FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
MIT Portugal
COMPETE 2020
PORTUGAL 2020
NORTE2020
PROGRAMA OPERACIONAL REGIONAL DO NORTE
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional
European Commission

Design and Optimization of Microbial Communities

Sophia Torres Santos

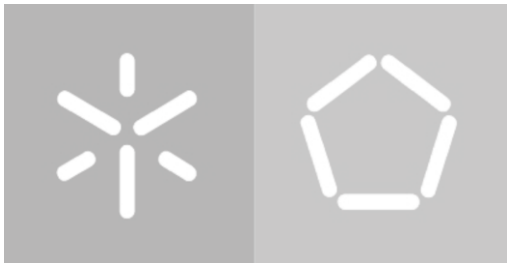
UMinho | 2024

Universidade do Minho
Escola de Engenharia

Sophia Torres Santos

Design and Optimization of Microbial Communities

abril de 2024



Universidade do Minho
Escola de Engenharia

SOPHIA TORRES SANTOS

DESIGN AND OPTIMIZATION OF MICROBIAL COMMUNITIES

TESE DE DOUTORAMENTO
DOUTORAMENTO EM BIOENGENHARIA

TRABALHO REALIZADO SOB ORIENTAÇÃO DE:
PROFESSORA DOUTORA ISABEL CRISTINA DE ALMEIDA PEREIRA DA ROCHA
E DE
DOUTOR OSCAR MANUEL LIMA DIAS

abril 2024

AGRADECIMENTOS

E que aventura fantástica está prestes a terminar! Mas antes de terminar quero agradecer a todos aqueles que tornaram este trabalho possível.

Agradeço à Professora Doutora Isabel Rocha. Obrigada por todos os conhecimentos científicos e pessoais transmitidos, pela constante disponibilidade e acompanhamento. Pelo incentivo, motivação e confiança depositada. Obrigada por me ajudar a crescer e a abrir horizontes. Será sempre um exemplo para mim.

Agradeço ao Doutor Oscar. Por ter aceitado o desafio de me orientar, de me aconselhar, de me motivar. Por me estimular a perseguir outros desafios e a dar sempre mais de mim em cada tarefa proposta.

Agradeço à instituição de acolhimento, Centro de Engenharia Biológica da Universidade do Minho, por me ter proporcionado as condições necessárias à realização deste trabalho.

Agradeço também à Fundação para a Ciência e Tecnologia pelo financiamento atribuído através da bolsa de doutoramento SFRH/BD/121695/2016.

Agradeço em especial a todos os parceiros do Instituto de Tecnologia Química e Biológica António Xavier, Instituto Gulbenkian e Ciência e Universidade dos Açores, em especial à Ana e ao Ricardo, que sem o seu trabalho e ajuda este trabalho não seria possível.

Agradeço a todos os elementos do grupo BioSystems, presentes e passados. Obrigada pelo bom ambiente de trabalho que sempre proporcionaram, pelas discussões científicas e partilha de ideias. Um agradecimento ao Vítor Pereira pela ajuda na parte da implementação dos métodos de otimização no contexto de comunidade microbianas.

À Joana que me acompanhou desde o início nesta caminhada doutoral. Obrigada pela partilha de conhecimentos, pelas longas horas de conversa, pelas gargalhadas, pelas conversas sérias e pela amizade conquistada.

À Ana e à Sara e "às frizes" que bebemos e que espero que continuemos a beber.

À minha família e amigos. Por estarem sempre presentes e serem um presente na minha vida.

À minha mãe, a quem devo todos os meus valores e educação. Por sempre me encorajar a ser melhor.

Ao Dinho, pelo amor, dedicação e paciência. Por ser o meu porto seguro e aconchego. Pelo incentivo a me superar todos os dias. Por nunca me deixar desistir.

Ao Tomé, que renovou a minha vida!

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, April 15th, 2024

_____

Sophia Torres Santos

ABSTRACT

Design and Optimization of Microbial Communities

Microbial communities directly affect surrounding environments and are an important biological process, with potential applications in a variety of fields, such as biotechnology, environmental, and human health. However, the overall understanding of interactions and dynamics in microbial communities remains a challenge. Synergies between computational methods and genome-scale metabolic models have been explored in the last years, as a way to unravel community interactions and behavior, as demonstrated by the numerous simulation methods developed for application in the context of microbial communities. The available simulation methods, with application to microbial communities, were here evaluated and revealed good predictions for phenotypic behavior. However, few studies are available in terms of optimization tools in the community context. Hence, this work describes the implementation of algorithms for the optimization of minimal medium composition, as well as genes/reactions for the production of target compounds. These tools were implemented in MEWpy to transform it into an integrative Python workbench for metabolic engineering to explore constraint-based models of microbial communities.

Five hydrothermal samples from the São Miguel Island, Azores, were analyzed to determine prokaryotic community composition to further reconstruct individual and community genome-scale metabolic models, and through simulation and design methods try to unveil possible routes to produce compounds with industrial interest. The first manually curated genome-scale metabolic model for the thermophilic bacterium *Sulfurihydrogenibium azorense* Az-Fu1 was developed, uncovering the details of its metabolic capabilities and suggesting for the first time that *S. azorense* Az-Fu1 may have metabolic potential for bacterial cellulose production. Moreover, the microbial communities of the different samples were modeled, and co-culture optimization was performed using the implemented methods. Among other results, it was shown that *S. azorense* Az-Fu1 can enhance its cellulose production capabilities when fed with acetate produced by another organism.

Keywords: Design, extremophile environments, genome-scale metabolic modeling, microbial communities, optimization.

RESUMO

Design e Optimização de Comunidades Microbianas

As comunidades microbianas afetam diretamente os ambientes circundantes e são um processo biológico de enorme relevância, com aplicações potenciais em vários campos, como biotecnologia, meio ambiente e saúde humana. No entanto, a compreensão geral das interações e dinâmicas nas comunidades microbianas continua um desafio. Têm sido exploradas nos últimos anos diversas sinergias entre métodos computacionais e modelos metabólicos à escala genómica como forma de desvendar interações e comportamentos de comunidades, tal como demonstrado pelos inúmeros métodos de simulação desenvolvidos para aplicação neste contexto. Os métodos de simulação disponíveis com aplicação a comunidades microbianas foram avaliados e revelaram boa capacidade preditiva do comportamento fenotípico. No entanto, poucos estudos estão disponíveis em termos de ferramentas de otimização no contexto de comunidades microbianas. Assim, este trabalho descreve a implementação de algoritmos para a otimização da composição mínima do meio, bem como genes/reações para a produção de compostos alvo. Essas ferramentas foram implementadas no MEWpy de forma a transformá-lo em um ambiente de trabalho Python integrado para engenharia metabólica de comunidades microbianas.

Cinco amostras hidrotermais da ilha de São Miguel, Açores, foram analisadas para determinar a composição das comunidades procarióticas de forma a reconstruir modelos metabólicos individuais e comunitários à escala genómica e, através de métodos de simulação e design, tentar desvendar possíveis formas de produzir compostos com interesse industrial. Foi desenvolvido o primeiro modelo metabólico à escala genómica manualmente curado para a bactéria termofílica *Sulfurihydrogenibium azorense* Az-Fu1, revelando detalhes das suas capacidades metabólicas e sugerindo, pela primeira vez, que *S. azorense* Az-Fu1 pode ter potencial metabólico para produção de celulose bacteriana. Além disso, foram modeladas as comunidades microbianas das diferentes amostras e a otimização de co-culturas foi realizada usando os métodos implementados. Entre outros resultados, foi demonstrado que *S. azorense* Az-Fu1 pode aumentar a sua capacidade de produção de celulose quando suplementado com acetato produzido por outro organism.

Palavras-chave: Ambientes extremófilos, comunidades microbianas, modelação metabólica à escala genómica, otimização.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

Motivation, Objectives, and Thesis Outline

"Our unity is our strength, and diversity is our power"

Kamala Harris

Microbial communities are widespread in nature and have been recognized to be more adequate in industrial settings than pure cultures, especially regarding the robustness towards contaminations. Moreover, the metabolic capabilities can be greatly extended when compared with individual species. However, very few applications have been described besides the production of food products using natural communities. In order to use fully the industrial potential of microbial communities it is important to be able to customize their behavior by optimizing the composition both in terms of species and genes, as well as optimizing the environmental conditions.

This thesis aims to contribute to developing optimized microbial communities for the production of a target compound by improving both modeling and simulation methods and developing metabolic engineering tools. The *in silico* results will be applied to the optimization of the production of cellulose by *Sulfuhydrogenibium azorense* Az-Fu in a co-culture system.

Context and Motivation

Microbial communities play pivotal roles throughout all environments showing unique biological properties in maintaining Earth's biosphere (Fierer, 2017; Gilbert et al., 2018; Sunagawa et al., 2015). Microbial communities in environments that once were described as unhabitable (Rothschild et al., 2001) are the ones that nowadays are gaining huge interest due to their remarkable diversity of morphology, biochemistry, genomics, and biosynthesis of many adaptive compounds (Durvasula et al., 2018). Moreover, microbial communities in high-temperature environments are generally less diverse (Inskeep et al., 2013), making hydrothermal habitats an ideal model system for studying principles of community structure and function (Sahm et al., 2013).

Genome-scale metabolic models (GSMM) are a relevant example with proven results in accurately predicting metabolic states for an increasing number of single organisms (Feist et al., 2009) as well as for microbial communities (Colarusso et al., 2021), reducing the time and cost implicated in experimental tasks. Together with the growing availability of 'omics' data and multiple algorithms to process and make knowledge from metagenomic raw data, these metabolic models are an indispensable systems biology tool to try to understand function, interaction, and dynamics within these microbial consortia (Zaramela et al., 2021).

GSMMs are usually used to compute metabolic phenotypes of an organism in response to environmental and genetic perturbations (Feist et al., 2010; Rocha et al., 2008) most commonly through simulation methods, such as Flux Balance Analysis (FBA) (Varma & Palsson, 1994). However, the construction of GSMMs with good phenotypic predictions needs manual curation and experimental validation (Lieven et al., 2020), which is still a laborious task even with available semi-automatic reconstruction pipelines (Arkin et al., 2018; Machado et al., 2018). But more challenging than that is trying to define biologically relevant objective functions (García-Jiménez et al., 2021) in a microbial community context.

A myriad of extension methods for FBA with application to community GSMMs is available (Chan et al., 2017; Gomez et al., 2014; Mahadevan et al., 2002; Zelezniak et al., 2015), with proven reliability when predicting growth, response to nutrients, and gene essentiality in single organisms and even microbial communities (Chng et al., 2020; Machado et al., 2021; Nayfach et al., 2020). However, not much research has been made on computational optimization methods to rationally design microbial

communities using GSMMs, which is one of the most promising features for the use of community models for the production of desired compounds in industrially relevant amounts (García-Jiménez et al., 2021; Tsoi et al., 2019; Zaramela et al., 2021). This is particularly relevant given that microbial communities are often easier to use than pure cultures in industrial settings.

Research Aims

The present thesis aims at reconstructing extremophilic microbial communities' metabolic models and performing simulations with the ultimate objective of using those models to design communities better fitted for industrial aims. Microbial communities are often more adequate than pure cultures for industrial applications and have, at the same time, great potential in terms of metabolic capabilities by combining individual capacities of single strains.

The specific objectives are:

- Identifying, through metagenomic taxonomic profiling approaches, extremophile organisms on hydrothermal samples from São Miguel, Azores;
- Developing genome-scale metabolic models of the extremophile organisms identified to further combine individual models to obtain an accurate representation of the capabilities of a microbial community;
- Evaluating existing tools and methods with an emphasis on Flux Balance Analysis based techniques, and using the most promising ones to perform simulations using as a case study the well-established nitrification bioprocess catalyzed by *Nitrosomonas euroapaea* and *Nitrobacter vulgaris*.
- Adapting metaheuristics algorithms previously developed at the Biosystems group to be applied to the design of improved communities for industrial aims, by allowing the manipulation of the genome of untargeted and target species, and the environmental conditions;
- Use the selected simulation tools and methods and metaheuristics algorithms developed to find design strategies for the production of cellulose, using extremophile community models.

Outline of the Thesis

To address the above-mentioned objectives, this thesis has been structured into 7 Chapters:

- In **Chapter 1** a comprehensive review of systems biology approaches on microbial communities was conducted, trying to give an overall view of the metagenomic analysis process and how systems biology using genome-scale metabolic reconstruction, simulation, and optimization can improve knowledge on function, interaction, and dynamics within these microbial consortia.
- In **Chapter 2** a metagenomic study of five hydrothermal samples from São Miguel, Azores, was performed to determine their prokaryotic community compositions. Taxonomic profiling using assembly-based and read-based analysis was combined and the most abundant organisms were predicted for each sample. *Sulfurihydrogenibium azorense* Az-Fu1 was one of the most abundant predicted organisms found in two of the hydrothermal samples.
- In **Chapter 3** the first manually curated genome-scale metabolic model for the thermophilic bacterium *Sulfurihydrogenibium azorense* Az-Fu1 is presented. The reconstruction revealed the presence of the main components of the bacterial cellulose operon and its regulators, suggesting that *Sulfurihydrogenibium azorense* Az-Fu1 may have metabolic potential for cellulose production.
- In **Chapter 4** a systematic evaluation of different steady-state simulation methods applied to microbial communities has been used to model the well-established nitrification bioprocess catalyzed by *Nitrosomonas euroapaea* and *Nitrobacter vulgaris*. Methods performances were compared to assess which ones should be used in a specific community-level context.
- In **Chapter 5** The implementation of metabolic engineering methods in MEWPy to explore constraint-based models of microbial communities allowing the optimization of microbial communities using Evolutionary Algorithms is described. MEWpy also allows the use of the simulation methods for microbial communities included in the REFRAMED library.
- In **Chapter 6** the microbial communities of the different samples were modeled and both untargeted and targeted co-culture optimization was performed using the methods described in the previous chapter to evaluate *Sulfurihydrogenibium azorense* Az-Fu1 capabilities of cellulose production.
- Finally, in **Chapter 7** the main conclusions of this thesis are summarized and some perspectives on future work based on the questions raised throughout this work are presented.

Scientific Output

The main scientific outputs of this thesis are listed below and include peer-reviewed publications and oral presentations at international conferences.

Peer-reviewed Publications

Santos, S., Dias, O., Rocha, I. *Genome-scale metabolic model of thermophilic bacterium Sulfurihydrogenibium azorense Az-Fu1* (in preparation).

Santos, S., Pereira, V., Dias, O., Rocha, M., Rocha, I. *Optimization of Microbial Communities* using MEWpy (in preparation).

Santos, S., Dias, O., Rocha, I. *Modeling and Design of Microbial Communities from Extremophilic Environments in the Azores* (in preparation).

van den Berg, N.I., Machado, D., **Santos, S**. et al. *Ecological modelling approaches for predicting emergent properties in microbial communities*. Nat Ecol Evol (2022).

Oral Presentations

S. Santos, S. Correia, I. Rocha. *Inferring optimal minimal medium on genome-scale metabolic models using evolutionary algorithms.* – Metabolic Pathway Analysis Conference, Riga, Latvia, 12-16 August 2019.

Chapter 1

Systems Biology in Microbial Communities

"Coming together is a beginning. Keeping together is progress.
Working together is success"
Henry Ford

Microbial communities are widespread in nature and have been recognized to be more adequate in industrial settings than pure cultures, especially regarding the robustness towards contaminations. Moreover, the metabolic capabilities can be greatly extended when compared with individual species. However, very few applications have been described besides the production of food products using natural communities. It is important to be able to customize the microbial communities' behavior by optimizing the composition both in terms of species and genes, as well as optimizing the environmental conditions to use fully the industrial potential.

This chapter will review systems biology approaches applied to microbial communities giving insights into how can genome-scale metabolic reconstruction, simulation, and optimization improve knowledge of function, interaction, and dynamics within these microbial consortia.

1.1 Introduction

Microbial communities serve as integrated systems in many natural processes playing important roles in a wide range of areas such as industrial (Park et al., 2018), environmental biotechnology (Zhou et al., 2011), and human health (Jansma et al., 2021). Moreover, the use of microbial communities has been largely promoted as an alternative to improve limitations of rational design of single organisms to produce target chemicals (Sgobba et al., 2020; Wang et al., 2020). However, regardless of the growing availability of 'omics' data, it is still difficult to fully understand the function, interaction, and dynamics within these microbial consortia (Zaramela et al., 2021) essentially due to the lack of experimental studies.

Genome-scale metabolic models (GSMM) have become a key and valuable tool for the study of metabolic systems biology from biomedical to industrial research (Gu et al., 2019). Since the publication of the first metabolic model (Schilling et al., 2000) more than twenty years ago, others are becoming available for an increasing number of single organisms (Feist et al., 2009) and more recently also for microbial communities (Colarusso et al., 2021). The main purpose of metabolic models is to predict cellular behavior under different genetic and environmental conditions (Woolston et al., 2013) reducing the time and cost implicated in experimental tasks.

The construction of GSMM for microbial communities, as for single organisms, can be a laborious task. In general, the exact composition of a community is not known, and even when 'omics' data are available, manual curation and experimental validation are needed to create a GSMM with good phenotypic predictions (Lieven et al., 2020). Considering these difficulties in having reliable data, several community models include only the central carbon metabolism (Stolyar et al., 2007), use semi-automatic reconstruction pipelines (Arkin et al., 2018; Machado et al., 2018), or represent artificial communities of very well-characterized organisms for which detailed individual models are already, such as *Escherichia coli* and *Saccharomyces cerevisiae* (Brenner et al., 2008; Hanly et al., 2011).

