *Alimentary Pharmacology and Therapeutics* 27.2 (2008), pp. 104–119.
[166] Harcombe, W. "Novel Cooperation Experimentally Evolved between Species". In: *Evolution* 64.7 (2010), pp. 2166–2172.

[167] Harcombe, W. R. et al. "Metabolic Resource Allocation in Individual Microbes Determines Ecosystem Interactions and Spatial Dynamics". In: *Cell Reports* 7.4 (2014), pp. 1104–1115.

[168] Hastings, Janna et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". In: *Nucleic acids research* 41.D1 (2013), pp. D456–D463.

[169] Hatzimanikatis, Vassily et al. "Exploring the diversity of complex metabolic networks". In: *Bioinformatics* 21.8 (2005), pp. 1603–1609.

[170] Heavner, Benjamin D et al. "Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance". In: *Database* 2013 (2013), bat059.

[171] Heck, Ruben GA van et al. "Efficient Reconstruction of Predictive Consensus Metabolic Network Models". In: *PLOS Comput Biol* 12.8 (2016), e1005085.

[172] Heinemann, Matthias et al. "In silico genome-scale reconstruction and validation of the Staphylococcus aureus metabolic network". In: *Biotechnology and bioengineering* 92.7 (2005), pp. 850–864.

[173] Heinken, Almut et al. "Functional metabolic map of Faecalibacterium prausnitzii, a beneficial human gut microbe". In: *Journal of bacteriology* 196.18 (2014), pp. 3289–3302.

[174] Helliwell, Katherine E et al. "Unraveling vitamin B12-responsive gene regulation in algae". In: *Plant physiology* 165.1 (2014), pp. 388–397.

[175] Henry, Christopher S. et al. "iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations". In: *Genome Biology* 10.6 (2009).

[176] Henry, Christopher S. et al. "High-throughput generation, optimization and analysis of genome-scale metabolic models". In: *Nature Biotechnology* 28.9 (2010), 977–U22.

[177] Herrgard, Markus J. et al. "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology". In: *Nature Biotechnology* 26.10 (2008), pp. 1155–1160.

[178] Hettne, Kristina M. et al. "A dictionary to identify small molecules and drugs in free text". In: *Bioinformatics* 25.22 (2009), pp. 2983–2991.

[179] Hidese, Ryota et al. "Escherichia coli dihydropyrimidine dehydrogenase is a novel NAD-dependent heterotetramer essential for the production of 5, 6-dihydrouracil". In: *Journal of bacteriology* 193.4 (2011), pp. 989–993.

[180] Hildebrand, Mark et al. "Metabolic and cellular organization in evolutionarily diverse microalgae as related to biofuels production". In: *Current opinion in chemical biology* 17.3 (2013), pp. 506–514.

[181] Ho, Shih-Hsin et al. "Bioethanol production using carbohydrate-rich microalgae biomass as feedstock". In: *Bioresource Technology* 135 (2013), pp. 191–198.

[182] Hoskisson, Paul A and Hobbs, Glyn. "Continuous culture–making a comeback?" In: *Microbiology* 151.10 (2005), pp. 3153–3159.

[183] Hucka, M. et al. "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models". In: *Bioinformatics* 19.4 (2003), pp. 524–531.

[184] Hui, Sheng et al. "Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria". In: *Molecular systems biology* 11.2 (2015), p. 784.
[185] Hyatt, Doug et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification". In: *BMC bioinformatics* 11.1 (2010), p. 1.
[186] Ibarra, Rafael U., Edwards, Jeremy S., and Palsson, Bernhard O. "Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth". In: *Nature* 420.6912 (2002), pp. 186–9.
[187] Imam, Saheed et al. "iRsp1095: a genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network". In: *BMC systems biology* 5.1 (2011), p. 116.
[188] Imam, Saheed et al. "Quantifying the effects of light intensity on bioproduction and maintenance energy during photosynthetic growth of Rhodobacter sphaeroides". In: *Photosynthesis research* 123.2 (2015), pp. 167–182.
[189] Jamshidi, Neema and Palsson, Bernhard O. "Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets". In: *Bmc Systems Biology* 1 (2007).
[190] Jensen, Paul A, Lutz, Kyla A, and Papin, Jason A. "TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks". In: *BMC systems biology* 5.1 (2011), p. 147.
[191] Ji, B. and Nielsen, J. "From next-generation sequencing to systematic modeling of the gut microbiome". In: *Front Genet* 6 (2015), p. 219.
[192] Jiménez, José I et al. "Deciphering the genetic determinants for aerobic nicotinic acid degradation: the nic cluster from Pseudomonas putida KT2440". In: *Proceedings of the National Academy of Sciences* 105.32 (2008), pp. 11329–11334.
[193] Jiménez, José Ignacio et al. "Genomic analysis of the aromatic catabolic pathways from Pseudomonas putida KT2440". In: *Environmental microbiology* 4.12 (2002), pp. 824–841.
[194] Jones, Philip et al. "InterProScan 5: genome-scale protein function classification". In: *Bioinformatics* 30.9 (2014), pp. 1236–1240.
[195] Jong, Louis E Den Dooren de. "Bijdrage tot de kennis van het mineralisatieproces". PhD thesis. TU Delft, Delft University of Technology, 1926.
[196] Jorquera, Orlando et al. "Comparative energy life-cycle analyses of microalgal biomass production in open ponds and photobioreactors". In: *Bioresource Technology* 101.4 (2010), pp. 1406–1413.
[197] Jude, Florence et al. "Posttranscriptional control of quorum-sensing-dependent virulence genes by DksA in Pseudomonas aeruginosa". In: *Journal of bacteriology* 185.12 (2003), pp. 3558–3566.
[198] Kaasen, Inga et al. "Molecular cloning and physical mapping of the otsBA genes, which encode the osmoregulatory trehalose pathway of Escherichia coli: evidence that transcription is activated by katF (AppR)". In: *Journal of bacteriology* 174.3 (1992), pp. 889–898.
[199] Kajiyama, Yusuke et al. "Complex formation by the mrpABCDEFG gene products, which constitute a principal Na+/H+ antiporter in Bacillus subtilis". In: *Journal of bacteriology* 189.20 (2007), pp. 7511–7514.
[200] Kanehisa, Minoru et al. "Data, information, knowledge and principle: back to metabolism in KEGG". In: *Nucleic acids research* 42.D1 (2014), pp. D199–D205.

[201] Kanehisa, Minoru et al. "KEGG: new perspectives on genomes, pathways, diseases and drugs". In: *Nucleic Acids Research* 45.D1 (2017), pp. D353–D361.

[202] Karpe, Peter D, Latendresse, Mario, and Caspi, Ron. "The pathway tools pathway prediction algorithm". In: *Standards in genomic sciences* 5.3 (2011), p. 424.

[203] Karr, J. R. et al. "A whole-cell computational model predicts phenotype from genotype". In: *Cell* 150.2 (2012), pp. 389–401.

[204] Kaushik, Jai K and Bhat, Rajiv. "Why is trehalose an exceptional protein stabilizer? An analysis of the thermal stability of proteins in the presence of the compatible osmolyte trehalose". In: *Journal of Biological Chemistry* 278.29 (2003), pp. 26458–26465.

[205] Kell, Douglas B and Oliver, Stephen G. "How drugs get into cells: tested and testable predictions to help discriminate between transporter-mediated uptake and lipoidal bilayer diffusion". In: *Frontiers in pharmacology* 5 (2014), p. 231.

[206] Keller, Oliver et al. "A novel hybrid gene prediction method employing protein multiple sequence alignments". In: *Bioinformatics* 27.6 (2011), pp. 757–763.

[207] Kelly, C. R. et al. "Fecal microbiota transplant for treatment of Clostridium difficile infection in immunocompromised patients". In: *Am J Gastroenterol* 109.7 (2014), pp. 1065–71.

[208] Kersey, Paul Julian et al. "Ensembl Genomes 2016: more genomes, more complexity". In: *Nucleic acids research* 44.D1 (2016), pp. D574–D580.

[209] Keseler, Ingrid M et al. "EcoCyc: fusing model organism databases with systems biology". In: *Nucleic acids research* 41.D1 (2013), pp. D605–D612.

[210] Kevles, Daniel J. "Ananda Chakrabarty wins a patent: Biotechnology, law, and society, 1972-1980". In: *Historical studies in the Physical and Biological sciences* 25.1 (1994), pp. 111–135.

[211] Kilian, Oliver et al. "High-efficiency homologous recombination in the oil-producing alga Nannochloropsis sp." In: *Proceedings of the National Academy of Sciences* 108.52 (2011), pp. 21265–21269.

[212] Kim, J. and Reed, J. L. "Refining metabolic models and accounting for regulatory effects". In: *Current Opinion in Biotechnology* 29 (2014), pp. 34–38.

[213] Kim, Joonhoon, Reed, Jennifer L, and Maravelias, Christos T. "Large-scale bi-level strain design approaches and mixed-integer programming solution techniques". In: *PLoS One* 6.9 (2011), e24162.

[214] Kim, Juhyun et al. "Transcriptomic fingerprinting of Pseudomonas putida under alternative physiological regimes". In: *Environmental microbiology reports* 5.6 (2013), pp. 883–891.

[215] Kim, Min Kyung and Lun, Desmond S. "Methods for integration of transcriptomic data in genome-scale metabolic models". In: *Computational and Structural Biotechnology Journal* 11.18 (2014), pp. 59–65.

[216] Kim, Young Hwan et al. "Analysis of aromatic catabolic pathways in Pseudomonas putida KT 2440 using a combined proteomic approach: 2-DE/MS and cleavable isotope-coded affinity tag analysis". In: *Proteomics* 6.4 (2006), pp. 1301–1318.