GSMMs are usually used to compute metabolic phenotypes of an organism in response to environmental and genetic perturbations (Feist et al., 2010; Rocha et al., 2008). The most commonly used simulation methods are Flux Balance Analysis (FBA) (Varma et al., 1994) or FBA extensions (Chan et al., 2017; Gomez et al., 2014; Mahadevan et al., 2002; Zelezniak et al., 2015). These same simulation methods developed for single-organism GSMMs have been adapted and employed for the simulation with community GSMMs. However, the major challenge for the use of FBA-based methods with community

models is the selection of biologically relevant objective functions (García-Jiménez et al., 2021). In fact, while for individual microbes, the assumption of growth maximization applies, an abstraction of the aim of a community is difficult to devise. Also, although the use of these approaches has given good phenotypic predictions using relatively large-scale community models, problems may arise with the increased complexity of the metabolic network or a higher number of organisms within the communities (Zomorrodi et al., 2016). Thus, the development of efficient predictive mathematical modeling approaches and scalable tools that give good phenotypic predictions, as well as the improvement of community GSMM and the subsequent experimental validation is mandatory such that these modeling frameworks can be applied to real communities that often have a high degree of complexity.

Moreover, one of the most promising features for the use of community models is the rational design of microbial communities that could turn them capable of producing desired compounds in industrially relevant amounts (García-Jiménez et al., 2021; Tsoi et al., 2019; Zaramela et al., 2021). This is particularly relevant given that microbial communities are often easier to use than pure cultures in industrial settings. Also, it would be possible in principle to optimise the production of specific compounds using simple substrates taking advantage of the unique metabolic capabilities of a community that single organisms are not capable of covering. The metabolic engineering or design problem could be simply formulated as the maximization of the production of a target compound by manipulating either environmental conditions, community composition in terms of species, or by performing genetic manipulations in targeted organisms of the community. However, not much research has been performed on extremophiles in this direction and many details still need to be further investigated for the successful implementation of such an approach.

1.2 Systems Biology

The huge development of molecular biology and technology was the basis for the appearance of the fields of study known as omics. Despite the large amount of data that these approaches generate every day, living systems are complex and their behavior is thus difficult to predict over time under various conditions. Systems biology is an interdisciplinary field that studies complex interactions over all omics levels of a biological system. Systems biology aims to understand the cell as a whole instead of the individual parts, studying its structure, dynamics, control, and design methods (Kitano, 2002). Integrating

computational tools and theoretical approaches with experimental efforts enabled the development of predictive and quantitative models to understand how the different parts of a biological system interact with each other. The goal of such models is to predict the behavior of the cells under various genetic and environmental conditions rather than mimicking cell behavior (Isalan, 2012). The increase in high-throughput data led to the development of new algorithms that integrate omics data into models, some of them use machine learning approaches (Antonakoudis et al., 2020), which improve (Ramon et al., 2018) or reduce the model's network (Singh et al., 2020).

Strain design has been given special attention in the past years due to the increasing demand for the biological production of chemicals, pharmaceuticals, food ingredients, and enzymes (Stephanopoulos, 1999). In the early days, strain development was carried out by random mutagenesis, phenotype screening, and selection, or target modifications. All of these methods have a high degree of failure and are time-consuming and expensive (Woolston et al., 2013). Metabolic models allied with bioinformatics tools have been used successfully to identify genetic targets towards improved phenotypes, rationally guiding laboratory experiments (Maia et al., 2016). However, the production of high-valued compounds is not always cost-effective and great efforts must be made to improve yield and productivity (García-Jiménez et al., 2021).

Recently, advances in the big data areas of metagenomics, metatranscriptomics, and metaproteomics, led to engineering of natural and synthetic microbial communities. These are described as better platforms gaining from their heterogeneity, division of labor, and feedstock utilization (Sgobba et al., 2020). Hence, the use of metabolic models in a microbial community context has become the new frontier in systems biology to generate testable hypotheses (Colarusso et al., 2021).

1.3 Genome-Scale Metabolic Modelling

GSMMs are mathematical representations of the complete set of biochemical reactions encoded in the genome of an organism. They can be used to simulate the metabolic phenotype of an organism under different experimental conditions and its response to multiple environmental and genetic perturbations (Feist et al., 2010; Rocha et al., 2008). Hence, these models have become widely adopted *in silico* tools in the context of biotechnological applications such as microbial strain design, drug discovery, and more recently understanding microbial community interactions (Fang et al., 2020).

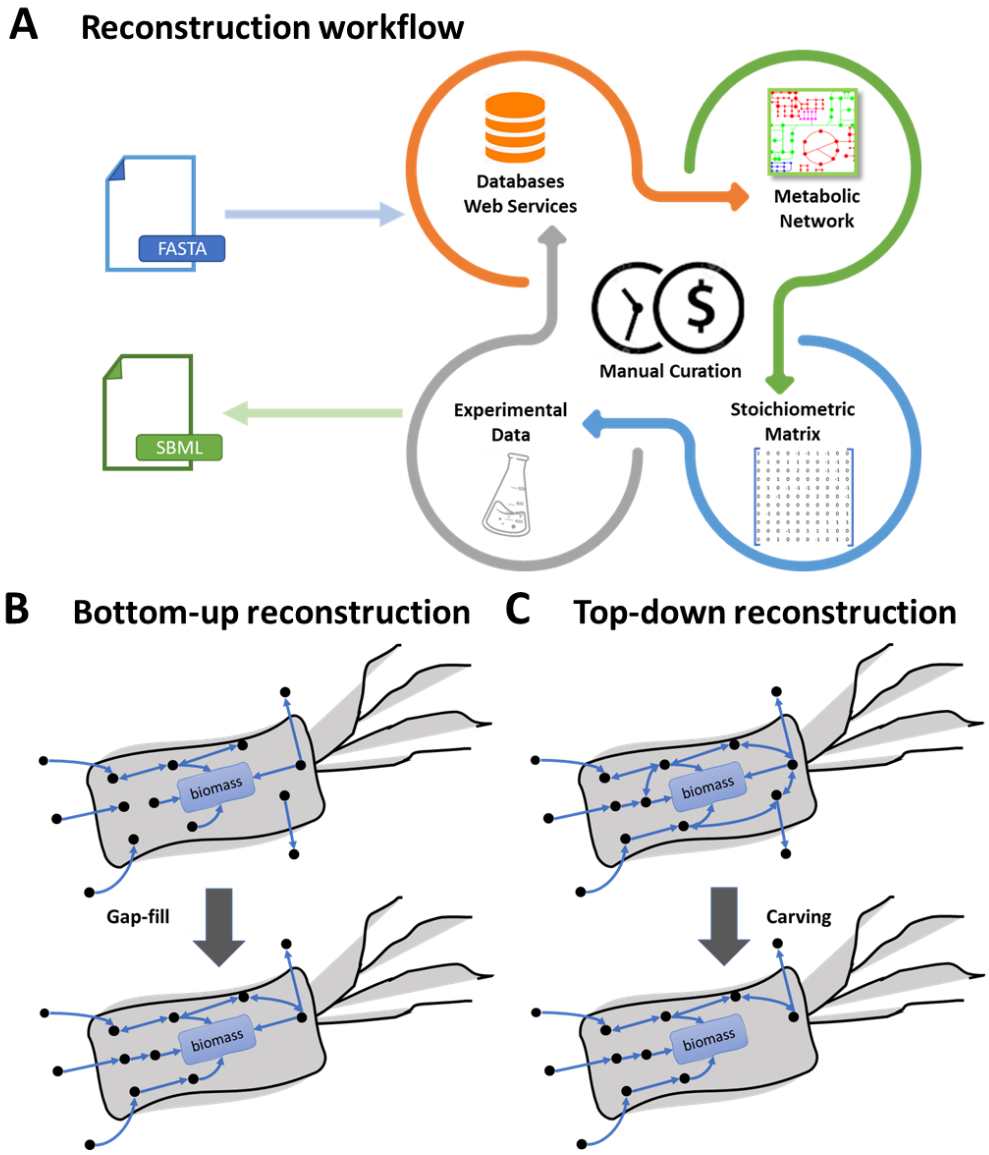


Figure 1.1. General overview of genome-scale metabolic model reconstruction process (A) This is an iterative process that needs a FASTA file with the organism genes sequences. Information from functional annotation and reactions are retrieved from databases and/or web services to assembly a metabolic network. Then the metabolic network is transformed into a stoichiometric model when constraints are added. Finally, *in silico* results are compared with experimental data. Once predictions comply with experimental data the model is exported in a standard format (SBML) to be used in further applications (B) In a bottom-up reconstruction approach new reactions are iteratively added to the network for gap-filling purposes. (C) In a top-down reconstruction, gapless pathways are inferred from genetic evidence, and the ones with low evidence are removed from the pre-existent universal model.

1.3.1 Genome-Scale Metabolic Models Reconstruction

Genome-scale metabolic reconstructions are usually arduous (Figure 1.1) and time-consuming, as demonstrated by a detailed protocol published in 2010 (Thiele et al., 2010), with 96 steps. Since then,

several tools have been published that assist and automate most of the steps of the reconstruction process (Mendoza et al., 2019).

Most of the published tools follow a bottom-up approach consisting of the following main steps: genome annotation, assembling of a metabolic network from the genome, conversion of the network to a stoichiometric model, and validation of the metabolic model. CarveMe (Machado et al., 2018) has a unique top-down approach that involves the creation of models from a BiGG-based (King et al., 2016a) manually curated universal template. However, the output model obtained from all these tools must be, to some extent, manually curated to be able to predict accurately phenotypic behavior (Lieven et al., 2020). To facilitate the reconstruction process, online databases are essential to retrieve information or assist during manual curation. The main databases usually used throughout the GSMM reconstruction process are presented in Table 1.1. The main steps of a GSMM reconstruction including the main databases used for each step are concisely described next.

Table 1.1 Main online data sources used for the reconstruction of genome-scale metabolic models.

Database	Description	Reference
BioCyc	BioCyc is a collection of Pathways/Genome Databases (PGDBs) for a vast number of prokaryotes and eukaryotes. It also includes software tools for exploring each PGDBs.	(Karp et al., 2019)
BIGG	Biochemical, Genetic, and Genomic (BIGG) is a knowledge base of genome-scale metabolic models. Contains information about genes, reactions, and metabolites included in published manually curated models. All information is mapped to external databases (NCBI, KEGG, and others)	(King et al., 2016a)
BRENDA	BRaunschweig ENzyme Database (BRENDA) is a database with enzyme functional data. Contains functional and molecular information of enzymes, based on primary literature.	(A. Chang et al., 2021)
ExPASy	Expert Protein Analysis System (ExPASy) is the Swiss Institute of Bioinformatics Resource Portal in different areas of the life sciences including genomics, proteomics, and structural biology.	(Ison et al., 2013)

GOLD	Genomes OnLine Database (GOLD) is a manually curated collection of genome projects and their metadata.	(Mukherjee et al., 2021)
KEGG	Kyoto Encyclopedia of Genes and Genomes (KEGG) is an online public repository that is a combined collection of information on genes, metabolites, reactions, and pathways.	(Kanehisa et al., 2016)
MetaCyc	MetaCyc is a database of non-redundant metabolic pathways. MetaCyc is curated from the scientific literature and contains pathways involved in primary and secondary metabolism, as well as associated compounds, enzymes, and genes.	(Caspi et al., 2014)
ModelSEED	ModelSEED is a resource for the reconstruction, exploration, comparison, and analysis of metabolic models. Contains information about genes, reactions, and metabolites included in GSMMs and links for external databases (BiGG, KEGG, MetaCyc, and others).	(Seaver et al., 2021)
NCBI	The National Center for Biotechnology Information (NCBI) is a repository of several databases that provide analysis, visualization and retrieval resources for biomedical, genomic and other biological data made available through the NCBI website.	(NCBI Resource Coordinators, 2018)
TCDB	Transporter Classification Database (TCDB) comprehends a classification system for membrane transporter proteins known as the Transporter Classification system.	(Saier et al., 2021)
UniProt	Universal Protein Resource Knowledgebase (UniProtKB) is the central hub for the collection of accurate, rich and consistent functional information on proteins.	(Consortium et al., 2021)

1.3.1.1 Genome Annotation

Genome-scale models' reconstructions start with the genome annotation process of the target organism. This is a crucial step as the assignment of an incorrect gene or enzymatic function can greatly impact the model performance. Metabolic gene products are assigned with functions and, if available,

unique identifiers such as Enzyme Commission (EC) (Barrett, 1997) and Transporter Classification (TC) (Saier et al., 2021) numbers. Genes involved in regulatory or signaling processes are not included in GSMM reconstructions. Public repositories of genomic data have available annotated genomes, such as NCBI and KEGG. However, in specific cases annotation of genes may be incorrect or even missing; therefore, a re-annotation is always recommended. Information from phylogenetically closely related organisms can also be used to improve genome annotation.

1.3.1.2 Assembling the Metabolic Network

The assembly of a metabolic network begins with the identification and collection of all reactions present in the organism in study, from biological databases such as KEGG and BiGG. All reactions catalyzed by enzymes encoded by genes annotated with EC numbers in the previous step must be incorporated into the model, as well as spontaneous reactions. Reactions should also be balanced (generic metabolites and/or metabolites without formula must be curated), and reversibility must be confirmed to avoid mispredictions. Although these steps are automated in most reconstruction tools, manual curation based on curated information from literature or databases, such as MetaCyc, BRENDA, and KEGG is highly recommended.

Compartmentalization

The next step in the reconstruction process is the compartmentation of the reactions, identifying all the organelles in which enzymes can operate. In prokaryotic organisms, compartments are typically limited to extracellular space, cytosol, and periplasm (in gram-negative bacteria). However, for instance, in Fungi, reactions can occur in up to sixteen compartments (Lu et al., 2019), and, for higher eukaryotes, reactions should be differentiated between tissues. Information on the location of reactions is available in the literature, but several bioinformatics tools have been developed to predict enzymatic location from protein sequences (Jiang et al., 2021)

Genes, Proteins, and Reactions

A high-quality GSMM should include curated Gene-Proetin-Reactions relationships (GPRs), which allows the accurate prediction of the effect of genetic modifications. These associations are usually defined according to databases and literature (Rocha et al., 2008; Thiele et al., 2010). Most GSMM

reconstruction tools, such as *merlin* (Capela et al., 2021), ModelSEED (Seaver et al., 2021), and CarveMe (Machado et al., 2018) do automatically include GPRs rules in the model.

Transport Reactions

After compartmentalization, reactions that transport metabolites across the inner and outer membranes must be added to the model. Again, literature information and databases contain these reactions. However, tools such as *TranSyT* (Lagoa et al., 2021) can use genome information, to generate predicted system-specific transport reactions, providing the associated gene-protein-reaction (GPR) rule.

1.3.1.3 Converting the Metabolic Network to a Stoichiometric Model

In this step, the metabolic reaction set is converted into a stoichiometric matrix, and other constraints are added to the model. These constraints include the definition of an abstraction of the biomass in the form of a mathematical equation that represents the drain of macromolecules to generate a new unit (gram) of biomass, energetic requirements through the inclusion of equations that represent the depletion of ATP for cell growth and maintenance, and constraints that represent the environmental conditions.

Biomass Composition

The composition of the biomass should be experimentally determined through chemostat experiments, or alternatively during the log phase of the cell's growth. However, in the absence of organism-specific experimental information, data from genome information (particularly nucleotides, deoxynucleotides, and amino acids) can be used or adapted from phylogenetically related organisms. The importance of an accurate biomass composition determination has been reported (Santos et al., 2016), showing that even small differences in biomass content coefficients may considerably impact GSMMs' predictions. Therefore, a well-defined biomass equation is a crucial step of the GSMM reconstruction process.

For *n* biomass constituents, the biomass equation can be formulated as:

$$\sum_{i=1}^{n} c_i X_i \rightarrow Biomass$$

where c_i is the coefficient of each macromolecule or building block, X_i, considered in the biomass. The units of all the coefficients are defined in mmol per gram of dry weight (mmol/gDW) and the biomass units are defined per hour (h^{-1}).

The formulation of the biomass objective function can be obtained at different levels of detail: basic level (define the macromolecular content on the cell, i.e., protein, RNA, DNA, lipids), intermediate level (calculate the necessary biosynthetic energy), and advanced level (detailing the necessary vitamins, elements, and cofactors) (Feist et al., 2010). If a biomass precursor is not considered in the biomass reaction, synthesis reactions may not be required for growth, as well as associated genes, which play an important role in *in silico* gene deletion experiments (Thiele et al., 2010).

Growth and Maintenance ATP Requirements

Growth-associated ATP maintenance (GAM) reaction accounts for the energy required to replicate a cell, including the synthesis of macromolecules (e.g., Proteins, DNA, and RNA). GAM should be experimentally determined through chemostat experiments. When no experimental data is available, GAM can be estimated by calculating the number of phosphate bonds for macromolecular synthesis (Neidhardt et al., 1990).

Non-growth associated ATP (NGAM) maintenance reaction represents the ATP requirements to the cell to maintain for example its membrane leakage (Feist et al., 2007), represented by n ATP hydrolysis reaction, which transforms ATP into ADP and phosphate. The following reaction should then be included in the model's reaction set:

$$x\ ATP + x\ H_2O \rightarrow x\ ADP + x\ P_i + x\ H^+$$

where x is the number of phosphate bounds and the NGAM value is set as its reaction rate. As for GAM, the value for the NGAM reaction rate should be experimentally determined or can either be found in literature or estimated by fitting the mode results to experimental data (Rocha et al., 2008).

Medium Constraints

Transport fluxes for nutrients in the medium should be constrained between zero and infinity. These constraints allow replicating the organism's physiological behavior under limited substrate availability or maximal uptake rates. The limiting substrate must be constrained to a specific rate value while the remaining ones are usually left unconstrained. Nutrients not available in the medium should be constrained to zero (Rocha et al., 2008).