[217] King, Zachary A et al. "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". In: *Nucleic acids research* 44.D1 (2016), pp. D515–D522.

[218] Kirst, Henning et al. "Truncated photosystem chlorophyll antenna size in the green microalga Chlamydomonas reinhardtii upon deletion of the TLA3-CpSRP43 gene". In: *Plant physiology* 160.4 (2012), pp. 2251–2260.

[219] Kleerebezem, Michiel et al. "Complete genome sequence of Lactobacillus plantarum WCFS1". In: *Proceedings of the National Academy of Sciences* 100.4 (2003), pp. 1990–1995.

[220] Kliphuis, Anna MJ et al. "Metabolic modeling of Chlamydomonas reinhardtii: energy requirements for photoautotrophic growth and maintenance". In: *Journal of applied phycology* 24.2 (2012), pp. 253–266.

[221] Klitgord, N. and Segre, D. "Environments that Induce Synthetic Microbial Ecosystems". In: *Plos Computational Biology* 6.11 (2010).

[222] Klok, Anne J et al. "Simultaneous growth and neutral lipid accumulation in microalgae". In: *Bioresource technology* 134 (2013), pp. 233–243.

[223] Kobayashi, Kazuo et al. "Gene analysis of trehalose-producing enzymes from hyperthermophilic archaea in Sulfolobales". In: *Bioscience, biotechnology, and biochemistry* 60.10 (1996), pp. 1720–1723.

[224] Koehorst, Jasper J et al. "Comparison of 432 Pseudomonas strains through integration of genomic, functional, metabolic and expression data". In: *Scientific Reports* 6 (2016).

[225] Koehorst, Jasper J et al. "Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics". In: *F1000Research* 5 (2016).

[226] Kolinko, S. et al. "Single-cell genomics of uncultivated deep-branching magnetotactic bacteria reveals a conserved set of magnetosome genes". In: *Environ Microbiol* (2015).

[227] Korshunov, Sergei and Imlay, James A. "Two sources of endogenous hydrogen peroxide in Escherichia coli". In: *Molecular microbiology* 75.6 (2010), pp. 1389–1401.

[228] Kourmpetis, Yiannis AI et al. "Bayesian Markov Random Field analysis for protein function prediction based on network data". In: *PloS one* 5.2 (2010), e9293.

[229] Kow, Rebecca L et al. "Disruption of the proton relay network in the class 2 dihydroorotate dehydrogenase from Escherichia coli". In: *Biochemistry* 48.41 (2009), pp. 9801–9809.

[230] Krogh, Anders et al. "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes". In: *Journal of molecular biology* 305.3 (2001), pp. 567–580.

[231] Krumholz, Elias W et al. "Genome-wide metabolic network reconstruction of the picoalga Ostreococcus". In: *Journal of experimental botany* (2011), err407.

[232] Kuepfer, Lars, Sauer, Uwe, and Blank, Lars M. "Metabolic functions of duplicate genes in Saccharomyces cerevisiae". In: *Genome Research* 15.10 (2005), pp. 1421–1430.

[233] Kumar, Akhil, Suthers, Patrick F., and Maranas, Costas D. "MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases". In: *Bmc Bioinformatics* 13 (2012).

[234] Kumar, Vinay Satish, Dasika, Madhukar S, and Maranas, Costas D. "Optimization based automated curation of metabolic reconstructions". In: *BMC bioinformatics* 8.1 (2007), p. 1.

[235] Kumar, Vinay Satish and Maranas, Costas D. "GrowMatch: an automated method for reconciling in silico/in vivo growth predictions". In: *PLoS Comput Biol* 5.3 (2009), e1000308.

[236] La Rosa, Ruggero, Nogales, Juan, and Rojo, Fernando. "The Crc/CrcZ-CrcY global regulatory system helps the integration of gluconeogenic and glycolytic metabolism in Pseudomonas putida". In: *Environmental microbiology* 17.9 (2015), pp. 3362–3378.

[237] Lagesen, Karin et al. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes". In: *Nucleic acids research* 35.9 (2007), pp. 3100–3108.

[238] Lai, Bin et al. "Anoxic metabolism and biochemical production in Pseudomonas putida F1 driven by a bioelectrochemical system". In: *Biotechnology for biofuels* 9.1 (2016), p. 1.

[239] Lasken, R. S. "Genomic sequencing of uncultured microorganisms from single cells". In: *Nature Reviews Microbiology* 10.9 (2012), pp. 631–640.

[240] Lawrence, Michael et al. "Software for computing and annotating genomic ranges". In: *PLoS Comput Biol* 9.8 (2013), e1003118.

[241] Lee, Byungwook and Lee, Doheon. "Protein comparison at the domain architecture level". In: *BMC bioinformatics* 10.15 (2009), p. 1.

[242] Lee, Jin-Ho et al. "Cloning and expression of a trehalose synthase from Pseudomonas stutzeri CJ38 in Escherichia coli for the production of trehalose". In: *Applied microbiology and biotechnology* 68.2 (2005), pp. 213–219.

[243] Lee, Samuel A et al. "General and condition-specific essential functions of Pseudomonas aeruginosa". In: *Proceedings of the National Academy of Sciences* 112.16 (2015), pp. 5189–5194.

[244] Lehr, Florian and Posten, Clemens. "Closed photo-bioreactors as tools for biofuel production". In: *Current opinion in biotechnology* 20.3 (2009), pp. 280–285.

[245] Leprince, Audrey et al. "Streamlining of a Pseudomonas putida genome using a combinatorial deletion method based on minitransposon insertion and the Flp-FRT recombination system". In: *Synthetic Gene Networks: Methods and Protocols* (2012), pp. 249–266.

[246] Levy, R. and Borenstein, E. "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules". In: *Proc Natl Acad Sci U S A* 110.31 (2013), pp. 12804–9.

[247] Lewinson, Oded, Padan, Etana, and Bibi, Eitan. "Alkalitolerance: a biological function for a multidrug transporter in pH homeostasis". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.39 (2004), pp. 14073–14078.

[248] Lewis, N. E., Nagarajan, H., and Palsson, B. O. "Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods". In: *Nature Reviews Microbiology* 10.4 (2012), pp. 291–305.

[249] Lewis, Nathan E et al. "Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models". In: *Molecular systems biology* 6.1 (2010), p. 390.

[250] Li, Heng et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[251] Liberati, Nicole T et al. "An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.8 (2006), pp. 2833–2838.

[252] Loeschcke, Anita and Thies, Stephan. "Pseudomonas putida—a versatile host for the production of natural products". In: *Applied microbiology and biotechnology* 99.15 (2015), pp. 6197–6214.

[253] Loper, Joyce E et al. "Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions". In: *PLoS Genet* 8.7 (2012), e1002784.

[254] Lopez, David et al. "Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data". In: *BMC bioinformatics* 12.1 (2011), p. 282.

[255] Lowe, Todd M and Eddy, Sean R. "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence". In: *Nucleic acids research* 25.5 (1997), pp. 955–964.

[256] Lubitz, Timo et al. "SBtab: a flexible table format for data exchange in systems biology". In: *Bioinformatics* (2016), btw179.

[257] Lund, Peter, Tramonti, Angela, and De Biase, Daniela. "Coping with low pH: molecular strategies in neutralophilic bacteria". In: *FEMS microbiology reviews* 38.6 (2014), pp. 1091–1125.

[258] M. Mata, T. eresa, A. Martins, A. ntonio, and S. Caetano, N. idia. "Microalgae for biodiesel production and other applications: A review". In: *Renewable and Sustainable Energy Rev.* 14.1 (2010), pp. 217–232.

[259] Machado, Daniel and Herrgård, Markus. "Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism". In: *PLoS Comput Biol* 10.4 (2014), e1003580.

[260] MacMillan, Susan V et al. "The ion coupling and organic substrate specificities of osmoregulatory transporter ProP in Escherichia coli". In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1420.1 (1999), pp. 30–44.

[261] Mahadevan, R. and Schilling, C. H. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". In: *Metabolic Engineering* 5.4 (2003), pp. 264–276.

[262] Maia, Paulo, Rocha, Miguel, and Rocha, Isabel. "In silico constraint-based strain optimization methods: the quest for optimal cell factories". In: *Microbiology and Molecular Biology Reviews* 80.1 (2016), pp. 45–67.

[263] Manichaikul, Ani et al. "Metabolic network analysis integrated with transcript verification for sequenced genomes". In: *Nature methods* 6.8 (2009), p. 589.

[264] Martínez-García, Esteban et al. "Pseudomonas 2.0: genetic upgrading of P. putida KT2440 as an enhanced host for heterologous gene expression". In: *Microbial cell factories* 13.1 (2014), p. 1.

[265] MATLAB. Natick, Massachusetts, 2015.

[266] Matthews, Brian W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[267] Mattozzi, Matthew d et al. "Expression of the sub-pathways of the Chloroflexus aurantiacus 3-hydroxypropionate carbon fixation bicycle in E. coli: Toward horizontal transfer of autotrophic growth". In: *Metabolic engineering* 16 (2013), pp. 130–139.

[268] Maurice, C. F., Haiser, H. J., and Turnbaugh, P. J. "Xenobiotics Shape the Physiology and Gene Expression of the Active Human Gut Microbiome". In: *Cell* 152.1-2 (2013), pp. 39–50.

[269] May, Patrick et al. "Metabolomics-and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii". In: *Genetics* 179.1 (2008), pp. 157–166.

[270] May, Patrick et al. "ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii". In: *Bmc Genomics* 10.1 (2009), p. 209.

[271] McCloskey, Douglas, Palsson, Bernhard Ø, and Feist, Adam M. "Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli". In: *Molecular systems biology* 9.1 (2013), p. 661.