1.3.1.4 Metabolic Model Curation and Validation

The model curation step is an iterative process that continues until simulation results match experimental data in the literature. Tools, such as *merlin* (Dias, Rocha, Ferreira, & Rocha, 2018), provide a graphical user interface that allows to easily perform re-annotation, correct reactions stoichiometric balance and directionality, include and/or exclude a reaction from the model, and ultimately to export the model in a standard and computational friendly format, such as in Systems Biology Markup Language (SMBL) format.

Gap-filling

The metabolic network must be screened for possible gaps. The presence of gaps can compromise the synthesis of biomass components and therefore the predictability of the model. Literature and databases (KEGG, MetaCyc, BRENDA, for instance) information should be used to assist the gap-filling process. Several tools can automate the network gap-filling (Mendoza et al., 2019). However, manual curation should follow such procedure. This is an iterative process that is repeated until all biomass precursors and other relevant compounds can be synthesized, and a feasible model is obtained.

Metabolic Model Consistency Tests and Comparison to Experimental Data

Once gap-filling is performed, consistency tests can verify the actual phenotypic prediction accuracy of the GSMM. Firstly, the metabolic model must be able to compute growth under specific conditions and the predicted rates assessed to published characterization studies (including growth, secretion, and uptake rates). Growth under a defined growth medium is strongly advised whenever possible (Feist et al.,

2008). Specific growth conditions can include growth in limiting substrates, aerobic and anaerobic conditions, and different growth conditions reported in the literature. Also, the analysis of active pathways under specific growth conditions can be performed for model validation. If GSMM predictions are not in accordance with the experimental results, the model should be examined, potentially missing reactions included, and incorrect reactions removed. Another approach for validating GSMMs is assessing the simulation results of experiments performed with deletion mutants. This approach can provide valuable insights into the predictive capabilities of the model and a good training set may be of great value for model debugging.

After this validation step, the high-quality GSMM should be exported in a standard format, such as Systems Biology Markup Language (SBML) to be used on GSMM simulation platforms. A high-quality standardized test of a GSMM can be accessed using the MEMOTE test suit (Lieven et al., 2020), which tests fundamental SBML semantic and conceptual requirements. A variety of platforms for simulation with GSMMs are available, such as the user-friendly java tool Optflux (Rocha et al., 2010), Matlab® through the COBRA toolbox (Heirendt et al., 2019)), and Python through COBRApy (Ebrahim et al., 2013) or Reframed packages.

1.3.2 Simulation with Genome-Scale Metabolic Models

The most common GSMM simulation approach, Flux Balance Analysis (FBA) (Varma & Palsson, 1994), calculates intracellular reaction rates (also known as flux distributions), under a steady-state assumption (no accumulation of internal metabolites), along with nutrient uptake and secretion, and the cellular growth rate (Orth et al., 2010). Since the mathematical representation of a GSMM forms an underdetermined system of linear equations, the determination of the most biologically plausible solution requires the specification of an objective function. The metabolic network can be represented by a stoichiometric matrix S , of dimensions $m \times n$, where m corresponds to the total number of metabolites and n to the total number of reactions in the network. The coefficients of the S matrix define the relationship between the reactions and compounds of the metabolic network. An optimal solution, consistent with the known constraints and the maximization or minimization of a given objective function (Z), can be obtained by solving the linear problem:

$$Maximize \quad Z$$

$$subject\ to \quad Sv = 0$$

$$\propto_j \leq v_j \leq \beta_j, \qquad j = 1, \dots N$$

where v is a vector of fluxes of each reaction, and $\propto_j$ and β_j are the lower and upper limits for the fluxes, respectively. These limits are used to set the reversibility of the reactions, to limit uptake and secretion rates, and eventually to specify measured fluxes (Rocha et al., 2008). The solution to a FBA problem with respect to a given objective function is a set of optimal metabolic flux distributions. Although the objective value found is unique, flux distributions rarely are. To tackle this problem a second optimization criterion can be added such as found in the FBA variants of Parsimonious enzyme usage Flux Balance Analysis (pFBA) (Lewis et al., 2010), which attempts to find a more biologically viable flux distribution, by trying to minimize the absolute value of the sum of all fluxes through all reaction in the network while maintaining the reached optimum or Flux Variability Analysis (FVA) (Mahadevan et al., 2003), which sets a constraint that requires the objective flux to be equal to its optimal value, and assess the robustness of a flux distribution regarding its production capability of the target compounds. Robustness is analyzed by maximizing and minimizing each reaction flux. High robustness is verified when the predicted maximum and minimum flux values differ slightly.

The major challenge in FBA is the definition of an objective function, with biological relevance (Gianchandani et al., 2010). A variety of objective functions have been used to define an FBA problem. However, in the case of microbes, the maximization of growth rate is the most commonly used assumption based on selection pressure (Feist et al., 2010). Other objective functions can be used, such as minimizing ATP production or maximizing the desired compound. FBA has a wide range of applications, such as the optimization of bio-processes in industries or the identification of drug targets (Raman et al., 2009).

1.3.3 Optimization using Genome-Scale Metabolic Models

Genetic manipulation of microbes for industrial purposes is nowadays widely used to produce bio-based, sustainable, environmentally friendly, and viable compounds (Stephanopoulos, 1999). GSMMs are in that matter a commonly used tool for rational strain design by biotechnology companies with

successful applications (Jullesson et al., 2015) aided by the development of computational strain optimization methods. These methods try to identify a set of possible genetic modifications towards a desired phenotypical trait. Typically, the objective is to optimize the production of the desired compound, maintaining cell viability. Based on this bi-level objective, OptKnock (Burgard et al., 2003) established the foundation for the development of many other computational strain optimization methods. A variety of computational strain design optimization methods have been developed to search for non-intuitive genetic designs in more efficient and scalable ways. Today over 30 different computational strain optimization methods are published and can be divided into 3 main branches: bi-level mixed-integer programming, metaheuristics, and elementary-mode analysis-based methods. These methods can use different options for optimization targets (genes, reactions) or tasks (deletions, insertions, up/down-regulation) (Maia et al., 2016). Most of the mixed-integer programming and elementary-mode analysis methods guarantee to reach a global optimal solution. However, such approaches do not scale well with larger models or a high number of perturbations. Alternatively, metaheuristic-based methods, such as Evolutionary Algorithms and Simulated Annealing, not only usually have good scalability but are more flexible to the specification of objective functions (Rocha et al., 2008). However, these methods do not guarantee that a globally optimal solution is found.

Accessing and using these methods is not always straightforward as the implementation of some methods is not available (Maia et al., 2016). Platforms such as Opflux (Rocha et al., 2010), and CAMEO (Cardoso et al., 2018) incorporate metaheuristic (OptGene (Patil et al., 2005) family methods) and deterministic methods (OptKnock) (Burgard et al., 2003), but are currently restricted to the use of GSMM containing only metabolic information. The recently developed python package MEWpy (Pereira et al., 2021), offers a practical interface to several optimization heuristics allowing to model and optimize microbial production on GSMMs with metabolic, transcriptional, and translational information.

1.4 Microbial communities

Microbial communities are immensely abundant throughout all environments and their composition is equally diverse. Microorganisms within their community share not only the same ecosystem but also the same resources. From extreme environments to host symbionts, these communities play pivotal roles in a multitude of processes and have unique biological properties

maintaining Earth's biosphere and contributing to plant and animal physiology and health (Fierer, 2017; Gilbert et al., 2018; Sunagawa et al., 2015). Thus, the possibility of controlling and engineering natural and synthetic microbial communities would be of huge interest to the environmental, health, and industrial processes (García-Jiménez et al., 2021).

1.4.1 Historical Perspective

The first attempt to "measure the invisible" microorganisms was made by Antoine Van Leeuwenhoek in 1676, using an indigenous microscope to observe the oral microbiome (Hamarneh, 1960). But only in the late 1800s, Robert Koch was able to isolate and cultivate microorganisms in a solid phase. His studies not only helped to understand microbial physiology but also to establish causality between microorganisms and disease (Blevins & Bronze, 2010). Since then, and until the emergence of the microbial ecology concept, based on Winogradsky's works in the 1920s and 1930s to foment the growth of symbiotic microbes (Dworkin & Gutnick, 2012), studies were focused on pure cultures.

In fact, the study of microorganisms and their environmental roles within communities, initiated by Winogradsky, revolutionized microbiology and these culture-based approaches were widely applied to microbial communities from all environments (Ackert, 2013). Identification of microorganisms was the next step in microbiological history, and indeed a vast number of methods were developed over time, such as the gram-staining technique, the agar selective and differential media, and biochemical tests. However, these culture-dependent methods besides being time-consuming and labor-intensive are also inadequate for identifying phenotypically similar species (Buszewski et al., 2017).

Perhaps the breakthroughs in these almost 400 years of history (Figure 1.2) were the works of Carl Woese, Fred Sanger, and Kary Mullis. The emergence of 16s rRNA subunits as molecular taxonomy markers (Woese et al., 1977) and the development of automated DNA sequencing methods (Sanger et al., 1977) and techniques, such as polymerase chain reaction (PCR) (Mullis et al., 1986) allowed not only the identification and classification of microorganisms but also to unveiled microbial communities unculturable species.

Supported by the development of these techniques, metagenomics as a research field surfaced at the beginning of the 1990s, with the first description of a method for the identification of species in a natural microbial community using 16s rRNA libraries (Ward et al., 1990). However, the term

"Metagenomics" was only proposed later that decade, disclosing all the industrial potential of accessing genomes of unculturable organisms (Handelsman et al., 1998). Although the amount of data that was generated by these techniques and instruments was already vast, limited throughput and the high costs of sequencing were still barriers identified by the Human Genome Project development (Goodwin et al., 2016). Next-generation sequencing, presented in the mid-2000s, is a truly high-throughput sequencing platform, greatly reducing the necessary reaction volume while dramatically extending the number of sequencing reactions, and reducing the time, costs, and complexity required to sequence large amounts of DNA (Schuster, 2007).

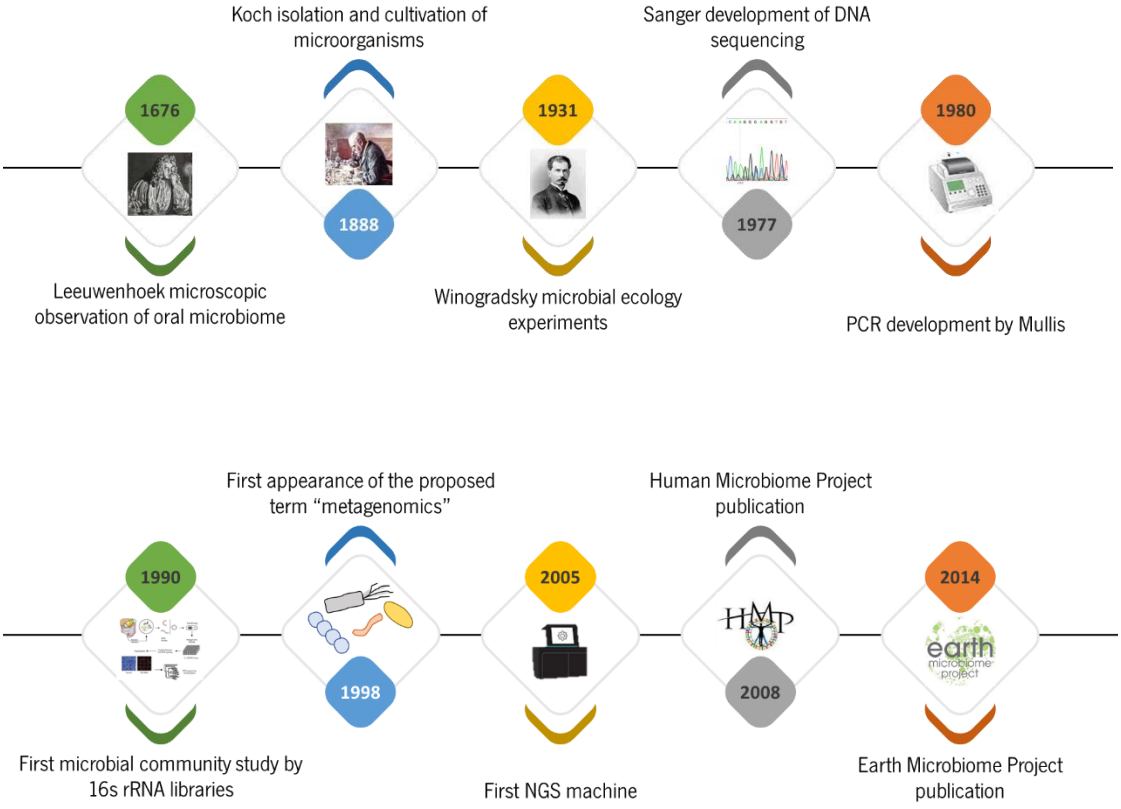


Figure 1.2. Metagenomics timeline and milestones. Timeline showing advances in microbial communities' studies. Adapted from (Escobar-Zepeda et al., 2015)

These new achievements contributed to the exponential growth in the size of sequencing repositories and launched metagenomics as a new scientific field. Since then, a global effort has been

made to catalog the uncultured microbial diversity through the Global Ocean Sampling Expedition (Rusch et al., 2007), the Human Microbiome Project (Turnbaugh et al., 2007), and more recently the Earth Microbiome Project (Gilbert et al., 2014). Metagenomics has reshaped our view of the tree of life (Hug et al., 2016; Parks et al., 2018), led to the identification of deeply rooted and metabolically diverse organisms (Jaffe et al., 2020; Wilson et al., 2014), and leveraged the characterization of new biosynthetic pathways and products (Crits-Christoph et al., 2018; Freeman et al., 2017).

1.4.2 Microbial Community Organisms' Identification

One of the primary objectives of studying microbial communities is determining their composition at the species or strain level, ideally with quantitative information on occurrences (Frioux et al., 2020). The emergence of high-throughput sequencing technologies and new computational pipelines, combined with shotgun metagenomics methods allowed exploring genomic information of all organisms in a microbial community, not only the culturable ones (Zorrilla et al., 2021). In fact, metagenomics has enabled large-scale investigation of complex microbiomes and unveiled species that carry out complete nitrification of ammonia (Daims et al., 2015), or the widespread presence of antibiotic genes in commensal gut bacteria (Donia et al., 2014), for instance. However, the process of identifying microorganisms has some limitations due to the yet large amount of unculturable organisms without a reference genome available (Zorrilla et al., 2021).

The identification of organisms and extraction of microbial abundances from the raw data is usually a complex and multiple-step procedure. The process can be summarized in four steps: (i) Experimental pipeline; (ii) Pre-processing analysis; (iii) Sequence analysis and (iv) Post-processing analysis (Quince et al., 2017). For each of these steps, numerous experimental and computational approaches are available, each of which with limitations and challenges. A short description of each step is available below.

1.4.2.1 Experimental pipeline

The experimental pipeline consists of collecting, processing, and sequencing the metagenome samples. The importance of this step is often underestimated in metagenomics. Sample collection and preservation protocols are extremely important steps once any miscalculation can affect both the quality and the accuracy of metagenomics data. Moreover, careful preliminary work is often required to ensure

that the best collection and storage methods are being applied to the sample as different samples may have different optimal methods.

For metagenomic DNA extraction, the key objectives are to collect sufficient biomass and minimize the contamination of samples. A variety of DNA enrichment protocols are currently available (Quince et al., 2017). However, these procedures can introduce bias into sequence data (Probst et al., 2015). DNA extraction methods should be used according to the microbial diversity in the sample (easy and hard to lyse microbes) (Yuan et al., 2012), the possibility of DNA loss during vigorous extraction techniques (Kennedy et al., 2014) and contamination (essentially on low-biomass samples) (Tanner et al., 1998).

Library preparation and sequencing methods are essentially selected considering the availability of materials and services, cost, ease of automation, and DNA sample quantification. However, it is still expensive to sequence and analyze large numbers of metagenomes without access to sequencing facilities (Quince et al., 2017). In shotgun metagenomics studies, the Illumina platform is the most used due to its wide availability, very high outputs, and high accuracy (Slatko et al., 2018).

1.4.2.2 Pre-processing analysis

Metagenome *de novo* assembly is a crucial analytical step, in which ideally full microorganism genomes are formed from shorter reads. However, contigs generated from the *de novo* assembly are usually very fragmented and are rarely longer than a few kilobase pairs (Deng et al., 2021). The first step in this process is to identify and remove low-quality sequences and contaminants, running a variety of computational tools for quality control, such as FastQC (Andrews, 2010). After quality control, the reads can either pass directly to taxonomic identifiers (see section 1.4.2.3) or be assembled into contigs, as described hereafter. The most popular metagenome assembly approach method is the Bruijn graph approach (Pevzner et. al, 2001), in which each read is broken into *k*mers of the same length (k), and a graph is formed using the *k*mers as vertices. Contigs are built by walking the graph from edges to nodes (Aylin et al., 2020).

Metagenome assembly presents unique challenges. Unlike single genomes, sequence coverage along different genomes is usually different as different species are present in communities in different abundancies. Low-abundance genomes can be ignored when the overall sequencing depth for graph formation is big. To assist in the recovery of low-abundant genomes, sequencing depth can be shortened,

though accurate contigs of high-abundant genomes can be difficult to obtain (Quince et al., 2017). Metagenome-specific assembly tools, such as Meta-IDBA (Peng et al., 2011), MetaSpades (Nurk et al., 2017), and MEGAHIT (Li et al., 2015), have been developed to overcome these challenges. MEGAHIT can be considered one of the top three metagenomics assembly tools, according to the Critical Assessment of Metagenomic Interpretation (CAMI) challenge (Sczyrba et al., 2017). Whichever assembly tool is used, the result will be potentially millions of contigs that need to be linked to the respective genomes.