[272] Médigue, Claudine et al. "The secE gene of Helicobacter pylori". In: *Journal of bacteriology* 184.10 (2002), pp. 2837–2840.

[273] Medini, Duccio et al. "The microbial pan-genome". In: *Current opinion in genetics & development* 15.6 (2005), pp. 589–594.

[274] Merchant, Sabeeha S et al. "The Chlamydomonas genome reveals the evolution of key animal and plant functions". In: *Science* 318.5848 (2007), pp. 245–250.

[275] Merchant, Sabeeha S et al. "TAG, You're it! Chlamydomonas as a reference organism for understanding algal triacylglycerol accumulation". In: *Current opinion in biotechnology* 23.3 (2012), pp. 352–363.

[276] Meyer, Folker, Overbeek, Ross, and Rodriguez, Alex. "FIGfams: yet another set of protein families". In: *Nucleic acids research* 37.20 (2009), pp. 6643–6654.

[277] Meyer, Pablo et al. "Verification of systems biology research in the age of collaborative competition". In: *Nature biotechnology* 29.9 (2011), p. 811.

[278] Mikoulinskaia, Galina V et al. "Identification, cloning, and expression of bacteriophage T5 dnk gene encoding a broad specificity deoxyribonucleoside monophosphate kinase (EC 2.7. 4.13)". In: *Protein expression and purification* 33.2 (2004), pp. 166–175.

[279] Milanesio, Paola et al. "Regulatory exaptation of the catabolite repression protein (Crp)–cAMP system in Pseudomonas putida". In: *Environmental microbiology* 13.2 (2011), pp. 324–339.

[280] Mintz-Oron, Shira et al. "Network-based prediction of metabolic enzymes' subcellular localization". In: *Bioinformatics* 25.12 (2009), pp. I247–I252.

[281] Mitchell, Alex et al. "The InterPro protein families database: the classification resource after 15 years". In: *Nucleic acids research* (2014), gku1243.

[282] Mo, Monica L., Palsson, Bernhard O., and Herrgard, Markus J. "Connecting extracellular metabolomic measurements to intracellular flux states in yeast". In: *Bmc Systems Biology* 3 (2009).

[283] Monk, Jonathan, Nogales, Juan, and Palsson, Bernhard O. "Optimizing genome-scale network reconstructions". In: *Nature Biotechnology* 32.5 (2014), pp. 447–452.

[284] Monk, Jonathan M. et al. "Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.50 (2013), pp. 20338–20343.

[285] Moore, Edward RB et al. "Nonmedical: pseudomonas". In: *The prokaryotes*. Springer, 2006, pp. 646–703.

[286] Morgat, Anne et al. "Updates in Rhea—a manually curated resource of biochemical reactions". In: *Nucleic acids research* (2014), gku961.

[287] Morris, Cindy E et al. "The life history of the plant pathogen Pseudomonas syringae is linked to the water cycle". In: *The ISME journal* 2.3 (2008), pp. 321–334.

[288] Mosquera-Rendón, Jeanneth et al. "Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species". In: *BMC genomics* 17.1 (2016), p. 1.

[289] Nancy, Y Yu et al. "PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea". In: *Nucleic acids research* (2010), gkq1093.

[290] Nelson, KE et al. "Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440". In: *Environmental microbiology* 4.12 (2002), pp. 799–808.

[291] Nielsen, H. B. et al. "Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes". In: *Nature Biotechnology* 32.8 (2014), pp. 822–828.

[292] Nikel, Pablo I and Lorenzo, Víctor de. "Engineering an anaerobic metabolic regime in Pseudomonas putida KT2440 for the anoxic biodegradation of 1, 3-dichloroprop-1-ene". In: *Metabolic engineering* 15 (2013), pp. 98–112.

[293] Nikel, Pablo I, Martínez-García, Esteban, and Lorenzo, Víctor de. "Biotechnological domestication of pseudomonads using synthetic biology". In: *Nature Reviews Microbiology* 12.5 (2014), pp. 368–379.

[294] Nikel, Pablo I, Pérez-Pantoja, Danilo, and Lorenzo, Víctor de. "Pyridine nucleotide transhydrogenases enable redox balance of Pseudomonas putida during biodegradation of aromatic compounds". In: *Environmental Microbiology* 18.10 (2016), pp. 3565–3582.

[295] Nikel, Pablo I et al. "Pseudomonas putida KT2440 strain metabolizes glucose through a cycle formed by enzymes of the Entner-Doudoroff, Embden-Meyerhof-Parnas, and pentose phosphate pathways". In: *Journal of Biological Chemistry* 290.43 (2015), pp. 25920–25932.

[296] Nikel, Pablo I et al. "From dirt to industrial applications: Pseudomonas putida as a Synthetic Biology chassis for hosting harsh biochemical reactions". In: *Current opinion in chemical biology* 34 (2016), pp. 20–29.

[297] Ning, Zemin, Cox, Anthony J, and Mullikin, James C. "SSAHA: a fast search method for large DNA databases". In: *Genome research* 11.10 (2001), pp. 1725–1729.

[298] Nogales, Juan, Palsson, Bernhard Ø, and Thiele, Ines. "A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: i JN746 as a cell factory". In: *BMC systems biology* 2.1 (2008), p. 1.

[299] Nogales, Juan et al. "Molecular characterization of the gallate dioxygenase from Pseudomonas putida KT2440 the prototype of a new subgroup of extradiol dioxygenases". In: *Journal of Biological Chemistry* 280.42 (2005), pp. 35382–35390.

[300] Nogales, Juan et al. "Unravelling the gallic acid degradation pathway in bacteria: the gal cluster from Pseudomonas putida". In: *Molecular microbiology* 79.2 (2011), pp. 359–374.

[301] Nogales, Juan et al. "Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis". In: *Proceedings of the National Academy of Sciences* 109.7 (2012), pp. 2678–2683.

[302] Nood, E. van et al. "Duodenal infusion of donor feces for recurrent Clostridium difficile". In: *N Engl J Med* 368.5 (2013), pp. 407–15.

[303] Nookaew, Intawat et al. "The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism". In: *Bmc Systems Biology* 2 (2008).

[304] Noor, Elad et al. "An integrated open framework for thermodynamics of reactions that combines accuracy and coverage". In: *Bioinformatics* 28.15 (2012), pp. 2037–2044.

[305] Noor, Elad et al. "Pathway thermodynamics highlights kinetic obstacles in central metabolism". In: *PLoS Comput Biol* 10.2 (2014), e1003483.

[306] Nørager, Sofie et al. "E. coli dihydroorotate dehydrogenase reveals structural and functional distinctions between different classes of dihydroorotate dehydrogenases". In: *Structure* 10.9 (2002), pp. 1211–1223.

[307] Notebaart, Richard A et al. "Correlation between sequence conservation and the genomic context after gene duplication". In: *Nucleic acids research* 33.19 (2005), pp. 6164–6171.

[308] Novichkov, P. S. et al. "RegPrecise 3.0-A resource for genome-scale exploration of transcriptional regulation in bacteria". In: *Bmc Genomics* 14 (2013).

[309] Oberhardt, Matthew A, Palsson, Bernhard Ø, and Papin, Jason A. "Applications of genome-scale metabolic reconstructions". In: *Molecular systems biology* 5.1 (2009), p. 320.

[310] Oberhardt, Matthew A et al. "Genome-scale metabolic network analysis of the opportunistic pathogen Pseudomonas aeruginosa PAO1". In: *Journal of bacteriology* 190.8 (2008), pp. 2790–2803.

[311] Oberhardt, Matthew A et al. "Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis". In: *PLoS Comput Biol* 7.3 (2011), e1001116.

[312] Oh, You-Kwan et al. "Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data". In: *Journal of Biological Chemistry* 282.39 (2007), pp. 28791–28799.

[313] Ohresser, Marc, Matagne, René F, and Loppes, Roland. "Expression of the arylsulphatase reporter gene under the control of the nit1 promoter in Chlamydomonas reinhardtii". In: *Current genetics* 31.3 (1997), pp. 264–271.

[314] Oliveira, Ana Paula, Nielsen, Jens, and Förster, Jochen. "Modeling Lactococcus lactis using a genome-scale flux model". In: *BMC microbiology* 5.1 (2005), p. 1.

[315] Oliveira Dal'Molin, Cristiana Gomes de et al. "AlgaGEM–a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome". In: *BMC genomics* 12.4 (2011), p. 1.

[316] Orth, Jeffrey D and Palsson, Bernhard Ø. "Systematizing the generation of missing metabolic knowledge". In: *Biotechnology and bioengineering* 107.3 (2010), pp. 403–412.

[317] Orth, Jeffrey D and Palsson, BernhardØ. "Gap-filling analysis of the i JO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions". In: *BMC systems biology* 6.1 (2012), p. 1.

[318] Orth, Jeffrey D, Thiele, Ines, and Palsson, Bernhard Ø. "What is flux balance analysis?" In: *Nature biotechnology* 28.3 (2010), pp. 245–248.

[319] Orth, Jeffrey D et al. "A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011". In: *Molecular systems biology* 7.1 (2011), p. 535.

[320] Österlund, Tobias, Nookaew, Intawat, and Nielsen, Jens. "Fifteen years of large scale metabolic modeling of yeast: developments and impacts". In: *Biotechnology advances* 30.5 (2012), pp. 979–988.

[321] Osterman, Andrei. "A hidden metabolic pathway exposed". In: *Proceedings of the National Academy of Sciences* 103.15 (2006), pp. 5637–5638.

[322] Ottman, N. A. "Host immunostimulation and substrate utilization of the gut symbiont Akkermansia muciniphila". PhD thesis. WUR, Wageningen University, 2015.

[323] Overhage, Jörg, Priefert, Horst, and Steinbüchel, Alexander. "Biochemical and genetic analyses of ferulic acid catabolism in Pseudomonas sp. strain HR199". In: *Applied and environmental microbiology* 65.11 (1999), pp. 4837–4847.

[324] O'Brien, EdwardJ, Monk, JonathanM, and Palsson, BernhardO. "Using Genome-scale Models to Predict Biological Capabilities". In: *Cell* 161.5 (2015), pp. 971–987.

[325] Padan, Etana et al. "Alkaline pH homeostasis in bacteria: new insights". In: *Biochimica et biophysica acta (BBA)-biomembranes* 1717.2 (2005), pp. 67–88.

[326] Page, Andrew J et al. "Roary: rapid large-scale prokaryote pan genome analysis". In: *Bioinformatics* 31.22 (2015), pp. 3691–3693.

[327] Palleroni, Norberto J. "Pseudomonas". In: *Bergey's Manual of Systematics of Archaea and Bacteria* (1984).

[328] Parschat, Katja et al. "Xanthine dehydrogenase from Pseudomonas putida 86: specificity, oxidation–reduction potentials of its redox-active centers, and first EPR characterization". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* 1544.1 (2001), pp. 151–165.

[329] Pasek, Sophie, Risler, Jean-Loup, and Brézellec, Pierre. "Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins". In: *Bioinformatics* 22.12 (2006), pp. 1418–1423.

[330] Passell, Howard et al. "Algae biodiesel life cycle assessment using current commercial data". In: *Journal of environmental management* 129 (2013), pp. 103–111.

[331] Pedruzzi, Ivo et al. "HAMAP in 2015: updates to the protein family classification and annotation system". In: *Nucleic acids research* 43.D1 (2015), pp. D1064–D1070.

[332] Petersen, Thomas Nordahl et al. "SignalP 4.0: discriminating signal peptides from transmembrane regions". In: *Nature methods* 8.10 (2011), pp. 785–786.

[333] Peterson, J. et al. "The NIH Human Microbiome Project". In: *Genome Res* 19.12 (2009), pp. 2317–2323.

[334] Pharkya, Priti, Burgard, Anthony P, and Maranas, Costas D. "OptStrain: a computational framework for redesign of microbial production systems". In: *Genome research* 14.11 (2004), pp. 2367–2376.

[335] Pinner, Elhanan et al. "Physiological role of nhaB, a specific Na+/H+ antiporter in Escherichia coli." In: *Journal of Biological Chemistry* 268.3 (1993), pp. 1729–1734.

[336] Placzek, Sandra et al. "BRENDA in 2017: new perspectives and new tools in BRENDA". In: *Nucleic Acids Research* 45.D1 (2017), pp. D380–D388.

[337] Plata, Germán et al. "Reconstruction and flux-balance analysis of the Plasmodium falciparum metabolic network". In: *Molecular systems biology* 6.1 (2010), p. 408.

[338] Poblete-Castro, Ignacio et al. "Industrial biotechnology of Pseudomonas putida and related species". In: *Applied microbiology and biotechnology* 93.6 (2012), pp. 2279–2290.

[339] Prasant, M et al. "Pseudomonas syringae: an overview and its future as a "rain making bacteria"". In: *Int Res J Bio Sci* 4.2 (2015), pp. 70–77.

[340] Prieto, Ana, Canavate, J Pedro, and García-González, Mercedes. "Assessment of carotenoid production by Dunaliella salina in different culture systems and operation regimes". In: *Journal of biotechnology* 151.2 (2011), pp. 180–185.

[341] Puchałka, Jacek et al. "Genome-Scale Reconstruction and Analysis of the Pseudomonas putid a KT2440 Metabolic Network Facilitates Applications in Biotechnology". In: *PLoS Comput Biol* 4.10 (2008), e1000210.

[342] Puddu, A. et al. "Evidence for the Gut Microbiota Short-Chain Fatty Acids as Key Pathophysiological Molecules Improving Diabetes". In: *Mediators of Inflammation* (2014).

[343] Purton, S et al. "Genetic engineering of algal chloroplasts: progress and prospects". In: *Russian journal of plant physiology* 60.4 (2013), pp. 491–499.

[344] Qin, J. J. et al. "A human gut microbial gene catalogue established by metagenomic sequencing". In: *Nature* 464.7285 (2010), 59–U70.

[345] Quevrain, E. et al. "Identification of an anti-inflammatory protein from Faecalibacterium prausnitzii, a commensal bacterium deficient in Crohn's disease". In: *Gut* 65.3 (2016), pp. 415–425.

[346] Quinn, Jeanette M, Kropat, Janette, and Merchant, Sabeeha. "Copper response element and Crr1-dependent Ni2+-responsive promoter for induced, reversible gene expression in Chlamydomonas reinhardtii". In: *Eukaryotic Cell* 2.5 (2003), pp. 995–1002.

[347] Radivojac, Predrag et al. "A large-scale evaluation of computational protein function prediction". In: *Nature methods* 10.3 (2013), pp. 221–227.

[348] Radrich, Karin et al. “Integration of metabolic databases for the reconstruction of genome-scale metabolic networks”. In: *Bmc Systems Biology* 4 (2010).

[349] Rahav-Manor, O et al. “NhaR, a protein homologous to a family of bacterial regulatory proteins (LysR), regulates nhaA, the sodium proton antiporter gene in Escherichia coli.” In: *Journal of Biological Chemistry* 267.15 (1992), pp. 10433–10438.

[350] Rajilic-Stojanovic, M. and Vos, W. M. de. “The first 1000 cultured species of the human gastrointestinal microbiota”. In: *Fems Microbiology Reviews* 38.5 (2014), pp. 996–1047.

[351] Ramazzina, Ileana et al. “Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes”. In: *Nature chemical biology* 2.3 (2006), pp. 144–148.

[352] Ramazzina, Ileana et al. “Logical identification of an allantoinase analog (puuE) recruited from polysaccharide deacetylases”. In: *Journal of Biological Chemistry* 283.34 (2008), pp. 23295–23304.

[353] Ramos-González, María Isabel and Molin, Søren. “Cloning, sequencing, and phenotypic characterization of the rpoS gene from Pseudomonas putida KT2440”. In: *Journal of bacteriology* 180.13 (1998), pp. 3421–3431.

[354] Ramsey, D. M. and Wozniak, D. J. “Understanding the control of Pseudomonas aeruginosa alginate synthesis and the prospects for management of chronic infections in cystic fibrosis”. In: *Mol Microbiol* 56.2 (2005), pp. 309–22.

[355] Ramundo, Silvia et al. “Repression of essential chloroplast genes reveals new signaling pathways and regulatory feedback loops in Chlamydomonas”. In: *The Plant Cell* 25.1 (2013), pp. 167–186.

[356] Rasala, Beth A et al. “Robust expression and secretion of Xylanase1 in Chlamydomonas reinhardtii by fusion to a selection gene and processing with the FMDV 2A peptide”. In: *PloS one* 7.8 (2012), e43349.

[357] Rath, C. M. et al. “Molecular analysis of model gut microbiotas by imaging mass spectrometry and nanodesorption electrospray ionization reveals dietary metabolite transformations”. In: *Anal Chem* 84.21 (2012), pp. 9259–67.

[358] Reed, Jennifer L et al. “Towards multidimensional genome annotation”. In: *Nature Reviews Genetics* 7.2 (2006), pp. 130–141.

[359] Regenhardt, D et al. “Pedigree and taxonomic credentials of Pseudomonas putida strain KT2440”. In: *Environmental microbiology* 4.12 (2002), pp. 912–915.

[360] Rentzsch, Robert and Orengo, Christine A. “Protein function prediction–the power of multiplicity”. In: *Trends in biotechnology* 27.4 (2009), pp. 210–219.

[361] Reynes, Jean-Paul et al. “Escherichia coli thymidylate kinase: molecular cloning, nucleotide sequence, and genetic organization of the corresponding tmk locus.” In: *Journal of bacteriology* 178.10 (1996), pp. 2804–2812.

[362] Ridlon, J. M. et al. “Bile acids and the gut microbiome”. In: *Current Opinion in Gastroenterology* 30.3 (2014), pp. 332–338.

[363] Rienksma, Rienk A et al. “Systems-level modeling of mycobacterial metabolism for the identification of new (multi-) drug targets”. In: *Seminars in immunology*. Vol. 26. 6. Elsevier. 2014, pp. 610–622.

[364] Ritari, J. et al. “Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database”. In: *Bmc Genomics* 16 (2015).

[365] Rivas, Mariella O., Vargas, Pedro, and Riquelme, Carlos E. "Interactions of Botryococcus braunii Cultures with Bacterial Biofilms". In: *Microbial Ecology* 60.3 (2010), pp. 628–635.

[366] Rkenes, Torunn P, Lamark, Trond, and Strøm, Arne R. "DNA-binding properties of the BetI repressor protein of Escherichia coli: the inducer choline stimulates BetI-DNA complex formation." In: *Journal of bacteriology* 178.6 (1996), pp. 1663–1670.

[367] Rodgers-Melnick, Eli, Culp, Mark, and DiFazio, Stephen P. "Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS". In: *BMC genomics* 14.1 (2013), p. 1.

[368] Rosenberg, Julian N et al. "Comparative analyses of three Chlorella species in response to light and sugar reveal distinctive lipid accumulation patterns in the microalga C. sorokiniana". In: *PloS one* 9.4 (2014), e92460.

[369] Ruhal, Rohit, Kataria, Rashmi, and Choudhury, Bijan. "Trends in bacterial trehalose metabolism and significant nodes of metabolic pathway in the direction of trehalose accumulation". In: *Microbial biotechnology* 6.5 (2013), pp. 493–502.

[370] Salimi, F., Zhuang, K., and Mahadevan, R. "Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing". In: *Biotechnology Journal* 5.7 (2010), pp. 726–738.