1.4.2.3 Sequence analysis

Sequence analysis consists of trying to profile taxonomic, functional, and genomic features of the microbiome. In this primary analysis of metagenomic data, two approaches can be used: assembly-based analysis (binning) and read-based analysis. Both have strengths and weaknesses, and the success of either approach depends on the microbial community composition and complexity. Whenever possible, recommendations for using both approaches for sequence analysis have been performed (Quince et al., 2017), as these complete and validate each other.

Assembly-based metagenomic profiling

The first step of sequence analysis is to group the contigs obtained in the last step into species. However, metagenome assemblies are highly fragmented, making it impossible to know which contigs belong to a genome, or even how many genomes are present. Binning aims to group contigs into species. For that, supervised and unsupervised methods may be used. Whereas the first method assigns contigs to taxonomic classes through sequenced genome databases, the second seeks natural groups according to the data or statistical properties, using clustering. Both use a metric to define similarity between contigs allocated to a bin, and an algorithm to convert similarities into assignments (Quince et al., 2017). Unlike supervised methods that are based on contig homology to known genomes, unsupervised methods do not require prior knowledge about genomes in a sample. In fact, most microbial species have not been sequenced yet, making it impossible to map most fragments to reference genomes. Unsupervised methods rely on features, such as GC content or di- and higher-order nucleotide frequencies, that usually vary between taxonomic lineages. Clustering can then be used, and the sequence of new candidate phyla

be unveiled (Hug et al., 2016). Several recent methods, such as CONCOCT (Alneberg et al., 2014), use a combination of these two features.

A validation step should always be performed after binning, as each bin can contain more than a taxonomic group, which is not ideal. Completeness, as well as contamination, are two of the most tested features. For instance, when marker genes are missing, the genome is probably incomplete and when marker genes appear multiple times, the genome is eventually contaminated (Breitwieser et al., 2019).

Assembly-free metagenomic profiling

Besides identifying microbial species in a metagenome, taxonomic profiling of metagenomes estimates their abundance. Based on simply mapping reads to reference genomes and environmental-specific assemblies, published in databases (Nielsen et al., 2014), this method can mitigate assembly problems, speed up computational time, and enable profiling low abundance organisms that cannot be assembled *de novo*. The main limitation of this approach is the profiling of uncharacterized organisms essentially in samples from soil and ocean environments, which are hampered by a lack of representative reference genomes. In these cases, the use of assembly methods is generally advisable (Breitwieser et al., 2019). However, the number of available reference genomes is increasing every day, even for difficult-to-grow species, assisted by the development of new cultivation methods, and single-cell sequencing methods (Rinke et al., 2013). Also, the vast number of metagenomics studies of human gut samples, led to the extensive availability of reference genomes for this environment, allowing to use of assembly-free metagenomic profiling as an efficient and successful strategy, even for low-abundant microbes (Nelson et al., 2010).

1.4.2.4 Post-processing analysis

Post-processing analysis consists of using statistical tools to interpret the outputs of the methods described above. These outputs comprise data matrices of samples and/or microbial features, such as species, taxa, genes, and pathways. Most of the statistical approaches are not specific to metagenomic studies, including unsupervised and supervised methods. Unsupervised methods are used to infer ecological relationships within the community (Faust et al., 2012), by applying clustering, correlation, and visualization techniques, such as heat maps and ordinations (e.g. principal component analysis (PCA)).

Supervised methods include statistical methods, such as multivariate analysis of variance (ANOVA) or machine learning classifiers (Pasolli et al., 2016). Multivariate analysis is usually applied for direct hypothesis testing of differences between groups, whereas machine learning is used to train models that label groups of samples. These methods treat the community as a whole. For more precise testing of specific taxa or functional genes, the complexity, and the huge amount of data of metagenomic datasets may be a burden. To address that, methods of correction for multiple comparisons (White et al., 2009) or effect size estimation (Segata et al., 2011) are commonly used.

1.4.3 Microbial interactions

The microbial community's structure and function rely on complex microbe-microbe or microbe-environment interactions. The increasing attention that microbial communities have had in the last years is potentiated by the need to identify, understand, and enhance these interactions (Faust, 2018). In fact, various studies have been published revealing the impact of these interactions on our health, e.g., via our microbiome (Glowacki et al., 2020; Shi, 2019), our planet, e.g., via biogeochemical cycles (Nazaries et al., 2013), and even our food (Bokulich et al., 2016). Although a huge effort to understand metabolic interactions in these communities is undergoing, a great deal is still unknown (Zelezniak et al., 2015).

Microbial interactions can be classified as positive (cooperative metabolite exchange), negative (competition for resources), or neutral (no effect on the interacting species) (Figure 1.3). The possible combinations of positive, negative, and neutral outcomes for two interaction partners allow the classification of various interaction types (Faust et al., 2012).

Mutualism is characterized by a win-win relationship, in the case of bacterial cooperation to build a biofilm or in cases of cross-feeding (syntrophy). Commensalism is labeled as a win-neutral relationship in which commensals cross-feed on compounds that are produced by other community members. Parasitism is a classical host-parasite situation, where one wins on the loss of the other (e.g., bacteria-bacteriophage interaction). When the production of one species by-product influences the environment and harms other species (e.g. lowering the pH of the environment), characterized by a loss-neutral relationship, it is considered amensalism. In competition, a loss-loss relationship is established, where the growth of both organisms is affected in the presence of each other.

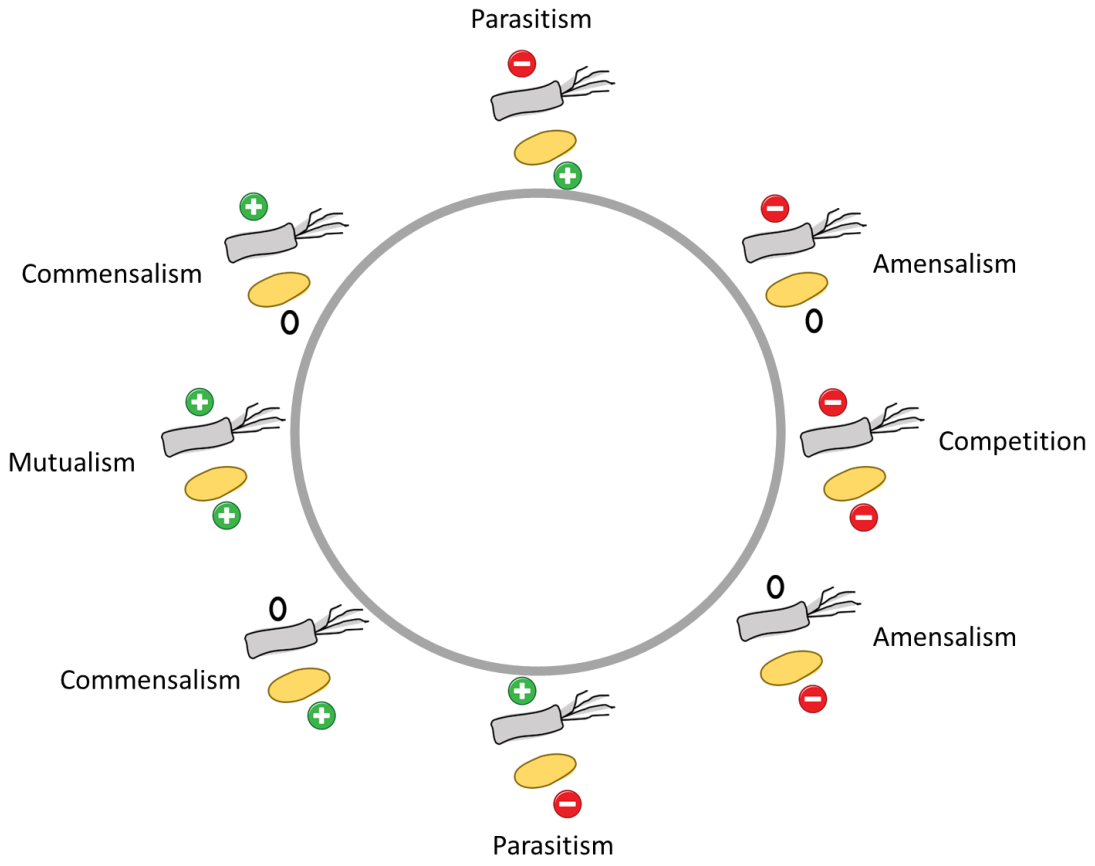


Figure 1.3. Summary of possible microbe-microbe interactions. For each interaction partner, there are three possible outcomes: positive (+), negative (–), and neutral (0). Adapted from (Faust et al., 2012).

Identifying and classifying these interactions is a difficult task, as metabolites cannot be easily attributed to a particular species or even to the environment. Moreover, species may excrete a large number of metabolites, interfering with the disclosure of metabolite excretion (Zelezniak et al., 2015). Tools have been developed to predict these microbe-microbe and host-microbe interactions, such as SMETANA (Zelezniak et al., 2015) and MICOM (Diener et al., 2020), where GSMMs are used as input and with demonstrated applicability (Machado et al., 2021).

1.4.4 Extremophiles

Over the last few years, the scientific community has been intrigued by finding life in environments that once were described as unhabitable (Rothschild et al., 2001). Metagenomic studies on extreme environments such as Antarctic ecosystems, saline lakes, geothermal springs, or deep-sea hydrothermal

vents, revealed a large number of organisms known as extremophiles (Nicolaus et al., 2010). Extremophiles are capable of sustaining biological life under extreme environmental stressors and are known for their remarkable diversity of morphology, biochemistry, genomics, and biosynthesis of adaptive compounds (Durvasula et al., 2018).

Table 1.2. Types of extremophiles based on their habitat. Adapted from (Durvasula et al., 2018).

Environmental Variable	Type	Characteristic	Example	Reference
Temperature	Hyperthermophile	Growth > 80°C Upper limit 130°C	*Geogemma barossii*	(Kashefi et al., 2003)
	Thermophile	Growth > 45°C Upper limit 80°C	*Sulfurihydrogenibium azorense*	(Aguiar et al., 2004)
	Mesophile	Growth > 20°C Upper limit 45°C	*Escherichia coli*	(Rudolph et al., 2010)
	Psychrophile	Growth <20°C	*Polaribacter gangjinensis*	(Lee et al., 2011)
Radiation		Gamma > 15 kGy UV > 600 J m^{-2}	*Deinococcus radiodurans*	(Bauermeister et al., 2011)
Pressure	Piezophile	Growth > 10 MPa Upper limit 130 MPa	*Thermococcus piezophilus.*	(Dalmasso et al., 2016)
Salinity	Halophile	Growth > 0.2 M salt Upper limit 5.5 M	*Haloterrigena thermotolerans*	(Montalvo-Rodríguez et al., 2000)
pH	Alkaliphile	Growth > pH 9 Upper limit pH 13	*Natronococcus occultus*	(Jones et al., 1998)
	Acidophile	Growth < pH 5	*Picrophilus oshimae*	(Schleper et al., 1995)

The tree main branches of the tree of life (Bacteria, Archaea, and Eukarya) are represented in all of Earth's extreme environments, showing the remarkable and wide adaptability of these organisms. Known growth conditions' limits range from negative 12°C to positive 130°C for temperature, pH around 0 to 13, pressure of over 100Mpa, beyond saturation conditions of NaCl and KCl, and high levels of ultraviolet and gamma radiation (Table 1.2). Harsh environmental conditions urged microorganisms to

evolve, including structural and biochemical adaptations, which allowed growth and survival under conditions that imposed a lack of nutrients and energy (Ando et al., 2021). Often these adaptations are a consequence of key changes in the organism's enzymes amino acid sequences, which are translated into variations in the structure, flexibility, charge, and/or hydrophobicity (Sarmiento et al., 2015). Therefore, extremozymes can offer exciting industrial opportunities when compared with organic synthesis and even mesophilic biocatalysts, allying better chemical precision, sustainability, and cost-effectiveness, to harsh enzymatic conditions capable of withstanding industrial demands (Van den Burg, 2003). Indeed, extremophiles' enzymes that have been isolated and functionally characterized, led to the identification of the thermostable DNA polymerase used in the polymerase chain reaction (PCR) (Ishino et al., 2014) and enzymes used in the biofuel (Barnard et al., 2010), food, paper and cosmetic industries (Sarmiento et al., 2015), for instance.

Until now, hyperthermophilic and thermophilic extremophiles/extremozymes have attracted the most attention. Prokaryotic diversity in hydrothermal ecosystems has been extensively studied (Ando et al., 2021). Moreover, the temperature endured by thermophilic enzymes, besides increasing reaction rates during high-temperature processing, reduces microbial contamination, lowers substrate viscosity, and increases the solubility of many polymeric substrates (Van den Burg, 2003).

1.4.5 Genome-scale Metabolic Modelling of Microbial Communities

The increased interest in studying microbial communities due to their key roles in human and environmental health was highly supported by the recent advances in high-throughput multi-omic technologies (metagenomics, metatranscriptomics, metaproteomics) (Zaramela et al., 2021). However, the use of these techniques to understand the individual role of microbes and their interactions, with each other and the community, is still a challenge. The use of GSMMs through computational and mathematical modeling approaches is nowadays a widely used platform to generate testable hypotheses on microbial community behavior at the taxon and community levels (Colarusso et al., 2021). Here a brief description of the current existent methods and tools for reconstruction, simulation, and optimization of microbial communities using GSMMs will be performed, including an assessment and a discussion of current limitations and challenges.

1.4.5.1 Microbial Community Genome-Scale Metabolic Model Reconstruction

Microbial community GSMMs' reconstruction is based on the reconstruction process of a single species. Two main approaches are used for these reconstructions: the mixed-bag and the compartmentalized network approaches (Faria et al., 2017).

The mixed-bag network approach considers the microbial community a single supra-organism (Abubucker et al., 2012), constructing a model composed of one cytosolic and one extracellular compartment, ignoring species boundaries. In this case, the resulting model will be analogous to a single-species prokaryotic network. Species-level resolution is ignored, and these networks are usually used to analyze community—environment interactions (Henry et al., 2016). The reconstruction of a mixed-bag network requires either the annotated genome sequences of all species present in the community or the whole deep metagenome sequence with annotated reads. From this stage, the network reconstruction is performed similarly to a single species network (see section 1.3.1). In the final model, redundant reactions (reactions included in at least two of the community species) are ignored.

The compartmentalized network approach considers the different species of the community in distinct compartments. For each organism in the community, a single species metabolic model is constructed based on its genome sequence. The community model will be composed of as many compartments as the number of organisms in the community, plus an extracellular pool to allow metabolite exchange. Species-level resolution is obtained and therefore cross-species metabolic interactions can be predicted and optimized. Metabolic capabilities can be associated with the respective species within the community, which will allow microbial community design (Faust, 2018). However, in this approach, microbial community composition in terms of species must be known. Hence, reconstructing models from metagenomic samples, with undefined composition, demands an initial step of species identification from the metagenome(see section 1.4.2).

The selection of an approach to reconstruct a microbial community network will depend on data and resource availability, as well as on the purpose. The limitations and challenges of each approach will be detailed in section 1.4.5.4.

1.4.5.2 Simulation of Microbial Communities using Genome-Scale Metabolic Models

The application of GSMMs to the simulation of microbial communities was pioneered by Stolyar et al, who have applied them in the simulation of a two-species methanogenic community (Stolyar et al., 2007). Several other simulation methods have since been published, but a comprehensive and systematic analysis will be detailed in Chapter 4 of this thesis. Here, pioneering methods for microbial community simulation will be discussed. As demonstrated before, multiple GSMMs can be easily merged into a compartmentalized multi-species GSMM, where organisms are allowed to exchange compounds through a common extracellular pool. However, formulating a biologically and ecologically meaningful objective function for a microbial community model is not trivial, and is one of the main differentiating features of different simulation methods. Two other main differentiating features are the inclusion of temporal and spatial components.

The simplest class of methods is used for the simulation of steady-state flux distributions in microbial communities, allowing the prediction of individual growth rates and cross-feeding interactions in community equilibrium. OptCom (Zomorrodi et al., 2012) implements a bi-level objective function where the inner objective is to maximize the growth rate of individual species, and the outer objective is to maximize the total community biomass. The underlying assumption is that the community will optimize the management of available resources while still considering the individualistic objectives of its members. Other methods, such as cFBA (Khandelwal et al., 2013), SteadyCom (Chan et al., 2017), and MICOM (Diener et al., 2020), implement different objective functions following similar assumptions. One advantage of steady-state methods is that they do not require parameterization. However, they are characterized by a large uncertainty in the space of optimal solutions.

The second class of methods considers the temporal component when performing time-course simulations. These are generally implemented as multi-species extensions of dynamic FBA (dFBA) (Zhuang et al., 2011), where time is divided into discrete steps and an FBA simulation is performed for each species at each time step. The d-OptCom (Zomorrodi et al., 2014) method combines dFBA with OptCom to implement dynamic multi-objective optimization. Recently, a faster implementation of multi-species dFBA, μbialSim (Popp et al., 2020), was able to scale up to communities with hundreds of species. These methods account for the temporal variation in metabolite concentrations and species abundance, which can play a role in community assembly, due to sequential species colonization, and allow simulating the response to changes in initial conditions and other environmental perturbations.

However, every organism requires the characterization of the substrate's uptake kinetics for all metabolites that are consumed (either from the growth medium or through cross-feeding). The impractical nature of such a massive *in vitro* characterization results in the adoption of default parameter values for all species and compounds, which limits the predictive ability of these methods.

The third class of methods adds the spatial component as well, resulting in spatiotemporal simulations. These methods are especially relevant in situations of heterogeneous or spatially segregated environments, such as biofilm colonies or the tract along the human gut, where nutrient diffusion and access to resources play a role in community assembly and the potential for cross-feeding interactions. The implementation of spatiotemporal methods mostly differs in the way the spatial component is modeled. COMETS (Harcombe et al., 2014) extends dFBA with a 2D-grid where each square represents a population of cells at a given point in time. BacArena (Bauer et al., 2017) also uses a 2D-grid structure but follows an individual-based modeling (IBM) approach, where each cell represents one individual, allowing for a more fine-grained resolution of inter and intra-species interactions. IndiMeSH (Borer et al., 2019) also uses an IBM approach, but the spatial structure is based on angular pore networks, which allows the creation of more complex structured environments. In addition to the parameters on uptake kinetics mentioned before, these methods also require the specification of diffusion rates for cells and metabolites.

These three main classes of methods represent increasing levels of simulation detail but also require more experimental data for model setup and, naturally, more computational power. Therefore, the selection of the most suitable kind of method depends on the complexity of the microbial community under study, as well as the amount of available experimental data. Regardless of this choice, GSMM-based simulation allows for a mechanistic interpretation of observed community phenotypes and of the competitive and cooperative behaviors contained therein. Multi-omics data (transcriptomics, proteomics, metabolomics) can be used to constrain the solution space of metabolic models further highlighting the metabolic pathways active in each condition, and analyzing how different kinds of perturbations will elicit a coordinated response comprising the molecular, cellular, and population levels.

1.4.5.3 Microbial Community Genome-Scale Metabolic Model Optimization

The biological properties of microbial communities are closely related to the organisms present within them and resource availability (Johns et al., 2016). If the community composition (species or

medium) is disturbed, its functional capabilities can be affected. Such findings challenged the scientific community to manipulate microbiomes to improve crop productivity (O'Connell et al., 1996), to deal with contaminated groundwater (Löffler et al., 2006), or even to recover valuable resources from wastewater (McCarty et al., 2011), showing the huge potential of engineering microbial communities. Two main approaches have been currently used for engineering microbial communities: the top-down approach (Widder et al., 2016), which aims to control metabolic processes for stabilizing complex and natural communities; and the bottom-up approach (Lindemann et al., 2016), which aims at designing defined, synthetic communities with desired functionalities (Zerfaß et al., 2018). When combined, these approaches can offer complementary strategies to design defined, synthetic communities used to impact and engineer the behavior of complex communities (Lawson et al., 2019).

Given the success of computational strain optimization methods to rationally design single organisms using GSMMs, already addressed in section 1.3.3, it is reasonable to assume that the same strategies may lead to success using microbial communities. Microbes rarely live in isolation, thus engineering microbial communities is a strategy closer to natural and physiologic behavior, than for isolated organisms. Moreover, the engineered product will be a cooperative effort from a structured community able to carry out complex interactions and synthesize complex molecules (Bosi et al., 2017; García-Jiménez et al., 2021).

Until now, the number of studies that aim at developing algorithms for designing microbial communities using GSMMs has been low. Most of the tools developed use strategies to optimize medium composition and/or best community configuration for a given objective. Among them, the SIM algorithm (Klitgord et al., 2010), developed over the MATLAB® platform, uses a mixed-integer linear programming approach, to optimize a medium composition that supports the growth of multispecies co-cultures in a specific condition. An initial minimal set of potential exchanged metabolites that allows a positive growth rate of both organisms is identified. The core function of the SIM algorithm identifies all potential metabolites that can be switched in such a medium and the initially available source for a given element (carbon, nitrogen, sulfur, or phosphate). Hence, the chemical formula of each metabolite present in the GSMMs must be available. The algorithm is based on the identification of a symbiosis-inducing media, assuming that each species will secrete what the other needs.

Similarly, FLYCOP (García-Jiménez et al., 2018), supports the design and engineering of microbial communities by selecting a consortium configuration that optimizes a given goal (community growth rate,

stability, medium composition, etc). Hence, FLYCOP selects and evaluates candidate consortium configurations, through an iterated local search algorithm (SMAC (Hutter et al., 2011)). For strain design, individual strain engineering must be performed beforehand, optimizing only the conditions for fermentation.

Using a different approach, DOLMN (Thommes et al., 2019) explores the space of feasible single-strain or multistrain metabolic networks, by systematically limiting the number of intracellular and transport reactions (performing reaction knockouts) in each metabolic model, using a MILP formulation. However, optimal solutions result in deleting a large number of reactions, which is not applicable for practical purposes.

The increasing interest in designing microbial communities demands that model-guided microbial community engineering should trend toward the development of technologies capable of predicting potential genetic modifications at the community level, like individual-level design platforms, such as OptKnock. These new implementations of microbial community design methods may focus not only on optimizing exchange reactions (species cross-feeding interactions, medium composition) but also on untargeted and targeted (directed to specific species in the community) genes/reactions, as well as identifying the best community composition for a given objective, using top-down and bottom-up approaches.

1.4.5.4 Genome-Scale Metabolic Modelling of Microbial Communities - Limitations and Challenges

Although microbial community modeling is widely established, there are known limitations and challenges. A good GSMM must be manually curated and validated with experimental data to accurately predict phenotypic behavior. The process of reconstructing a community GSMM can be similar to that of a single metabolic model. In this approach, one would start by annotating all genes in a metagenomics sample (mixed bag approach). Alternatively, one could initially perform the identification of all organisms present in the community and then reconstruct the corresponding GSMMs (compartmentalized approach) (Frioux et al., 2020). In the latter approach, depending on the number of organisms and the availability of curated GSMMs in databases, such as the BiGG knowledgebase (King et al., 2016a), the time it takes to process, collect, and eventually reconstruct GSMMs can increase exponentially.

In the last years, pipelines that semi-automatically reconstruct prokaryotic organisms' GSMMs in a matter of minutes, such as KBase (Arkin et al., 2018), the python package CarveMe (Machado et al., 2018) or the AGORA knowledgebase which is focused in studying the human microbiota or host-pathogen interactions (Almut Heinken et al., 2020), have been published. These models have already shown their reliability in predicting growth, response to nutrients, and gene essentiality in single organisms and microbial communities (Chng et al., 2020; Machado et al., 2021; Nayfach et al., 2020). However, when the objective is the rational design of microbial communities, these models should be used with caution, as they often have some inaccuracies (Lieven et al., 2020), requiring further manual curation.

To accurately study a microbial community, it is essential to correctly identify the available species. Besides wet-lab limitations, the lack of a complete reference genome and missing functional annotation of microbial genes in current databases (Bharti et al., 2021; Quince et al., 2017) impair the reconstruction of community GSMM.

When reconstructing microbial community's GSMMs, the struggle starts with how to compare simulation results. Several methods for simulating microbial communities using GSMMs have been published lately, although the full potential of those methods has not been achieved. The current experimental techniques are designed for individual organisms, with still limited applicability to microbial communities, as not enough experimental data on abiotic and biotic interactions, and perturbations is yet available (Machado et al., 2018).

Although one of the main objectives of studying microbial communities is the possibility of controlling and engineering natural and synthetic microbial communities, few studies are available in terms of optimization tools capable of predicting potential genetic modifications at the community level (García-Jiménez et al., 2021). The challenge here is the development of sophisticated and integrative platforms that support different levels of community optimization (e.g. medium, interactions, community composition, targeted organism design within the community) to unveil the full potential of microbial communities (Eng et al., 2019).

Chapter 2

Metagenomic study of thermophilic and hyperthermophilic environments from Azores

"You are never alone.
You are externally connected with everyone"
Amit Ray

Five hydrothermal samples from hydrothermal vents at São Miguel, Azores, were analyzed to determine prokaryotic community composition. Taxonomic profiling of all samples was performed using assembly-based and read-based analysis showing combined results specifically for the predicted most abundant organisms in each sample. Samples showed to be abundant in members of the *Aquificales* and *Crenarchaeota* orders, in specific, *Sulfurihydrogenibium azorense* Az-Fu1, *Desulfurococcus amylolyticus* DSM 16532, *Pyrobaculum islandicum* DSM 4184/ *Pyrobaculum aerophilum* str. IM2 and *Thermofilum adornatus* 1505.

S. azorense Az-Fu1 was first isolated at São Miguel, Azores but in a different location (Furnas). Thus, the metabolic capabilities of *S. azorense* Az-Fu1 in isolation and within its microbial community, hence a systems biology approach, is worthy to be investigated.

Sophia Santos[1], Ana Alão Freitas[2], Ricardo B. Leite[3], Duarte N. Toubarro[4], José B. Leal[5], Nelson Simões[4], Oscar Dias[1], Isabel Rocha[2]

[1] Centre of Biological Engineering of University of Minho, Braga, Portugal

[2] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa (ITQB-NOVA),

[3] Instituto Gulbenkian de Ciência, Oeiras, Portugal

[4] Centro de Biotecnologia dos Açores, Departamento de Biologia, Universidade dos Açores, Ponta Delgada, Portugal

[5] Ophiomics, Lisboa, Portugal

Autors' contributions

Sophia Santos conceived the study, carried out the read-based taxonomic profiling, analyzed the results, and drafted this chapter. Duarte N. Toubarro carried out the sample collection and DNA extraction. Ana Alão Freitas and Ricardo B. Leite carried out the metagenomics and bioinformatics pipeline. Nelson Simões and José B. Pereira Leal participated in the design of the study. Oscar Dias and Isabel Rocha conceived the study participated in its design and coordination and helped to draft this chapter.

2.1 Introduction

From extreme environments to host-symbionts, microbial communities play pivotal roles in a multitude of processes and have unique biological properties maintaining Earth's biosphere, contributing to plant and animal physiology, as well as human health (Fierer, 2017; Gilbert et al., 2018; Sunagawa et al., 2015). The appearance of Metagenomics as a field of study has transformed the way these microbial communities are perceived, being nowadays possible to control and engineer natural and synthetic microbial communities, with applications on environmental, health, and industrial processes (García-Jiménez et al., 2021).

The study of microbial communities in environments that once were described as unhabitable (Rothschild et al., 2001) – extremophiles – which are capable of sustaining biological life under one or more environmental stressors and are known for their remarkable diversity of morphology, biochemistry, genomics, and biosynthesis of many adaptive compounds (Durvasula et al., 2018) has nowadays gained huge importance. In harsh environmental conditions, microorganisms evolved several structural and biochemical adaptations that allowed them to grow and survive under conditions that impose, among others, a lack of nutrients and energy (Ando et al., 2021; Sarmiento et al., 2015). Therefore, the study of extreme microorganisms through their outstanding metabolic capabilities or their extremozymes can offer exciting industrial opportunities (Van den Burg, 2003).

When studying microbiomes via metagenomics, one of the primary objectives is the identification of the microorganisms, ideally with quantitative information on their occurrences (Frioux et al., 2020). Despite the existence of multiple algorithms, the identification of organisms and extraction of microbial abundances from the raw data is still a complex and challenging procedure. However, microbial communities in high-temperature environments are generally less diverse (Inskeep et al., 2013) making hydrothermal habitats an ideal model system to study principles of community structure and function (Sahm et al., 2013). In fact, prokaryotic diversity in hydrothermal ecosystems has been extensively studied at hydrothermal sites, such as Yellowstone National Park (Inskeep et al., 2013; Jay et al., 2016), repeatedly revealing characteristic taxonomic groups, like the main presence of *Aquificales* and *Crenarcheota* members (Strazzulli et al., 2017). Being widespread in nature, in marine and terrestrial geothermal systems, members of both phyla are essentially thermophilic and hyperthermophilic microorganisms, almost exclusively obligate or facultative autotrophs (Hedlund et al., 2015) and are also

capable of using a wide range of energy sources, such as H_2, elemental sulfur, and thiosulfate (Takacs-Vesbach et al., 2013, Yamanaka, 2008).

This work aims to characterize samples in the hydrothermal sites at São Miguel, Azores. Azores archipelago, composed of nine isolated islands almost exclusively of volcanic origin, is geographically situated in the Atlantic Ocean. In the Azores, namely at Furnas Valley, valuable research has been made on the search for thermophilic microorganisms and thermostable enzymes with biotechnological applications (Albuquerque et al., 2012; Albuquerque et al., 2010; França et al., 2006; Riessen et al., 2001). However, little data is available on other Azorean volcanic sites and overall diversity (Hamamura et al., 2013).

Applying metagenomic approaches, this work intends to provide an in-depth study of bacterial and archaeal diversity as a quantitative basis for understanding individual metabolic capabilities and metabolic interactions within the prokaryotic community present in extreme environments.

2.2 Methods

2.2.1 Sample collection

Samples were collected inside and at different hydrothermal springs sites, at São Miguel, Azores. A description of the different sampling sites is provided in Table 2.1.

Table 2.1 Physical and chemical characteristics of Azorean hydrothermal springs sample sites.

Location	Sample	Temperature (ºC)	pH
Caldeira Velha	CV	93.7	2.23
Nascente da Ponte	NP	95.8	6.88
Esguicho de Maio	ESG	98	7.29
Fumarola Caldeiras da Ribeira Grande	FCRG	49.4	2.2
Piscina Caldeiras da Ribeira Grande	PCRG	40	3.53

As for other well-studied hydrothermal sites, São Miguel's hotspots have a huge diversity of thermal features that cover a wide range in pH (2–8) and temperature (40–98°C) sites. Water samples were

collected using glass bottles of a 1-liter beaker with a lid. For each sample temperature and pH were recorded at the sampling site. The samples were placed in a thermal box to prevent the descent of temperature.

2.2.2 DNA extraction

Water samples were passed through Miracloth filtration material (Calbiochem) to remove large debris and then filtered through a 0.22µm Millipore® Sterivex™ filter unit. The membranes were processed at Instituto Gulbenkian de Ciência (IGC) (Oeiras, Portugal). Genomic DNA was extracted from the individual filters using PowerWater Sterivex DNA Isolation Kit (MO BIO, Carlsbad, CA, USA) following the manufacturer's protocol. The amount of the DNA extracted was later quantified using a NanoDrop 1000 spectrophotometer (Thermo-Fisher Scientific, Wilmington, DE, USA) measuring the UV absorption at 260 nm and 280 nm wavelengths. DNA integrity was evaluated by agarose gel electrophoresis.

2.2.3 Bacterial and Archaeal Diversity: 16S rRNA Gene Amplicon Sequencing

To assess the bacterial and archaeal diversity in the São Miguel - Azores reservoirs, the 16S rRNA gene was used as a marker for biodiversity. The extracted environmental DNA was amplified using primers targeting the V3 and V4 hypervariable regions of the 16S rRNA gene and V9 region for the 18S (Caporaso et al., 2012). Whole genome libraries were produced using the Nextera XT DNA Library Preparation Kit according to the manufacturer's instructions. The purified DNA was sequenced with Illumina NextSeq using a 150 bp paired-end DNA library (Illumina, San Diego, CA, USA) to generate at least 240 million reads per sample sequenced. PCR amplification of the 464 bp fragments was performed with the general bacterial primer pair 341F/785R and the general archaeal primer pair 340F–1000R (Klindworth et al., 2013). Ribosomal RNA gene amplicon sequencing was performed by IGC, Gene Expression Facility (Oeiras, Portugal).

2.2.4 Metagenomics and Bioinformatics Pipeline

Based on the features of the obtained reads and the software available, a semi-automatic pipeline (Figure 2.1) was designed and implemented at IGC (by Ana Alão Freitas and Ricardo Leite) to identify and

"make meaning" of the sequenced short reads. Some of the software used was executed using the Linux command line while others were used online directly on the respective websites.

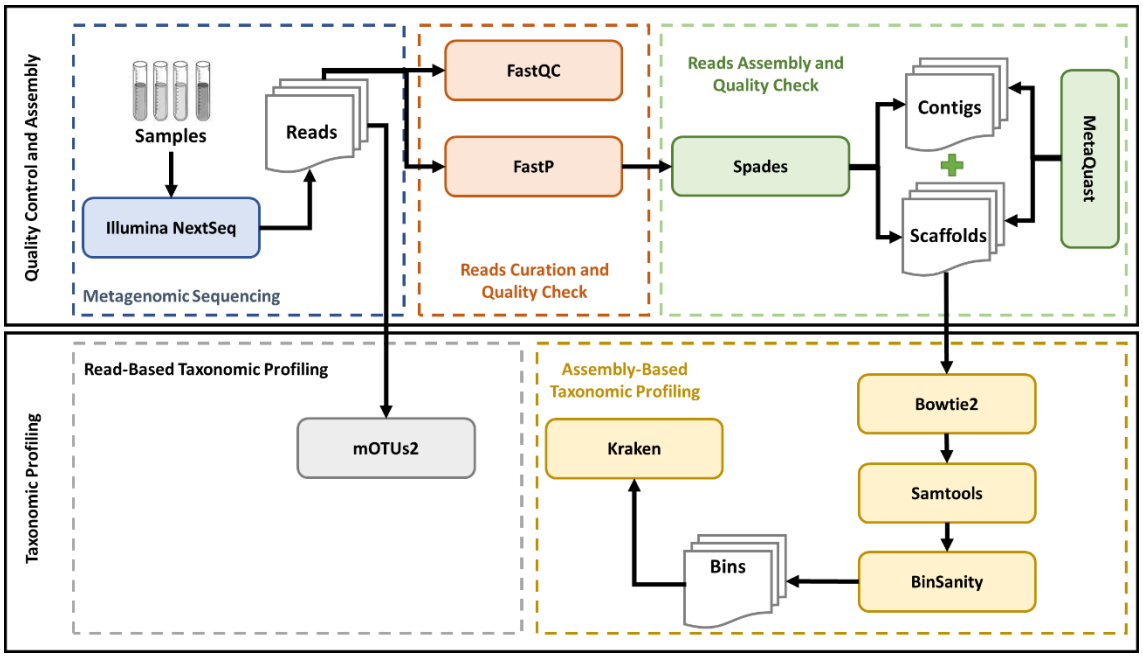


Figure 2.1. Metagenomics and bioinformatics semi-automatic pipeline implemented. Each step of the process is labeled with a different color. Tools are shown in boxes and data is shown as a multi-document flowchart. The read-based taxonomic profiling was exclusively done for this work.