[371] Santos, Filipe Branco dos, Vos, Willem M de, and Teusink, Bas. "Towards metagenome-scale models for industrial applications—the case of Lactic Acid Bacteria". In: *Current opinion in biotechnology* 24.2 (2013), pp. 200–206.

[372] Santos, VAP Martins dos et al. "Insights into the genomic basis of niche specificity of Pseudomonas putida KT2440". In: *Environmental Microbiology* 6.12 (2004), pp. 1264–1286.

[373] Saparov, Sapar M, Antonenko, Yuri N, and Pohl, Peter. "A new model of weak acid permeation through membranes revisited: does Overton still rule?" In: *Biophysical journal* 90.11 (2006), pp. L86–L88.

[374] Sassetti, Christopher M., Boyd, Dana H., and Rubin, Eric J. "Genes required for mycobacterial growth defined by high density mutagenesis". In: *Molecular Microbiology* 48.1 (2003), pp. 77–84.

[375] Saulnier, Delphine M et al. "Exploring metabolic pathway reconstruction and genome-wide expression profiling in Lactobacillus reuteri to define functional probiotic features". In: *PloS one* 6.4 (2011), e18783.

[376] Schellenberger, Jan et al. "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0". In: *Nature Protocols* 6.9 (2011), pp. 1290–1307.

[377] Schmitz, Simone et al. "Engineering mediator-based electroactivity in the obligate aerobic bacterium Pseudomonas putida KT2440". In: *Frontiers in microbiology* 6 (2015), p. 284.

[378] Schnackerz, Klaus D and Dobritzsch, Doreen. "Amidohydrolases of the reductive pyrimidine catabolic pathway: purification, characterization, structure, reaction mechanisms and enzyme deficiency". In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1784.3 (2008), pp. 431–444.

[379] Schuetz, Robert, Kuepfer, Lars, and Sauer, Uwe. "Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli". In: *Molecular systems biology* 3.1 (2007), p. 119.

[380] Schwander, Thomas et al. "A synthetic pathway for the fixation of carbon dioxide in vitro". In: *Science* 354.6314 (2016), pp. 900–904.

[381] Scott, Stuart A et al. "Biodiesel from algae: challenges and prospects". In: *Current opinion in biotechnology* 21.3 (2010), pp. 277–286.

[382] Scovill, William H, Schreier, Harold J, and Bayles, Kenneth W. "Identification and characterization of the pckA gene from Staphylococcus aureus." In: *Journal of bacteriology* 178.11 (1996), pp. 3362–3364.

[383] Seemann, Torsten. "Prokka: rapid prokaryotic genome annotation". In: *Bioinformatics* (2014), btu153.

[384] Shoaie, S. et al. "Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome". In: *Cell Metabolism* 22.2 (2015), pp. 320–331.

[385] Sigurdsson, Martin I et al. "A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1". In: *BMC systems biology* 4.1 (2010), p. 1.

[386] Silby, Mark W et al. "Pseudomonas genomes: diverse and adaptable". In: *FEMS microbiology reviews* 35.4 (2011), pp. 652–680.

[387] Sjöberg, Britt-Marie and Torrents, Eduard. "Shift in ribonucleotide reductase gene expression in Pseudomonas aeruginosa during infection". In: *Infection and immunity* 79.7 (2011), pp. 2663–2669.

[388] Smith, Adam Alexander Thil et al. "The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes". In: *PLoS Comput Biol* 8.5 (2012), e1002540.

[389] Smith, Linda T et al. "Osmotic control of glycine betaine biosynthesis and degradation in Rhizobium meliloti." In: *Journal of Bacteriology* 170.7 (1988), pp. 3142–3149.

[390] Smith, P. M. et al. "The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis". In: *Science* 341.6145 (2013), pp. 569–73.

[391] Smith, Temple F and Waterman, Michael S. "Identification of common molecular subsequences". In: *Journal of molecular biology* 147.1 (1981), pp. 195–197.

[392] Snipen, Lars, Almøy, Trygve, and Ussery, David W. "Microbial comparative pan-genomics using binomial mixture models". In: *BMC genomics* 10.1 (2009), p. 385.

[393] Snipen, Lars and Liland, Kristian Hovde. "Micropan: An R-package for microbial pan-genomics". In: *BMC bioinformatics* 16.1 (2015), p. 1.

[394] Snipen, Lars-Gustav and Ussery, David W. "A domain sequence approach to pangenomics: applications to Escherichia coli". In: *F1000Research* 1 (2013).

[395] Sohn, Seung Bum et al. "In silico genome-scale metabolic analysis of Pseudomonas putida KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival". In: *Biotechnology journal* 5.7 (2010), pp. 739–750.

[396] Stancik, Lauren M et al. "pH-dependent expression of periplasmic proteins and amino acid catabolism in Escherichia coli". In: *Journal of bacteriology* 184.15 (2002), pp. 4246–4258.

[397] Steeb, B. et al. "Parallel exploitation of diverse host nutrients enhances Salmonella virulence". In: *PLoS Pathog* 9.4 (2013), e1003301.

[398] Steen, Annika et al. "Construction and characterization of nitrate and nitrite respiring Pseudomonas putida KT2440 strains for anoxic biotechnical applications". In: *Journal of biotechnology* 163.2 (2013), pp. 155–165.

[399] Stobbe, Miranda D. et al. "Critical assessment of human metabolic pathway databases: a stepping stone for future integration". In: *Bmc Systems Biology* 5 (2011).

[400] Stobbe, Miranda D. et al. "Consensus and conflict cards for metabolic pathway databases". In: *Bmc Systems Biology* 7 (2013).

[401] Stolyar, Sergey et al. "Metabolic modeling of a mutualistic microbial community". In: *Molecular systems biology* 3.1 (2007), p. 92.

[402] Sudom, Athena et al. "Mechanisms of Activation of Phosphoenolpyruvate Carboxykinase from Escherichia coli by Ca2+ and of Desensitization by Trypsin". In: *Journal of bacteriology* 185.14 (2003), pp. 4233–4242.

[403] Sun, J. et al. "Constraint-based modeling analysis of the metabolism of two Pelobacter species". In: *Bmc Systems Biology* 4 (2010).

[404] Suzuki, Ryota and Shimodaira, Hidetoshi. "Pvclust: an R package for assessing the uncertainty in hierarchical clustering". In: *Bioinformatics* 22.12 (2006), pp. 1540–1542.

[405] Talukdar, Jayanta, Kalita, Mohan Chandra, and Goswami, Bhabesh Chandra. "Characterization of the biofuel potential of a newly isolated strain of the microalga Botryococcus braunii Kützing from Assam, India". In: *Bioresource technology* 149 (2013), pp. 268–275.

[406] Tan, Jie et al. "ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions". In: *mSystems* 1.1 (2016), e00025–15.

[407] Taniguchi, Yuichi et al. "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells". In: *Science* 329.5991 (2010), pp. 533–538.

[408] Tardif, Marianne et al. "PredAlgo: a new subcellular localization prediction tool dedicated to green algae". In: *Molecular biology and evolution* (2012), mss178.

[409] Tatusov, Roman L, Koonin, Eugene V, and Lipman, David J. "A genomic perspective on protein families". In: *Science* 278.5338 (1997), pp. 631–637.

[410] Team, R Core. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* 2013.

[411] Tepper, Naama and Shlomi, Tomer. "Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways". In: *Bioinformatics* 26.4 (2010), pp. 536–543.

[412] Terashima, Mia, Specht, Michael, and Hippler, Michael. "The chloroplast proteome: a survey from the Chlamydomonas reinhardtii perspective with a focus on distinctive features". In: *Current genetics* 57.3 (2011), pp. 151–168.

[413] Tettelin, Hervé et al. "Comparative genomics: the bacterial pan-genome". In: *Current opinion in microbiology* 11.5 (2008), pp. 472–477.

[414] Teusink, B. et al. "In silico reconstruction of the metabolic pathways of Lactobacillus plantarum: Comparing predictions of nutrient requirements with those

from growth experiments". In: *Applied and Environmental Microbiology* 71.11 (2005), pp. 7253–7262.

[415] Teusink, Bas and Smid, Eddy J. "Modelling strategies for the industrial exploitation of lactic acid bacteria". In: *Nature Reviews Microbiology* 4.1 (2006), pp. 46–56.

[416] Teusink, Bas et al. "Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model". In: *Journal of Biological Chemistry* 281.52 (2006), pp. 40041–40048.

[417] Teusink, Bas et al. "Understanding the adaptive growth strategy of Lactobacillus plantarum by in silico optimisation". In: *PLoS Comput Biol* 5.6 (2009), e1000410.

[418] Thiele, Ines and Palsson, Bernhard O. "A protocol for generating a high-quality genome-scale metabolic reconstruction". In: *Nature Protocols* 5.1 (2010), pp. 93–121.

[419] Thiele, Ines, Vlassis, Nikos, and Fleming, Ronan M. T. "FASTGAPFILL: efficient gap filling in metabolic networks". In: *Bioinformatics* 30.17 (2014), pp. 2529–2531.

[420] Thiele, Ines et al. "A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2". In: *Bmc Systems Biology* 5 (2011).

[421] Thiele, Ines et al. "A community-driven global reconstruction of human metabolism". In: *Nature Biotechnology* 31.5 (2013), pp. 419–+.

[422] Tikh, Ilya and Schmidt-Dannert, Claudia. "Towards Engineered Light Energy Conversion in Nonphotosynthetic Microorganisms". In: *Synthetic Biology: Tools and Applications* (2013).

[423] Timinskas, Kęstutis et al. "Comprehensive analysis of DNA polymerase III α subunits and their homologs in bacterial genomes". In: *Nucleic acids research* 42.3 (2014), pp. 1393–1413.