2.2.4.1 Quality Control and Assembly

As illustrated in Figure 2.1, the whole pipeline starts from the short reads obtained from the metagenomic sequencing step. The reads from each sample were submitted to this pipeline separately.

Sequencing resulting FASTQ files with short-reads were submitted to a quality check using FastQC software version v0.11.5 (Andrews, 2010). A pre-processing and filtering process of unqualified bases was performed using the FastP software version v0.23.2 (Chen et al., 2018), using the following parameters: -q 20, -u 40.

Reads were afterward assembled via SPAdes genome assembler v3.14.1 (Bankevich et al., 2012) using the options "–pe -1" for the file with left reads and "– pe -2" for the file with right reads. Values for the k-mer sizes used in the Bruijn graphs application were chosen based on the values advised for the

size of the obtained short-reads (option "-k" 21,33,55,77,99,127). Both contigs and scaffolds quality was evaluated with Quast Version: 5.0.2 option Meta (Mikheenko et al., 2016).

2.2.4.2 Taxonomic profiling

Taxonomic profiling was obtained by performing the assembly-based analysis (binning) and read-based analysis (operational taxonomic units). When analyzing environmental samples, the lack of available representative reference genomes is a drawback, so the use of both approaches for sequence analysis is recommended (Quince et al., 2017).

Assembly-based analysis (binning)

Binning of the pre-processed scaffolds was carried out by the BinSanity v0.3.1 (Graham et al., 2017) algorithm. The pre-processing was performed by Bowtie2 v2.4.4 (Langmead et al., 2012) and SAMtools v1.16.1 (Danecek et al., 2021). Taxonomic assignment was conducted by Kraken 2 (Wood et. al, 2014), using as the default library the Nacional Center for Biotechnology Information's (NCBI) RefSeq database (NCBI Resource Coordinators, 2018).

Operational taxonomic units (OTUs)

OTUs were generated with the mOTUs profiler v.2.0.0 (Milanese et al., 2019) using the following parameters: -l 75; -g 2; and -c. Profiles were filtered to focus on a set of species that were confidently detectable. Specifically, microbial species that did not exceed a maximum relative abundance of 1×10^{-2} (1%) were excluded from further analysis, together with the fraction of unmapped metagenomic reads.

2.3 Results and Discussion

2.3.1 Bacterial and Archaeal diversity

Samples were collected from five sites inside and near hydrothermal springs at São Miguel, Azores. *In situ,* temperatures ranged from 40 to 98°C, whereas pH values varied between 2 and 8 (Table 2.1).

Hot springs are considered low-diversity habitats due to their extreme physicochemical parameters. Regarding temperature, no correlation has yet been found on its effect on prokaryotic diversity showing even contradictory results (Sahm et al., 2013) that also depend on other factors, such as sulfide concentration (Skirnisdottir et al., 2000), pH values (Yim et al., 2006) and prokaryotic organisms with specific metabolism (Miller et al., 2009).

When analyzing all samples collected at São Miguel's hydrothermal springs (Table 2.2), results show that bacterial diversity is essentially affected at high temperatures and neutral pH (ESG sample – 98 °C, pH 7.29), maintaining a big diversity at all other temperatures and pH ranges. However, archaeal diversity seems to be more affected by changes in temperature and pH, showing higher diversity at temperatures between 40 and 50 °C and acidic pH (2 – 3.5) – PCRG and FCRG samples.

Table 2.2. Bacterial and Archaeal 16s rRNA sample detection. Low - +, Medium - ++, and High - +++. Samples CV – Caldeira Velha, NP – Nascente Poente, ESG – Esguicho de Maio, FCRG – Fumarola Caldeira da Ribeira Grande, PCRG – Piscina Caldeira da Ribeira Grande.

Sample	Bacterial 16s rRNA	Archaeal 16s rRNA	DNA (ng/µL)
CV	+++	+	10.6
NP	+++	+	9.7
ESG	+	+	24
FCRG	+++	+++	28.1
PCRG	+++	++	22.7

Nevertheless, a chemical analysis of the sampling sites was not made, which, as stated before, can at some level explain the prokaryotic diversity results obtained.

2.3.2 Metagenome Assembly

The metagenomics and bioinformatics pipeline performed at IGC and described in section 2.2.4, was applied to all five samples (CV, NP, ESG, FCRG, and PCRG). Quality check of reads, contigs, and scaffolds obtained by the FASQC, FastP, and MetaQuast tools are summarized in Supplementary Tables

S1, S2, and S3. The total number of reads, contigs, and scaffolds obtained by the implemented pipeline is presented in Table 2.3.

Table 2.3. Obtained number of reads, contigs, and scaffolds for each sample analyzed. Samples CV – Caldeira Velha, NP – Nascente Poente, ESG – Esguicho de Maio, FCRG – Fumarola Caldeira da Ribeira Grande, PCRG – Piscina Caldeira da Ribeira Grande.

Sample	No. of Reads (R1=R1; nx2)	No. of Contigs	No. of Scaffolds
CV	47663521	198245	196495
NP	30793369	53856	52145
ESG	31628129	35828	35466
FCRG	19960245	46840	45594
PCRG	15123250	17474	17019

2.3.3 Taxonomic profile

The taxonomic profile of all samples was performed using two approaches: assembly-based analysis and read-based analysis. Recommendations for the use of both approaches for sequence analysis whenever possible have been made (Quince et al., 2017), once they complement and validate each other.

2.3.3.1 Assembly-based analysis - Binning

Assembly-based analysis was performed by applying the Binsanity v0.3.1 binning algorithm to the Scaffolds files. Results were then checked for quality, analyzing completeness and contamination. BinSanity results are presented in Table 2.4 which also includes the Kraken taxonomic identification of the bins that achieved values for completeness and contamination higher than 97% and lower than 10%, respectively. Complete results of the BinSanity and Kraken algorithms are presented in Supplementary Tables S4 and S5.

Although binning results add little information on species identification (around 9% to 20%), some conclusions can be made that corroborate sample bacterial and archaeal diversity shown in section 2.3.1.

In fact, all samples appear to have high prokaryotic diversity, as shown by the total number of bins predicted for each sample. However, Kraken results are hugely dependent on the availability of marker genes and reference genomes in databases, making it difficult to profile organisms present in low quantities in samples, as well as uncharacterized organisms that are typically present in the soil, marine, and freshwater environments (Breitwieser et al., 2019), such as the ones being analyzed in this work.

Table 2.4. Binning algorithm prediction results. Number of total predicted bins, bins with high completeness (>97%) and low contamination (<10%) values, and respective species identification. Samples CV – Caldeira Velha, NP – Nascente Poente, ESG – Esguicho de Maio, FCRG – Fumarola Caldeira da Ribeira Grande, PCRG – Piscina Caldeira da Ribeira Grande.

Sample	Number of bins		Species Identification	Domain
	Total	High Completeness Low Contamination		
CV	17	2	*Sulfurihydrogenibium azorense* Az-Fu1	Bacteria
			Thiomonas intermedia K12	Bacteria
NP	12	2	*Pyrobaculum islandicum* DSM 4184	Archaea
			Thermus scotoductus SA-01	Bacteria
ESG	10	2	*Sulfurihydrogenibium azorense* Az-Fu1	Bacteria
			Pyrobaculum islandicum DSM 4184	Archaea
FCRG	11	1	*Thermoplasma acidophilum* DSM 1728	Archaea
PCRG	10	2	*Acidithiobacillus caldus* SM-1	Bacteria
			Acidimicrobium ferrooxidans DSM 10331	Bacteria

As mentioned before, only a small number of genomes could be identified using this approach. Given the methodology of this taxonomic profiling approach, the identified genomes have a fair probability of being the most abundant in each sample (Hug et al., 2016). Indeed, all samples except sample FCRG, which is also one of the samples that revealed higher archaeal diversity, show bacterial diversity in accordance with the results obtained by 16S rRNA gene amplicon sequencing (section 2.3.1). Binning was not able to predict any species identification of archaeal diversity in CV and PCRG samples, although species identification of around 80% of the predicted bins remains inconclusive in both samples. Moreover, the results are also in accordance with previous studies that report the high abundance of *Aquificales* (*Sulfurihydrogenibium azorense* Az-Fu1) and *Crenarchaeota* members (*Pyrobaculum islandicum* DSM 4184) in hydrothermal sites (Strazzulli et al., 2017).

2.3.3.2 Read-based analysis - Operational taxonomic units (OTUs)

Read-base taxonomic profiling of metagenomes more than identifying microbial species in a metagenome also estimates their abundance. Based on simply mapping reads to reference genomes, these methods are ideal for profiling low-abundance organisms (Nielsen et al., 2014). OTUs generated with the mOTUs profiler v.2.0.0 (Milanese et al., 2019) for each sample are presented in Table 2.5. Abundance percentage values presented were normalized for the species with an abundance percentage higher than 1%. Complete mOTUs results are presented in Supplementary Table S6. Accounting for all samples, a total of 12 organisms were identified.

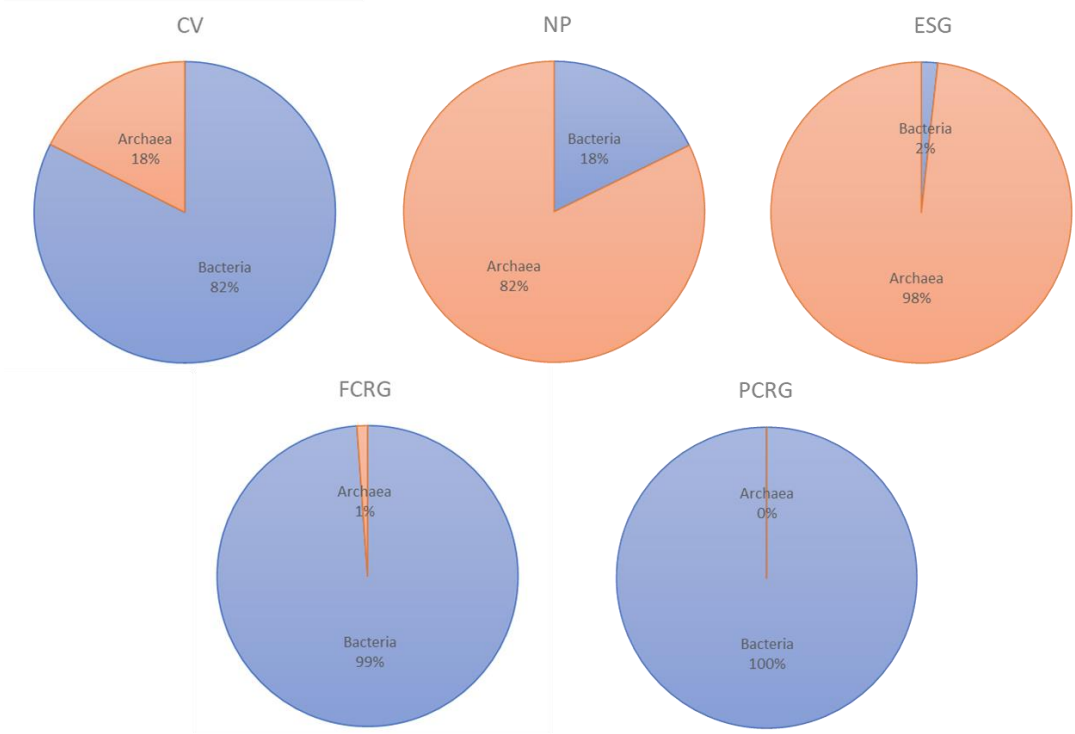


Figure 2.2. Domain-level composition of microbial communities based on mOTUs algorithm organism identification and respective abundance prediction.

The obtained results agree with the ones obtained in section 2.3.3.1, for both prokaryotic diversity and organism identification. As predicted by assembly-based analysis, all samples seem to have bacterial and archaeal diversity, exception made again for sample PCRG that is composed only of bacteria (Figure 2.2). Although mOTUs algorithm only predicted the presence of two organisms in the PCRG sample, the

binning algorithm predicted 10 bins, that can or not be mapped to actual different genomes (Hug et al., 2016). The identification procedure in the taxonomic profile process of both approaches is highly dependent on reference genomes in databases, so archaeal diversity detected using 16S rRNA gene amplicon sequencing (section 2.3.1) could be resultant of unknown archaeal organisms (Zorrilla et al., 2021). CV reveals to be the sample with higher prokaryotic diversity (composed of 9 different microorganisms), showing higher bacterial diversity than archaeal, which is in accordance with all previous results.

Considering organism identification, all samples show that the most abundant organisms identified by the read-based approach are also the ones identified by the binning approach. However, *Thiomonas intermedia* K12 and *Thermus scotoductus* SA-01 in samples CV and NP, respectively, were predicted by both approaches but are not included in the most abundant by the mOTUs algorithm. Moreover, results for both approaches exhibit remarkable similarity for samples with lower prokaryotic diversity (ESG, FCRG, and PCRG), showing that indeed read-based and assembly-based analysis approaches validate and complement each other. A unique divergence in the results has been spotted: the two approaches have identified in samples CV and ESG a different member of the *Pyrobaculum* genus, *Pyrobaculum islandicum* DSM 4184 by the binning approach and *Pyrobaculum aerophilum* str. IM2 by the read-based approach. In fact, the two organisms are genetically and metabolically similar (Feinberg et al., 2008) so these differences can be due to the use of different reference genome databases (mOTUs uses an internal reference genomes database while BinSanity uses NCBI RefSeq database) or even to incorrect fragment assignment which can prevent achieving accurate contigs of high-abundant genomes, before organism identification (Breitwieser et al., 2019).

Table 2.5. Operational Taxonomic Units algorithm results. For each sample only organisms with respective abundance higher than 1% are presented. Shaded organisms belong to the Domain Bacteria, and non-shaded organisms belong to the Domain Archaea. Samples CV – Caldeira Velha, NP – Nascente Poente, ESG – Esguicho de Maio, FCRG – Fumarola Caldeira da Ribeira Grande, PCRG – Piscina Caldeira da Ribeira Grande. Organisms: *Acidimicrobium ferrooxidans* DSM 10331 (*A. ferrooxidans* DSM 10331), *Acidithiobacillus caldus* SM-1 (*A. caldus* SM-1), *Desulfurococcus amylolyticus* DSM 16532 (*D. amylolyticus* DSM 16532), *Pyrobaculum aerophilum* str. IM2 (*P. aerophilum* str. IM2), *Sulfurihydrogenibium azorense* Az-Fu1 (*S. azorense* Az-Fu1), *Thermodesulfovibrio yellowstonii* DSM 11347 (*T. yellowstonii* DSM 11347), *Thermofilum adornatus* 1505 (*T. adornatus* 1505), *Thermoplasma acidophilum* DSM 1728 (*T. acidophilum* DSM 1728), *Thermorudis peleae* (*T. peleae*), *Thermus antranikianii* DSM 12462 (*T. antranikianii* DSM 12462), *Thermus scotoductus* SA-01 (*T. scotoductus* SA-01), *Thiomonas intermedia* K12 (*T. intermedia* K12).

CV		NP		ESG		FCRG		PCRG	
Organism	**%**	**Organism**	**%**	**Organism**	**%**	**Organism**	**%**	**Organism**	**%**
S. azorense Az-Fu1	58.6	*P. aerophilum* str. IM2	79.9	*P. aerophilum* str. IM2	96.0	*T. acidophilum* DSM 1728	98.7	*A. caldus* SM-1	64.9
T. adornatus 1505	9.9	*T. antranikianii* DSM 12462	12.4	*S. azorense* Az-Fu1	1.8	*A. caldus* SM-1	1.1	*A. ferrooxidans* DSM 10331	35.1
T. yellowstonii DSM 11347	7.2	*T. scotoductus* SA-01	5.3	*D. amylolyticus* DSM 16532	1.3				
A. caldus SM-1	7.0	*D. amylolyticus* DSM 16532	1.3						
A. ferrooxidans DSM 10331	4.9	*T. adornatus* 1505	1.1						
T. acidophilum DSM 1728	4.8								
P. aerophilum str. IM2	2.9								
T. peleae	2.5								
T. intermedia K12	2.1								

Once more, mTOUs algorithm, in accordance with the binning algorithm, revealed a high abundance of the hydrothermal characteristic taxonomic groups *Aquificales* (*Sulfurihydrogenibium azorense* Az-Fu1) and *Crenarchaeota* (*Desulfurococcus amylolyticus* DSM 16532, *Pyrobaculum aerophilum* str. IM2 and *Thermofilum adornatus* 1505) members (Strazzulli et al., 2017). In particular, *Sulfurihydrogenibium azorense* Az-Fu1 was identified in the samples CV and ESG, being the most abundant organism in the CV sample. Curiously, *S. azorense* Az-Fu1 was first isolated in Furnas, São Miguel, Azores (Aguiar et al., 2004) and observed in other metagenomics studies of the same site (Sahm et al., 2013). Therefore, the prediction of the presence of this bacterium on other hydrothermal sites suggests that this bacterium is well adapted to the physical and chemical environmental conditions of São Miguel Island. Being a representative of the *Aquificales* order, the metabolic capabilities of *S. azorense* Az-Fu1 in isolation, as well as within its microbial communities is therefore of huge interest.

2.4 Conclusions

Since the mid-2000s, with the genesis of next-generation sequencing, metagenomic studies have transformed the understanding of microbial communities' function and dynamics through the global effort to catalog the uncultured microbial diversity throughout the most diverse environments (Gilbert et al., 2014; Rusch et al., 2007; Turnbaugh et al., 2007). More recently, the scientific community has focused on the study of environments that once were described as unhabitable (Rothschild et al., 2001) where a large number of organisms, known as extremophiles (Nicolaus et al., 2010) were revealed, presenting remarkable genomic and metabolic attributes (Durvasula et al., 2018) with industrial applications.