[424] Timmis, Kenneth N. "Pseudomonas putida: a cosmopolitan opportunist par excellence". In: *Environmental microbiology* 4.12 (2002), pp. 779–781.

[425] Tomar, Namrata and De, Rajat K. "Comparing methods for metabolic network analysis and an application to metabolic engineering". In: *Gene* 521.1 (2013), pp. 1–14.

[426] Tormo, Antonio, Almirón, Marta, and Kolter, Roberto. "surA, an Escherichia coli gene essential for survival in stationary phase." In: *Journal of bacteriology* 172.8 (1990), pp. 4339–4347.

[427] Touchon, Marie et al. "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths". In: *PLoS genet* 5.1 (2009), e1000344.

[428] Trentacoste, Emily M et al. "Metabolic engineering of lipid catabolism increases microalgal lipid accumulation without compromising growth". In: *Proceedings of the National Academy of Sciences* 110.49 (2013), pp. 19748–19753.

[429] Tripp, H James et al. "Toward a standard in structural genome annotation for prokaryotes". In: *Standards in genomic sciences* 10.1 (2015), p. 1.

[430] Tzamali, E. et al. "A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities". In: *Bmc Systems Biology* 5 (2011).

[431] Udaondo, Zulema et al. "Analysis of the core genome and pangenome of Pseudomonas putida". In: *Environmental microbiology* (2015).

[432] Vallenet, David et al. "MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data". In: *Nucleic acids research* (2012), gks1194.

[433] Vander Wauven, Corinne et al. "Pseudomonas aeruginosa mutants affected in anaerobic growth on arginine: evidence for a four-gene cluster encoding the arginine deiminase pathway." In: *Journal of bacteriology* 160.3 (1984), pp. 928–934.

[434] Velasco-García, Roberto et al. "Betaine aldehyde dehydrogenase from Pseudomonas aeruginosa: cloning, over-expression in Escherichia coli, and regulation by choline and salt". In: *Archives of microbiology* 185.1 (2006), pp. 14–22.

[435] Ventura, M. et al. "Genome-scale analyses of health-promoting bacteria: probiogenomics". In: *Nature Reviews Microbiology* 7.1 (2009), 61–U77.

[436] Venturi, Vittorio. "Control of rpoS transcription in Escherichia coli and Pseudomonas: why so different?" In: *Molecular microbiology* 49.1 (2003), pp. 1–9.

[437] Volpers, Michael et al. "Integrated in silico analysis of pathway designs for synthetic photo-electro-autotrophy". In: *PloS one* 11.6 (2016), e0157851.

[438] Wargo, Matthew J. "Homeostasis and catabolism of choline and glycine betaine: lessons from Pseudomonas aeruginosa". In: *Applied and environmental microbiology* 79.7 (2013), pp. 2112–2120.

[439] Wargo, Matthew J and Hogan, Deborah A. "Identification of genes required for Pseudomonas aeruginosa carnitine catabolism". In: *Microbiology* 155.7 (2009), pp. 2411–2419.

[440] Wargo, Matthew J, Szwergold, Benjamin S, and Hogan, Deborah A. "Identification of two gene clusters and a transcriptional regulator required for Pseudomonas aeruginosa glycine betaine catabolism". In: *Journal of bacteriology* 190.8 (2008), pp. 2690–2699.

[441] Wegkamp, A et al. "Development of a minimal growth medium for Lactobacillus plantarum". In: *Letters in applied microbiology* 50.1 (2010), pp. 57–64.

[442] Weiss, Taylor L et al. "Genome size and phylogenetic analysis of the A and L races of Botryococcus braunii". In: *Journal of applied phycology* 23.5 (2011), pp. 833–839.

[443] Weiss, Taylor L et al. "Colony organization in the green alga Botryococcus braunii (Race B) is specified by a complex extracellular matrix". In: *Eukaryotic cell* 11.12 (2012), pp. 1424–1440.

[444] Weitzel, Michael et al. "13CFLUX2—high-performance software suite for 13C-metabolic flux analysis". In: *Bioinformatics* 29.1 (2013), pp. 143–145.

[445] West, Thomas P. "Pyrimidine base catabolism in Pseudomonas putida biotype B". In: *Antonie van Leeuwenhoek* 80.2 (2001), pp. 163–167.

[446] Whitney, Spencer M, Houtz, Robert L, and Alonso, Hernan. "Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco". In: *Plant Physiology* 155.1 (2011), pp. 27–35.

[447] Wiechert, Wolfgang. "C-13 metabolic flux analysis". In: *Metabolic Engineering* 3.3 (2001), pp. 195–206.

[448] Wienkoop, S. tefanie et al. "Targeted proteomics for Chlamydomonas reinhardtii combined with rapid subcellular protein fractionation, metabolomics and metabolic flux analyses". In: *Mol. Biosystems* 6.6 (2010), pp. 1018–1031.

[449] Wienkoop, Stefanie et al. "ProMEX–a mass spectral reference database for plant proteomics". In: *Frontiers in plant science* 3 (2012), p. 125.

[450] Wijffels, Rene H, Barbosa, Maria J, and Eppink, Michel HM. "Microalgae for the production of bulk chemicals and biofuels". In: *Biofuels, Bioproducts and Biorefining* 4.3 (2010), pp. 287–295.

[451] Winsor, Geoffrey L et al. "Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes". In: *Nucleic acids research* (2010), gkq869.

[452] Wong, Andrew and Shatkay, Hagit. "Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge". In: *BMC bioinformatics* 14.3 (2013), p. 1.

[453] Wu, Xiao et al. "Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida". In: *FEMS microbiology reviews* 35.2 (2011), pp. 299–323.

[454] Xavier, Joana C, Patil, Kiran Raosaheb, and Rocha, Isabel. "Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes". In: *Metabolic Engineering* (2016).

[455] Xiong, Wei et al. "CO2-fixing one-carbon metabolism in a cellulose-degrading bacterium Clostridium thermocellum". In: *Proceedings of the National Academy of Sciences* 113.46 (2016), pp. 13180–13185.

[456] Xu, Chuan et al. "Genome-scale metabolic model in guiding metabolic engineering of microbial improvement". In: *Applied microbiology and biotechnology* 97.2 (2013), pp. 519–539.

[457] Yang, Laurence et al. "Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data". In: *Proceedings of the National Academy of Sciences* 112.34 (2015), pp. 10810–10815.

[458] Yang, Song, Doolittle, Russell F, and Bourne, Philip E. "Phylogeny determined by protein domain content". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.2 (2005), pp. 373–378.

[459] Yang, Xiaowen et al. "Analysis of pan-genome to identify the core genes and essential genes of Brucella spp." In: *Molecular Genetics and Genomics* 291.2 (2016), pp. 905–912.

[460] Ye, Lumeng et al. "Draft genome sequence analysis of a Pseudomonas putida W15Oct28 Strain with antagonistic activity to Gram-positive and Pseudomonas sp. pathogens". In: *PloS one* 9.11 (2014), e110038.

[461] Yim, Harry et al. "Metabolic engineering of Escherichia coli for direct production of 1, 4-butanediol". In: *Nature chemical biology* 7.7 (2011), pp. 445–452.

[462] Yizhak, Keren et al. "Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model". In: *Bioinformatics* 26.12 (2010), pp. i255–i260.

[463] Yuan, Qianqian et al. "Pathway-Consensus Approach to Metabolic Network Reconstruction for Pseudomonas putida KT2440 by Systematic Comparison of Published Models". In: *PloS one* 12.1 (2017), e0169437.

[464] Zarecki, Raphy et al. "A novel nutritional predictor links microbial fastidiousness with lowered ubiquity, growth rate, and cooperativeness". In: *PLOS Comput Biol* 10.7 (2014), e1003726.

[465] Zelezniak, Aleksej et al. "Metabolic dependencies drive species co-occurrence in diverse microbial communities". In: *Proceedings of the National Academy of Sciences* 112.20 (2015), pp. 6449–6454.

[466] Zhang, Ying et al. "Three-dimensional structural view of the central metabolic network of Thermotoga maritima". In: *Science* 325.5947 (2009), pp. 1544–1549.

[467] Zhou, L. and Foster, J. A. "Psychobiotics and the gut-brain axis: in the pursuit of happiness". In: *Neuropsychiatr Dis Treat* 11 (2015), pp. 715–23.

[468] Ziegler, Christine, Bremer, Erhard, and Krämer, Reinhard. "The BCCT family of carriers: from physiology to crystal structure". In: *Molecular microbiology* 78.1 (2010), pp. 13–34.

[469] Zoetendal, E. G. and Vos, W. M. de. "Effect of diet on the intestinal microbiota and its activity". In: *Current Opinion in Gastroenterology* 30.2 (2014), pp. 189–195.

[470] Zomorrodi, Ali R, Islam, Mohammad Mazharul, and Maranas, Costas D. "d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities". In: *ACS synthetic biology* 3.4 (2014), pp. 247–257.

[471] Zomorrodi, Ali R and Maranas, Costas D. "OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities". In: *PLoS Comput Biol* 8.2 (2012), e1002363.

[472] Zomorrodi, Ali R et al. "Mathematical optimization applications in metabolic networks". In: *Metabolic engineering* 14.6 (2012), pp. 672–686.

[473] Zur, Hadas, Ruppin, Eytan, and Shlomi, Tomer. "iMAT: an integrative metabolic analysis tool". In: *Bioinformatics* 26.24 (2010), pp. 3140–3142.