In this work, the prokaryotic diversity of five samples from hydrothermal vents at São Miguel, Azores, was characterized, using metagenomic approaches. The taxonomic profiling of all samples was performed using assembly-based and read-based analysis algorithms. Although both approaches are usually used separately and selected depending on the origin of the sample in the study, the use of both is usually recommended (Quince et al., 2017) whenever possible. Indeed, both taxonomic profiling approaches presented very similar results demonstrating that the two approaches validate and complement each other. Differences were only spotted on the assignment of different members of the *Pyrobaculum* genus depending on the profiling algorithm, which can be justified by the use of different reference genome databases.

High abundant organisms were identified in all samples and domain-level composition was also predicted. The results of both taxonomic profiling approaches showed a similar profile to the experimental 16S rRNA gene amplicon sequencing profile obtained. Differences were only encountered in sample PCRG, which showed a considerable detection of archaeal 16s RNA, however, no archaeal organisms were predicted to be present when applying the taxonomic profiling algorithms. No conclusion can be made on the presence or absence of archaeal diversity in the PCRG sample, since the taxonomic profiling algorithms available have indeed some limitations, due to the still big amount of unculturable organisms without a reference genome available in databases (Zorrilla, Patil, et al., 2021).

Aquificales and *Crenarchaeota* members, regularly present in hydrothermal sites (Strazzulli et al., 2017), were predicted to be present in high abundance in samples CV, NP, and ESG, by both taxonomic profiling approaches. Specifically, *Sulfurihydrogenibium azorense* Az-Fu1 and *Pyrobaculum islandicum*

DSM 4184/ *Pyrobaculum aerophlium* str. IM2, *Aquificales,* and *Crenarchaeota* members, respectively, appear to be abundantly present in various samples.

Environments rich in *Aquificales* members have recently received scientific interest for being believed to be the earliest lineage within the domain *Bacteria* (Takacs-Vesbach et al., 2013) and for being a common presence in thermophilic and hyperthermophilic environments. Curiously, *S. azorense* Az-Fu1 was first isolated in January 2001 from terrestrial hot springs at Furnas, São Miguel Island, Azores, Portugal (Aguiar et al., 2004) and its re-occurrence at different sites with similar environmental conditions is worthy to be investigated. A systems biology approach to study *S. azorense* Az-Fu1 metabolic capabilities in isolation, as well as part of these communities, is suggested.

2.5 Supplemental Material

Additional file in Excel format: Chapter2_Supplementary_Material.xlsx

Link: DesignOptimizationMicrobialCommunities/Data

Table S1 FastQC Report Summary

Table S2 FastP Report Summary

Table S3 MetaQuast Quality Results

Table S4 BinSanity Complete Results

Table S5 Kraken Results for the bins with high completeness (>97%) and low contamination (<10%) values and respective species identification

Table S6 mOTUs Complete Results

Chapter 3

Genome-scale metabolic model of thermophilic bacterium *Sulfurihydrogenibium azorense Az-Fu1*

"All for one, one for all"

Alexandre Dumas

*i*SS352 is the first manually curated genome-scale metabolic model for the thermophilic bacterium *S. azorense* Az-Fu1, comprising 352 genes and 772 reactions. The construction of this model involved performing a reannotation, which revealed the presence of the main components of the bacterial cellulose operon and its regulators, suggesting that *S. azorense* Az-Fu1 may have the metabolic potential for cellulose production. The model also clarifies this chemolithoautotrophic organism's carbon fixation route, central carbon, and sulfur metabolisms. The *i*SS352 metabolic model will serve the ongoing fundamental research of chemolithoautotrophic metabolism in extreme environments, provide clues regarding new extremophilic enzymes, and studies of interactions with the identified environmental microbial community in which it was discovered.

3.1 Introduction

The *Aquilicales* order is of significant interest as it is believed to be the earliest-lineage within the domain *Bacteria* (Takacs-Vesbach et al., 2013). Being widespread in marine and terrestrial geothermal systems, members of *Aquificales* are thermophilic and hyperthermophilic bacteria almost exclusively obligate or facultative autotrophs (Hedlund et al., 2015). Considered to be primary producers, members of this order can also use a wide range of energy sources, such as H_2, elemental sulfur, and thiosulfate (Takacs-Vesbach *et al.*, 2013, Yamanaka, 2008). Moreover, the study of extremophiles and their enzymes has increased in the last years (Counts et al., 2017; Han et al., 2019; Kumar et al., 2019) due to the need for the industry to have more stable biocatalysts at higher temperatures and pressures or extreme pHs to resist to harsh industrial settings (Atalah et al., 2019) and also for using less costly purification technologies, such as heat treatment that does not involve chemical reagents, protein tags, nor adsorbents (Sun et al., 2012).

Sulfurihydrogenibium azorense Az-Fu1, a representative of the *Aquificales* order, commonly found as dominant taxa in hot spring vent communities, is a gram-negative, thermophilic, chemolithoautotrophic, and microaerophilic bacterium (Lalonde et al., 2005). The bacterium was isolated in January 2001 from terrestrial hot springs at Furnas, São Miguel Island, Azores, Portugal (Aguiar et al., 2004). *S. azorense* grows optimally at 68°C, pH 6, and at low concentrations of NaCl and can also grow heterotrophically (Nakagawa et al., 2005) and use elemental sulfur, thiosulfate, hydrogen, and ferrous iron as energy sources (Aguiar et al., 2004). As an *Aquificales* member, *S. azorense* is believed to fix CO_2 via reductive Tricarboxylic Acid Cycle (rTCA) (Hügler et al., 2007) to generate acetyl-CoA as an end product (Gupta et al., 2013), like its closely related relative *Sulfurihydrogenibium subterraneum*. In fact, the rTCA is reported to be the most plausible candidate for the first autotrophic metabolism in the earliest life (Kitadai et al., 2017). Also, (Lalonde et al., 2005) reported that, under stress conditions, this bacterium produces chemolithoautotrophically sufficient amounts of exopolysaccharides (EPS).

EPS are polymeric structures of repeated sugar units of the same or different types that microorganisms usually produce under unfavorable growth conditions. These structures have a high range of physiological roles, such as preventing the entering of harmful substances within the cell, contributing to providing nutrients and/or water or promoting adherence to surfaces (biofilm formation) (Rana & Upadhyay, 2020). In extremophiles, EPS accumulation functions as a relevant adaptation strategy of cell protection and membrane stabilization (Kambourova et al., 2016). Due to their adaptation

to extreme environmental conditions, EPS produced by extremophiles is usually more stable under a wide range of temperatures, pH, and saline conditions. Moreover, these EPS have novel physical and chemical properties, making these polymers good candidates for use in food, cosmetic, and pharmaceutical industries, and novel biomedicine areas (Nicolaus et al., 2010). Hence, the possibility of *S. azorense's* chemolithotrophic and thermophilic production of EPS suggests that the study of its metabolism is of significant interest.

One of the most used tools to capture the complex physiological characteristics, behavior, and metabolic capabilities of a cell as an integrated system is genome-scale metabolic models (GSMMs). These models are mathematical representations describing an organism, cell, tissue, or microbial community metabolism that can fully explore metabolic relationships between genotype and phenotype (Bordbar et al., 2014; Kim et al., 2008). These mathematical models are a valuable platform for the rapid testing of hypotheses. Since the publication of the first GSMM (Edwards et al., 1999), these models have received increasing interest, as shown by the increasing list of applications: metabolic engineering strategies, identification of gene functions, identification of drug targets in pathogens, and more recently the integration of multi-omics data to study metabolic rewiring of human cells or tissues, and the prediction of microbe-microbe/host-microbe interactions within microbial communities (Kim et al., 2017). The number of published GSMMs is increasing and is available in several databases, such as BioModels (Malik-Sheriff et al., 2020) and BiGG Models (King et al., 2016). Most of those models represent prokaryotic organisms; however, when related to chemolithoautotrophic organisms, little is still known about their metabolism, and consequently, few are published as GSMMs, as seen in Table 3.1.

High-quality GSMMs must be manually curated and validated with experimental data to predict a given phenotype, a process that, even with available tools to automate most of the reconstruction tasks, is still laborious and time-consuming (Thiele et al., 2010). Workflows that semi-automatically construct a GSMM of a prokaryotic organism in a matter of minutes and decrease the time spent in these reconstructions, such as the platform KBase (Arkin et al., 2018), or the python package CarveMe (Machado et al., 2018), have been published in the last years.

These models have already shown their reliability when predicting growth, response to nutrients, and gene essentiality in single organisms and even microbial communities (Chng et al., 2020; Machado et al., 2021; Nayfach et al., 2020). However, if the organism has particular nutritional requirements and

metabolism, and if the objective is the rational design, these models should be used with caution, as results may be inaccurate (Lieven et al., 2020), needing further manual curation.

Table 3.1. Chemolithoautotrophic organisms with published genome-scale metabolic models. CBB - Calvin-Benson-Bassham cycle, rTCA – reverse Tricarboxylic acid cycle, 3-HP/4-HB - 3-hydroxypropionate/4-hydroxybutyrate cycle, DC/4-HB - dicarboxylate/4-hydroxybutyrate cycle, WL - Wood-Ljungdahl pathway.

Organism	Domain	Bioenergetic process	Electron donor	Electron acceptor	Carbon fixation	Reference
Acidithiobacillus ferrooxidans	Bacteria	Sulfur and Iron Oxidizing	S, $S_2O_3^{2-}$, Fe^{2+}	O_2	CBB	(Campodonico et al., 2016)
Methanococcus maripaludis S2	Archaea	Methanogen	H_2	CO_2	WL	(Goyal et al., 2014)
Methanosarcina barkeri	Archaea	Methanogen	H_2	CO_2	WL	(Feist et al., 2006)
Nitrobacter winogradskyi	Bacteria	Nitrate Oxidizing	NO_2	O_2	CBB	(B L Mellbye et al., 2018)
Nitrosomonas europaea	Bacteria	Ammonia Oxidizing	NH_3	O_2	CBB	(B L Mellbye et al., 2018)
Nitrosopumilus maritimus SCM1	Archaea	Ammonia Oxidizing	NH_3	O_2	3-HP/4-HB	(F. Li et al., 2018)
Nitrospira moscoviensis	Bacteria	Nitrate Oxidizing	NO_2^-	O_2	rTCA	(Lawson et al., 2021)
Sulfolobus solfataricus P2	Archaea	Sulfur Oxidizing	S, $S_2O_3^{2-}$	O_2	3-HP/4-HB	(Ulas et al., 2012)

Here we present a manually curated and validated GSMM of the sulfur and hydrogen oxidizing thermophilic bacterium *S. azorense Az-Fu1*. The key metabolic capabilities were analyzed to give insights into the chemolithoautotrophic CO_2 fixation process and to elucidate the potential of this bacterium to produce extracellular polymeric substances.

3.2 Methods

3.2.1 Online Databases

Different online databases were used in each stage of this work, most of them through *merlin*'s framework. The National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators, 2018), was used to retrieve the genome sequence of *S. azorense Az-Fu1* (assembly accession number ASM2154v1), and all genome files were imported by *merlin*. Universal Protein Resource Knowledgebase (UniProtKB) (Consortium, 2021) and BRENDA (A. Chang et al., 2021) were used to obtain enzyme functional information through *merlin*'s re-annotation pipeline, Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016) was used to obtain reactions based on the enzyme commission numbers (EC numbers) of the annotated genome into *merlin*, and also to perform pathway analysis. MetaCyc (Caspi et al., 2014) and BiGG models (King et al., 2016) were used for network curation. ModelSEED (Seaver et al., 2021) was used for *merlin*'s workflow for correct reversibility of reactions. PSORTb 3.0 (Yu et al., 2010) was used to predict reactions compartments, while the Transporters Classification Database (TCDB) (Saier et al., 2021) was used to predict the model transport reactions through *merlin*'s TranSyT plug-in.

3.2.2 Metabolic Model Reconstruction

merlin (Dias et al., 2018) is a user-friendly framework that allows performing several steps of the reconstruction process semi-automatically, downloading relevant information from several databases (see section 3.2.1) and was used to assist the reconstruction of the *S. azorense* GSMM. Moreover, it has a graphical interface that facilitates GSMM information reviewing and manual curation. The main steps of the GSMM model reconstruction process are hereafter described.

3.2.2.1 Genome Annotation

merlin allows performing the functional annotation of a genome, using as similarity search engines the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) and Diamond (Buchfink et al., 2021) against databases that contain reviewed (such as UniProt/Swiss-Prot) and unreviewed (such as UniProt/TrEMBL) enzyme information. The EC numbers and enzymatic functions assigned to each gene

are scored based on the taxonomy and frequency of similar sequences, as described elsewhere (Dias et al., 2018). The genome annotation is a crucial step of the GSMM reconstruction process, as incorrect EC numbers and enzymatic function assignments can significantly impact the model performance. Once the similarity search is complete, *merlin*'s automatic annotation workflow feature (Capela et al., 2021) can prioritize gene products and EC numbers obtained. This operation considers a list of organisms ordered by phylogenetic similarity provided by the user and defines a confidence level (A to I) for each gene annotation when a match is found. In the case of *S. azorense,* due to the lack of available information on the (closely related) organism(s) and to maximize retrieval of reviewed information, the automatic workflow feature was configured to use genus instead of species as input. The ranked list of the closely related phylogenetic genus to *S. azorense* is presented below (Table 3.2). The phylogenetic tree (Supplementary Figure 3.1) was constructed from 16S RNA sequences reference organisms of each genus using EMBL-EBI Clustal OMEGA multiple sequence alignment tool (Sievers et al., 2011).

Table 3.2. List of phylogenetic similar organisms/genus to *S. azorense* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Sulfurihydrogenibium azorense Az-Fu1	A
genus *Sulfurihydrogenibium*	B
genus *Persephonella*	C
genus *Aquifex*	D
genus *Hydrogenobaculum*	E
genus *Thermus*	F
genus *Thermotoga*	G
genus *Deinococcus*	H
genus *Desulfobacterium*	I

3.2.2.2 Assembling the Metabolic Network

The assembly of a metabolic network starts with gathering all reactions present in the organism. *merlin* was used to retrieve those reactions by importing them from the KEGG database, based on the annotated EC numbers from the previous step, and spontaneous reactions. Reactions should also be balanced (generic metabolites and metabolites without formula must be curated), and reversibility must be confirmed to avoid mispredictions of the model. To assist these steps, *merlin* tools to check whether a reaction is balanced and to correct the reactions' reversibility were used. Although these steps are automated in *merlin*, manual curation was performed based on curated information from literature and databases, such as MetaCyc (Caspi et al., 2014).

Compartmentalization

The compartmentalization of the model was based on results obtained from PSORTb 3.0 (Yu et al., 2010). The "Long format" report generated was imported, and *merlin*'s compartments feature assigned each reaction to its specific compartment (Capela et al., 2021).

Transport Reactions

With the model compartmentalized, transport reactions between them must be defined. *merlin*'s TranSyT (Lagoa et al., 2021) was used to generate system-specific transport reactions associated with Gene-Protein-Reactions (GPRs) rules that were automatically integrated into the model.

Genes, proteins, and reactions

A high-quality GSMM requires GPRs' rules to predict genetic modifications accurately. These associations are usually defined according to databases and literature (Rocha et al., 2008; Thiele et al., 2010). *merlin*'s *"Gene-Protein-Reaction rules"* feature was used to automatically add GPRs' rules to the model. The algorithm used by *merlin* to implement these rules is described elsewhere (Dias et al., 2015).

3.2.2.3 Converting the Metabolic Network to a Stoichiometric Model

Biomass equation

Biomass composition must be experimentally determined in cells growing in the log phase before being included in the model. However, in the absence of organism-specific experimental information, data from genome information (particularly nucleotides, deoxynucleotides, and amino acids) can be used or adapted from phylogenetically related organisms.

For *S. azorense,* little information was found in the literature regarding biomass composition. The macromolecular composition was adapted from the gram-negative bacterium *Escherichia coli* (*E. coli*) (Feist et al., 2007). The composition of amino acids, nucleotides, and deoxynucleotides was estimated from *S. azorense Az-Fu1* genome information through *merlin*'s "e-Biomass Equation" feature. This feature also automatically includes cofactor composition based on a study of universal essential cofactors in prokaryotes (Xavier et al., 2017). The fatty acids composition was adopted from the closely related organism *Sulfurihydrogenibium subterraneum* (Takai et al., 2003). The lipids composition was adapted from the closely related organism *Hydrogenobacter thermophilus* (Yoshino et al., 2001), while cell wall components and carbohydrates were adapted from *E. coli*, considering enzyme annotation to include or exclude specific elements. When required, new coefficients were calculated, maintaining the relative abundances of the original data.

An alternative biomass equation was defined for simulations under anaerobic conditions, mimicking the environmental conditions where Heme is not required. Hence, this compound was removed from the Cofactor composition, and all other coefficients were recalculated as mentioned before.

Growth and maintenance ATP requirements

No information was found for *S. azorense* on growth and maintenance ATP. Therefore, such data was adapted from experimental data for *E. coli* (Feist et al., 2007; Neidhardt et al., 1990).

3.2.2.4 Metabolic Model Curation and Validation

The model curation is an iterative process that stops when simulation results match experimental data in the literature. *merlin*'s interface was used to efficiently perform re-annotations, correct reactions

stoichiometric balance and directionality, include or exclude reactions from the model, and finally, export models in Systems Biology Markup Language (SMBL) format to be used in simulation using platforms such as Optflux (I. Rocha et al., 2010), Matlab® or python (COBRApy (Ebrahim et al., 2013), REFRAMED - https://github.com/cdanielmachado/reframed)

Gap-filling

Before being ready for simulations, the metabolic network must be screened for possible gaps. The presence of gaps can deeply compromise the synthesis of biomass components and other relevant compounds. To assist in this process, *merlin*'s "BioISO" (Capela et al., 2021) was used to trace back the network to identify gaps that can be originated from errors in genome annotation, absence of enzymatic, transport, or exchange reactions, or incorrect reaction irreversibility or direction. Other features included in *merlin*, such "Blocked reactions", which identifies reactions that contain dead-end metabolites, and "Draw in Browser" which opens on a web browser a selected KEGG pathway, showing specifically highlighted enzymes and reactions present in the model, were used to facilitate the detection of gaps in the network (Capela et al., 2021). Literature and databases (KEGG, MetaCyc, BRENDA, for instance) were also used to assist the gap-filling process. This is an iterative process that is repeated until all biomass precursors, and other essential compounds, can be synthesized, and a feasible model is obtained.