List of publications

Matthijn C. Hesselman*, Dorett I. Odoni*, Brendan M. Ryback*, Suzette de Groot, **Ruben G. A. van Heck**, Jaap Keijsers, Pim Kolkman, David Nieuwenhuijse, Youri M. van Nuland, Erik Sebus,Rob Spee, Hugo de Vries, Marten T. Wapenaar, Colin J. Ingham, Karin Schroën, Vítor A. P. Martins dos Santos, Sebastiaan K. Spaans, Floor Hugenholtz, and Mark W. J. van Passel. "A Multi-Platform Flow Device for Microbial (Co-) Cultivation and Microscopic Analysis". In: *PloS one* 7(5) 2012.

Brendan M. Ryback*, Dorett I. Odoni*, **Ruben G. A. van Heck**, Youri van Nuland, Matthijn C. Hesselman, Vítor A. P. Martins dos Santos, Mark W. J. van Passel, and Floor Hugenholtz. "Design and analysis of a tunable synchronized oscillator". In: Journal of biological engineering 7(1) 2013.

Emiel B. J. ten Buren, Michiel A. P. Karrenbelt, Marit Lingemann, Shreyans Chordia, Ying Deng, JingJing Hu, Johanna M. Verest, Vincen Wu, Teresita J. Bello Gonzalez, **Ruben G. A. van Heck**, Dorett I. Odoni, Tom Schonewille, Laura van der Straat, Leo H. de Graaff, and Mark W. J. van Passel. "Toolkit for Visualization of the Cellular Structure and Organelles in *Aspergillus niger*". In: ACS Synth. Biol., 3(12) 2014.

Maarten J.M.F. Reijnders, **Ruben G. A. van Heck**, Carolyn M. C. Lam, Mark A. Scaife, Vitor A. P. Martins dos Santos, Alison G. Smith, and Peter J. Schaap. "Green genes: bioinformatics and systems-biology innovations drive algal biotechnology". In: Trends in Biotechnology, 32(12) 2014.

Ruben G. A. van Heck*, Mathias Ganter*, Vítor A. P. Martins dos Santos, and Joerg Stelling. "Efficient Reconstruction of Predictive Consensus Metabolic Network Models". In: PLoS Comput Biol 12(8) 2016.

*Equal contributions

Eugeni Belda*, **Ruben G. A. van Heck***, Maria Jose Lopez-Sanchez, Stephane Cruveiller, Valerie Barbe, Claire Fraser, Hans-Peter Klenk, Jorn Petersen, Anne Morgat, Pablo I. Nikel, David Vallenet, Zoe Rouy, Agnieszka Sekowska, Vitor A. P. Martins dos Santos, Victor de Lorenzo, Antoine Danchin, and Claudine Medigue. "The revisited genome of *Pseudomonas putida* KT2440 enlightens its value as a robust metabolic chassis". In: Environmental Microbiology 18(10) 2016.

Jasper J. Koehorst*, Jesse C. J. van Dam*, **Ruben G. A. van Heck**, Edoardo Saccenti, Vitor A. P. Martins dos Santos, Maria Suarez-Diez, and Peter J. Schaap. "Comparison of 432 *Pseudomonas* strains through integration of genomic, functional, metabolic and expression data". In: Scientific Reports 6 2016.

Dorett I. Odoni*, Juan A. Tamayo-Ramos*, Jasper Sloothaak, **Ruben G. A. van Heck**, Vitor A. P. Martins dos Santos, Leo H. de Graaff, Maria Suarez-Diez, and Peter J. Schaap. "Comparative proteomics of *Rhizopus delemar* ATCC 20344 unravels the role of amino acid catabolism in fumarate accumulation". In: PeerJ Preprints 2017.

Kees C. H. van der Ark*, **Ruben G. A. van Heck***, Vitor A. P. Martins dos Santos, Clara Belzer, and Willem M. de Vos. "More than just a gut feeling: Constraint-based genome-scale metabolic models for predicting functions of human intestinal microbes". Under review at Microbiome.

Ruben G. A. van Heck*, Henri van Kruistum*, Maria Suarez-Diez, and Nico Claassens. "Determining organism-specific synthetic pathways for CO2 fixation via metabolic modelling". In preparation

Ruben G. A. van Heck*, Stefano Donati*, Linde Kampers, Pablo I. Nikel, Maria Suarez-Diez, Vitor A. P. Martins dos Santos, Peter J. Schaap, Edoardo Saccenti. "*In silico* design of anaerobic *Pseudomonas putida*". In preparation

*Equal contributions

Overview of completed training activities

Discipline specific	**Year**
Linear and integer programming	2013
IPOP networks tutorial	2013
Synthetic Biology 6.0	2013
iGEM 2013 regional jamboree (judge)	2013
SB@NL 2013	2013
Gordon conference Synthetic Biology	2013
COBRA conference 2014	2014
Multiscale, Cell-based Modelling in Biological Development and Cancer	2014
iGEM 2014 world jamboree (supervisor)	2014
Integrative Cell Models	2015
EmPowerPutida Kick-off meeting	2015
EmPowerPutida 1st General Assembly	2015
EmPowerPutida 2nd General Assembly	2016
EmPowerPutida 3rd General Assembly	2016
BioSB 2016	2016
General	
VLAG PhD week	2013
Reviewing a scientific paper	2014
Scientific writing	2015
Teaching and Supervising Thesis Students	2015
Introduction to FAIR data management	2016
Responsible Research and Innovation	2016
Soft skills training with role play	2016
IPLAW1x Intellectual Property Law and Policy: Part 1	2016
IPLAW1x Intellectual Property Law and Policy: Part 2	2016
FAIR data management workshop	2017
Optional	
Preparation of research proposal	2013
Weekly group meetings	2013-2017
Seminar series	2013-2017
PhD trip 2015 organization	2014-2015
PhD trip 2015	2015
PhD trip 2017	2017

Acknowledgements

That's it. Done. This booklet marks the finale of four memorable years at the Laboratory of Systems and Synthetic Biology. Here, I have had the opportunity to work on a wide variety of interesting topics in a relaxed and diverse working environment. My work and work drive had - as is to be expected - their ups and downs during my PhD, but I have enjoyed consistently positive interactions with colleagues, friends, and family. I would like to take this opportunity to thank you all - especially those that got here after reading all preceding chapters - for greatly improving my quality of life these past few years.

First and foremost: **Maria**, thank you for everything. Originally you were not officially involved in my PhD, but you were always willing to make time and discuss how to best approach any presented challenges. About two years ago you *de facto* became my daily supervisor as I struggled to manage my ongoing projects. I have immensely benefited from your help these past years, as have many others in SSB. SSB would be a considerably less productive environment without your guidance. I was delighted to hear earlier this year that you were finally promoted to assistant professor; a promotion long overdue! **Maria**, you're great, keep doing what you do.

Vitor, thank you for the opportunity to do my PhD in your group. Even before I started the PhD, you enabled me to attend international conferences, workshops and meetings. These events provided me with a good overview of the *status quo* of the field, international collaborators, and several ideas. The development of these ideas into full-fledged projects was immensely enjoyable and rewarding, and would not have been possible without the freedom and support you grant us in SSB. I hope that SSB further develops according to your vision of tightly collaborating computational and experimental biologists.

Peter, thank you for your advice, support, and the interesting discussions we had over the years. I benefited from the fruitful collaborations you fostered within your Computational Systems Biology group, and I am grateful for your help in shaping my student projects.

Of all fellow PhD students of SSB and Microbiology (MIB) I worked together with, there is but one consideration for the first mention. He is always

first: First to officially register for a printed copy of this thesis, first in presentation skills at the VLAG PhD-week, first to call himself a bioinfomagician in all seriousness, first to be mentioned when describing interesting co-workers, and the first author on our first shared paper. **Maarten**, all jokes aside, it was great to work together with someone so undeniably efficient, confident and positive. At the time of this writing, there is about half a year left before your scheduled first Nobel prize, which will undoubtedly finally propel you to be the first **Maarten Reijnders** found in a Google search.

Jasper, **Jesse**, you were my go-to people - conveniently located in the same office - for any issues and questions I had with regards to annotation, databases, algorithms, and various bioinformatics methods; probably for any computer problems, really. Besides this, I also enjoyed the various projects we initiated together and 'paused' for the time being: **Jesse**, it was fun discovering together how big the problem with inconsistent metabolite naming really is; a remaining issue. **Jasper**, we somehow convinced each other it was a good idea for the two of us to attempt lab work together after years of not touching a pipette, only saved from our mistake when **Merlijn** - Thank you! - kindly offered to take over. Finally, I was happy to contribute to your shared project on the comparison of 432 *Pseudomonas* genomes.

Nico, it was really convenient to have you around whenever I was in need of a great example; it seems that somehow you can just do it all. It was a pleasure to be involved in your projects as a modeler as you have a great grasp on the possibilities and limitations of *in silico* methods. Near the end of our PhDs we designed and initiated a highly ambitious MSc thesis project on the design of synthetic CO_2 fixation pathways. **Henri**, our mutual student, far outperformed our expectations, and I found myself trying to push this project into my thesis less than two weeks before submission. **Nico**, I truly appreciate your help in finalizing that chapter.

Rita, you came into my office about a year ago to ask if there was anything I would like to have done in the lab. I really appreciated that offer and we quickly found ourselves putting together a project on metabolic engineering in *P. putida*. This project has been running well since, in good part also thanks to our mutual student **Christos**. Thank you for the opportunity!

Edoardo, I came to you several times these past years with questions or project proposals. Thank you for your readiness to help out and discuss immediately in each individual case. One of these discussions resulted in the start of the 'anaerobic *P. putida*' project, which has definitely been one of the highlights of my PhD. For this project we co-supervised **Stefano**, who rationally designed anaerobic *P. putida* strains based on *in silico* analyses, and laid the groundwork for the experimental evaluation.

Linde, I am extremely glad that you decided to further pursue the anaerobic *P. putida* project experimentally. It has been a delight to work with you

these past months, as well as to share the office. Somehow, you have convinced most of our other colleagues to bring candy - especially chocolate - to our office, which makes for a great office atmosphere. This is best exemplified by it regularly taking you over an hour to go from leaving to actually walking out the door due to extensive 'work' discussions.

Nhung, thank you for relieving me of all my unfinished work and granting me the freedom to enjoy life knowing its left in good hands. Your critical reading of several thesis chapters definitely helped me improve them, and I appreciate you thinking along on how to best end the PhD. For better or worse, I did not follow all your advice... I have also really enjoyed our many work discussions, especially those that spontaneously started in the office as soon as I put down other work for just one second.

Stamatios, at the start of our PhD we had some grand plans bringing together your experimental work and my computational work on *P. putida*. In the end, we must acknowledge these plans were somewhat overly ambitious as we both got caught up in the preparation phase. Regardless, we regularly had interesting discussions as you always seem to bring new great ideas to the table!

Some of my favorite work in the PhD has been directly supervising and working together with excellent BSs and MSc thesis students. **Michiel**, **Walter**, **Stefano**, **Henri**, and **Christos**, it was great to work with you!

The start of my journey into systems and synthetic biology was in 2011 when I joined the Wageningen UR team for the international Genetically Engineered Machine (iGEM) competition. **Pim**, thanks for convincing me to go to that first meeting. **Floor**, thanks for convincing me to stay on afterwards and for your great efforts and supervision together with **Mark**. **Dorett**, **Brendan**, **Matthijn**, **David**, **Rob**, **Hugo**, **Youri**, **Erik**, **Suzette**, **Martijn**, **Shi**, **Jaap**, **Mariana**, **Pim**, **Floor**, and **Mark**, thanks for that amazing experience; there was no other time during my studies, nor later on, where I learned so much in so little time.

Rienk, you first introduced me to metabolic modeling as part of a MSc course, and later supervised my MSc thesis together with **Maria**. I learned about metabolic modeling and the associated challenges from you, and I'm really grateful for all your help in finalizing that thesis.

Dorett, **Mark**, thanks for inviting me to join the iGEM supervision team in 2013. It was great experiencing the other side of iGEM together with the two of you. This was made easy by the amazing team we had that year: **Michiel**, **Emiel**, **Marjan**, **Vincen** (aka Zeus), **Shreyans**, **Marit**, **Jingjing**, and **Danny**. It was a real pleasure to be a part of this team!

Kees, together we pushed for the continuation of iGEM in Wageningen in 2014, and I was very glad when **Christian** decided to take charge of the whole operation. The three of us supervised iGEM that year together with

Rob, **Nico**, **Marnix** and **Stamatios**; another supervision team I thoroughly enjoyed being a part of. The iGEM students - **Wen**, **Rik**, **Jeremy**, **Walter**, **Marléne**, **Tjaša**, **Teresa**, **Max**, **Michiel**, **Bob**, **Miquel**, and **Kevin** - undeniably did a great job as they took second place in the overgraduate division! Thanks to you all for another great iGEM experience!

Another team that I enjoyed being a part of was the 2015 PhD-trip organization committee. **Maarten**, **Alex**, **Kees**, **Yue**, **Nico**, **Jasper**, and **Anna**, we spent a tremendous amount of time together to organize the trip to California, but it was always relaxed and fun. The trip itself was a blast with a solid scientific program and only a few minor incidents...

The PhD-trip is all about networking and fostering collaborations abroad, but also unveils the untapped potential for shared projects back in Wageningen. **Kees**, you identified a possible collaboration between us during the 2015 PhD-trip and quickly followed up on it. Thank you for that initiative and for involving me in one of your ongoing projects!

The next PhD-trip, in 2017, was again a great success with an interesting and diverse scientific program. **Wen**, **Erika**, **Prarthana**, **Hugo**, **Jeroen**, **Joyshree**, **Alex**, **Ioannis**, and **Johanna**, thank you for all your efforts in organizing this trip!

Although work often seems all-important during the PhD, it is the off-time spent with colleagues and good friends that form the fondest memories. **Dorett**, you are omnipresent in those. Thank you for all the good times we've shared these past years and for all your support. Working together, living together, traveling together, it has all been easy and consistently great. It was clear that you'd be one of my two paranymphs.

Similarly, there was no doubt who would be my other paranymph. **Wen**, its been wonderful to have you as a friend and colleague. I thoroughly enjoyed our climbing sessions, where you pushed me to my limit and regularly flew up when I came crashing down; thank you for saving and enriching my life. You really lighten up everyone around you with your positivism and radiant enthusiasm. Don't ever change.

Niru, it was great sharing an office with you both at the Dreijen and later on in the Helix. Those 2-3 days a week that you worked here were definitely more enjoyable than those where you 'worked elsewhere'. To be fair, coming to the office 2-3 days a week is quite impressive considering that you live in Utrecht and take your bike to work!

Bart, thank you for the good times in the office together at the Dreijen. You made sure the office was never quiet for too long, in good part to not a day going by without at least ten people coming to you for help. Also, thank you for your help with coding, with the servers, and for making sure that the rest of us do not accidentally lose all our work.

Niels, it was good fun attending the multi-scale modeling course together in Germany. During the course, potentially aided by The Original that you brought along, we set up an interesting MSc thesis project to co-supervise. Its a pity we never found a suitable student for the project, but we managed to keep our spirits high with some fun whiskey tasting events!

Benoit, meeting up was consistently a challenge due to our incompatible scheduling styles, but also consistently fun. I have thoroughly enjoyed the large diversity of cheese-based dinners we shared, as well as the usually following video games!

Nikolas, whether its video games in the evening, a well-organized LAN-party or the WeDay, you're always up for some fun. In all seriousness though, I've really enjoyed the gaming together!

Erika, thank you for recognizing a lacking skill set and subsequently personally initiating and overseeing my training in a field with clear practical applications: Vinology.

Bastian, as is to be expected from our fairly elected favorite PhD student, you truly were the social core of SSB. You took upon yourself the crucial task of making sure we had at least two coffee breaks a day, and you personally made sure everyone knew when they'd happen. Together with **Maarten** you ensured that the coffee breaks never had a dull moment, but also that I eventually really wanted to go back to work...

Talking about breaks, I would like to thank everyone from SSB and MIB for the very positive interactions during coffee breaks, lunch breaks, and lab activities.

Margo, you've been a great friend ever since we first started studying in Wageningen. Thank you for all the amazing pancakes, and for the countless hours of fun while climbing and playing 'extremely nerdy' board games.

Thomas, **Sanne**, I am really glad that you decided to move to Wageningen last year; It's been great spending considerably more time with you ever since. Our holidays to Finland together with **Dorett** were fantastic, just like our regular dinner and 'rikken' evenings!

Michael, I thought I'd never befriend a prince Carnival, but here we are. It was great having you around the first few years of the PhD, and you often organized something fun to do with the SSB group; games, laser tag, karting, kayaking, a rope park, climbing, etc... I really hope we keep up these short summer holidays together with **Dorett** and **Wen**!

Milad, **Dorett**, we have gone through great depths together during our friendship, but ultimately we emerged as certified divers from our holidays in Gozo. Thank you for that fantastic week!

Playing board/card games is one of my favorite things to do, and I'm glad I've had so many people to enjoy these games with during the last few years. **Margo**, **Dorett**, **Wen**, **Paul**, **Michael**, **Thomas**, **Sanne**, **Maria**, **David**,

Maaruthy, **Bastian**, **Niru**, **Rajaram**, **Teunke**, **Mark**, **Tim**, **Femke**, **Avalon**, **Ivar**, **Gerard**, **Mike**, **Ronald**, and all others who occasionally joined, thank you for all the great game days, afternoons and evenings!

Climbing and bouldering have been my other favorite pastimes these past years. **Margo**, **Dorett**, **Wen**, **Michael**, **Thijs**, **David**, **Alex**, **Floris**, and everyone else who joined from time to time, thank you for all the fun we've had climbing and bouldering in the gyms around Wageningen, as well as on our trip(s) to Fontainebleau!

Nong, **Marta**, **Dorett**, **Linde**, and **Nhung**, thank you for all the fun dinner evenings with amazing food!

Pooja, **Lekshmi**, **Gintas**, **Srivathsan**, **Maaruthy**, **Silvia**, **Sara**, **Filippo**, **Sumana**, and **Mathias**, thank you for the great time in Basel and for the occasional meet-up afterwards, let's keep it up!

I would also like to thank all the people I regularly play Magic with in Wageningen and Ede. I have really enjoyed being a part of this open community that helps each other to get better better at this great game, as well as the ease with which cards are being shared.

Ten slot wil ik graag al mijn familie bedanken voor het advies, het vertrouwen en de steun die ik de afgelopen jaren heb mogen genieten. **Mam**, jij in het bijzonder bedankt voor al je harde werk om mij, **Casper** en **Avalon** de beste kansen te geven in ons leven met de volledige vrijheid om het naar eigen wens in te vullen. **Pap**, bedankt dat je er altijd direct voor me bent als ik ergens hulp bij nodig heb. **Bert**, bedankt voor het meedenken en je advies gedurende de grotere beslissingen de afgelopen jaren. **Casper**, **Lily**, er is niks wat me meer verblijdt heeft dan jullie geweldige **Suki**. **Avalon**, bedankt voor alle leuke weekenden hier en in Londen en voor de talloze uren die je hebt gestoken in het nakijken en corrigeren van mijn bibliografie!

The research described in this thesis was financially supported by the IPOP program Systems Biology of Wageningen University and Research, and by the European Union's 7th Framework Programmes Microme (Project reference: 222886) and INFECT (Project reference: 321529), and by the European Union Horizon2020 project EmPowerPutida (Project reference: 635536).