Microaerobic and anaerobic metabolism

S. azorense is a facultative chemolithoautotrophic (Nakagawa et al., 2005), microaerophilic bacterium (Aguiar et al., 2004). Unlike plants, cyanobacteria, and other chemolithoautotrophic Proteobacteria that fix CO_2 through the Calvin-Benson-Bassham cycle, *S. azorense* metabolism was assessed to determine if it fixates CO_2 through the rTCA, such as its closely related organism *S. subterraneum* (Hügler et al., 2007) and most of *Aquificales* members. Anaerobic and microaerobic growth (specific growth rate of 0.28 h^{-1}) was also tested as reported in the literature (Aguiar et al., 2004; Nakagawa et al., 2005). EPS production (Lalonde et al., 2005) was also screened.

Carbon source utilization

S. azorense can grow heterotrophically using yeast extract, bactopeptone, trypticase peptone, and casamino acids (Nakagawa et al., 2005). Except for casamino acids, all other medium components reported as carbon sources are not chemically defined; thus, only casamino acids were tested for heterotrophic growth.

3.2.2.5 Simulations

Software

All simulations were performed using the python package REFRAMED, version 1.1.0, using CPLEX 12.8.0 as a solver, through the PyCharm Integrated Development Environment (IDE). REFRAMED provides tools for phenotype simulation such as Flux Balance Analysis (FBA) (Varma & Palsson, 1994)

Flux variability analysis

Quantitative evaluation of the model was performed using FVA (Mahadevan et al., 2003) to understand *S. azorense* EPS production capabilities. The analysis included setting the specific growth rate to at least 10% of the specific growth rate obtained with pFBA (Lewis et al., 2010) in the respective reference flux distribution.

3.3 Results and Discussion

3.3.1 Genome annotation

Assisted on the phylogenic tree developed in section 3.2.2.1 and *merlin*'s feature automatic workflow, genome re-annotation identified 774 metabolic genes (Table 3.3) out of 1657 candidates (Table 3.5), representing 47% of the whole *S. azorense* Az-Fu1 genome. The complete list of genes reviewed is available in Supplementary Table S1.

Manual curation was performed and the number of genes in the model decreased to 352 (21%) due, for instance, to the removal of pseudo and truncated genes, incomplete EC numbers, and blocked reactions with encoded genes.

Table 3.3. *S. azorense* genome annotation automatic workflow results taking into account phylogenetically related genus.

Organism	Gene count
Sulfurihydrogenibium azorense Az-Fu1	6
genus *Sulfurihydrogenibium*	101
genus *Persephonella*	13
genus *Aquifex*	202
genus *Hydrogenobaculum*	25
genus *Thermus*	21
genus *Thermotoga*	5
genus *Deinococcus*	12
genus *Desulfobacterium*	3
Default annotation	385

The analysis of the annotation results shows the presence of bacterial cellulose synthase (Bcs) subunits A (BcsA – SULAZ_1378) and B (BcsB – SULAZ_1376). As described elsewhere (Lalonde et al., 2005), *S. azorense* Az-Fu1 can produce sufficient amounts of EPS, whose composition is still unknown, under stress conditions. Hence, the presence of annotated subunits of cellulose synthase suggests that cellulose can be a part of *S. azorense* EPS. However, cellulose production is dependent not only on the catalytic subunits BcsA and BcsB but also on several additional BCS components and gene regulators (Krasteva et al., 2017; Römling, 2002). Hence, *S. azorense*'s genome was screened to identify other cellulose operon subunits missing from the current annotation.

The cellulose operon has three main types (Römling et al., 2015), as shown in Figure 3.1. All contain the catalytic subunits BcsA (cytosol) and BcsB (periplasm). These two subunits are essential for *in vitro* cellulose production. Subunits BcsC, present in type I and II operons, and BcsK, in type III are involved in cellulose transport, whereas subunit BcsD is present only in type I operon and is required for the Bcs complex arrangement. In type II and III operons, the BcsZ subunit regulates cellulose production. In type II operon, subunits BcsE and bcsG are required for optimal cellulose production, and subunit BcsQ seems to determine the cellular localization of the respective Bcs complexes (Römling et al., 2015). Other Bcs subunits have already been identified, but their function has not yet been unveiled. The regulation of the cellulose production process has also been described (Chang et al., 2001; Romling et al., 2013), showing the importance of bis-(3',5') cyclic diguanylic acid (c-di-GMP) molecule as an allosteric activator of subunit BcsA. In *Escherichia coli*, c-di-GMP is produced by GGDEF domain proteins such as AdrA or YedQ, and its degradation and shut-off of cellulose biosynthesis is achieved by specific phosphodiesterases, such as YhjH (Monteiro et al., 2009).

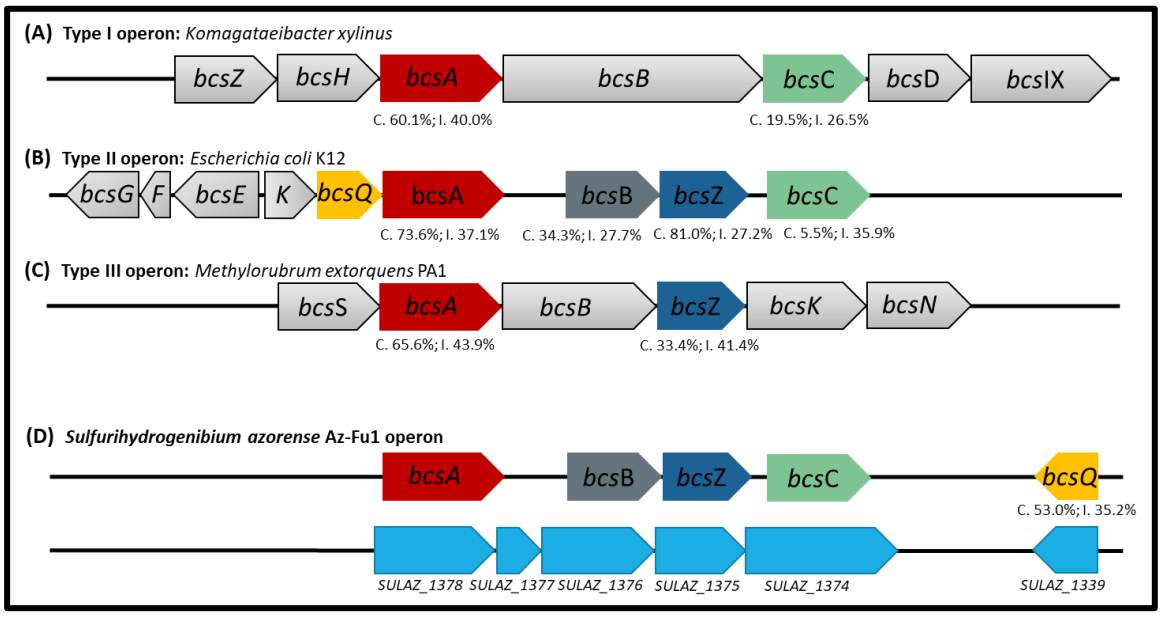


Figure 3.1. Comparison of cellulose operon types I, II, and III, respectively, (A), (B), and (C), with (D) - predicted *S. azorense* Az-Fu1 cellulose operon. Colors indicate that a match was found as a result of NCBI tblastn searches. Red – BcsA unit, Dark gray – BcsB unit, Dark blue – BcsZ, Green – BcsC, Yellow – BcsQ. (C) – coverage; (I) – identity.

As no information was found reporting the presence of any type of cellulose operon in *S. azorense*, gene accession numbers of all known units of the different BCS types and their regulators (Supplementary

Table 3.1) were analyzed. This analysis involved performing BLAST searches (NCBI tblastn) against *S. azorense* Az-Fu1's whole genome to identify the presence of BCS units and the respective BCS type, if possible. As a result, besides the already described presence of BcsA and BcsB, the BcsC and BcsZ units were found in the same order type II BCS operon, suggesting that if present, *S. azorense*'s cellulose operon will be a type II. Although all units were found to have adequate identity percentages (Figure 3.1), BcsB and BcsC show low coverage values. A gene sequence similar to the BcsQ unit was also found but not at the same position as in type II operons, which renders its identification inconclusive. Nevertheless, the re-annotation of genes SULAZ_1375 and SULAZ_1374 should be updated as bacterial cellulose synthase Z unit and bacterial cellulose synthase C unit, respectively. Regulators AdrA, YedQ, and YhjH were also identified with acceptable coverage and identity percentages (AdrA - C.50.3%, I. 35.2%; YedQ - C.42%, I.35.9%; YhjH – C.64.0%, I.28:7%). Although the main Bcs units and regulator factors were identified in the *S. azorense* Az-Fu1 genome, the bacterium's actual capability of producing cellulose must be *in vivo* experimentally assessed.

3.3.2 Biomass Composition

The biomass macromolecular composition, adapted from *E. coli* (Feist et al., 2007) is presented in Table 3.4. The detailed biomass composition is available in Supplementary Table S2. As mentioned before, the amino acid, deoxynucleotide, and nucleotide composition were calculated based on *S. azorense* Az-Fu1 genome's information using the e-Biomass feature in *merlin*.

This feature also automatically includes cofactor composition based on the study of universal essential cofactors in prokaryotes (Xavier et al., 2017). Usually, ubiquinone is included in the cofactors pool. However, most enzymes required for ubiquinone biosynthesis are not available in the genome annotation, and the defined medium (Aguiar et al., 2004) does not include ubiquinone; thus, this cofactor was omitted in the biomass formulation.

The lipids composition was adapted from the phylogenetically closely related organism *H. thermophilus* (Yoshino et al., 2001). However, *S. azorense* Az-Fu1 does not exhibit the enzymes responsible for producing phosphatidylserine, phosphatidylethanolamine, and phosphatidylcholine in the model. Hence, these compounds were excluded from biomass composition. The coefficients of the remaining compounds were recalculated, maintaining the relative abundances of the original data. The

average fatty acid composition was adopted from *S. subterraneum* (Takai et al., 2003) (Supplementary Table S3).

Table 3.4. Biomass macromolecular composition of the *S. azorense* Az-Fu1 model.

Biomass Composition	
	g gDW^{-1} (%)
Protein	53.3
DNA	2.7
RNA	13.6
Lipids	2.9
Carbohydrates	10.7
Cell Wall	6.8
Cofactors	10
Total	100

Cell wall and carbohydrate components were adapted from *E. coli* (Feist et al., 2007). Specifically for cell wall components, *E. coli*'s composition was reconciled between the KEGG reactions assigned through *merlin*'s annotation to its biosynthesis pathway (Supplementary Table S4).

The growth-associated energy (GAM) and non-growth-associated energy (NGAM) requirements for *S. azorense* are not been experimentally determined yet. Therefore the GAM requirements of 56.64 mmol $_{ATP}$ gDW^{-1} h^{-1} were estimated according to (Thiele et al., 2010) and based on data for *E.coli* (Neidhardt et al., 1990). The NGAM requirements of 8.39 mmol $_{ATP}$ gDW^{-1} h^{-1} were adopted from *E. coli* (Feist et al., 2007).

3.3.3 Metabolic Model

The genome-scale metabolic reconstruction of *S. azorense* Az-Fu1, *i*SS352, was generated through a bottom-up approach. The final reconstruction contains 352 genes, 772 reactions (62 exchange, 141

transport, and 25 gap-fill reactions), and 642 metabolites (Table 3.6), distributed over 70 subsystems and three compartments: extracellular, periplasm, and cytoplasm (Supplementary Tables S6, S7, and S8). The gene ratio is about 21%, a value in accordance with other published GSMMs for chemolithoautotrophic organisms (Supplementary Table S5), and 70% of all reactions have a GPR associated (Table 3.6). Reactions without GPR associations include spontaneous, exchange, outer membrane transport, and diffusion transport reactions of metabolites, such as CO_2, water, or O_2. The final GSMM, identified as *i*SS352, is included in Supplementary Material (*i*SS352.xml) in the SBML level 3 version 2 format. The model was able to score over 97% on all consistency tests on the MEMOTE (Lieven et al., 2020) test suit.

Table 3.5. *S. azorense* Az-Fu1 genome information

Genome Information	
Genome length (bp)	1640877
G+C content (%)	32.8
No. of ORFs	1657

Table 3.6. *i*SS352 metabolic model information

Metabolic Model	
Reactions	772
Metabolites	642
Genes	352
Gene Rules	546

3.3.4 Metabolism of *S. azorense* as represented in the *i*SS352 model

3.3.4.1 Carbon metabolism

Autotrophic CO_2 fixation is one of the most critical biosynthetic processes in nature. The Calvin-Benson-Bassham cycle is one of the most recognized pathways, mainly used by plants and cyanobacteria. Additional CO_2 fixation mechanisms that motivate the scientific community to explore possible alternative CO_2 fixation routes for reducing the human carbon footprint (Bar-Even et al., 2010; Liu et al., 2020) include five natural autotrophic pathways, namely the rTCA, dicarboxylate/4-hydroxybutyrate cycle, 3-hydroxypropionate cycle, 3-hydroxypropionate/4-hydroxybutyrate, and Wood-Ljungdahl cycles.

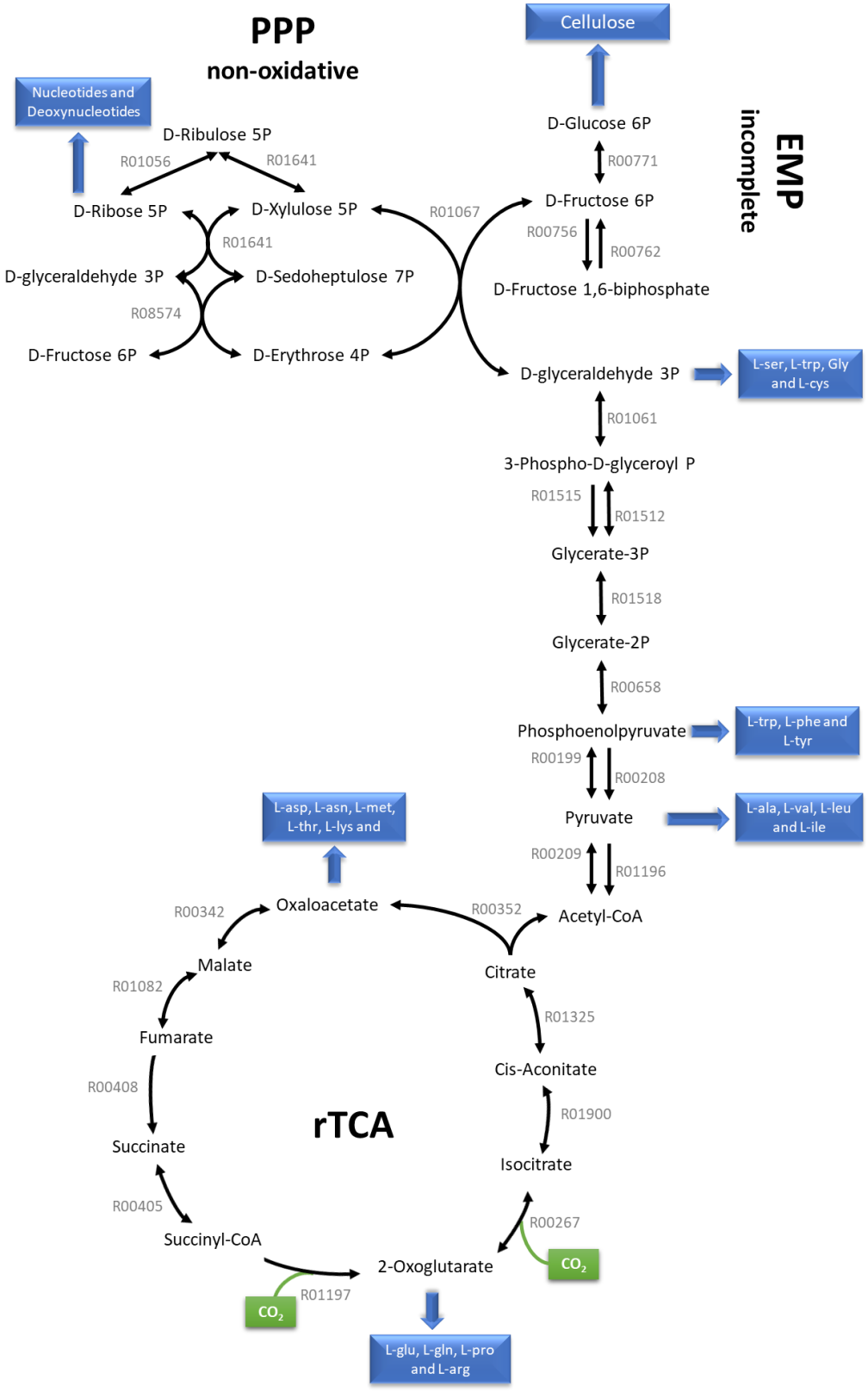


Figure 3.2. *Sulfurihydrogenibium azorense* Az-Fu1 proposed central carbon metabolism under chemolithoautotrophic growth and a potential route for cellulose production. rTCA – reverse Tricarboxylic Citrate Acid cycle; PPP – Pentose Phosphate Pathway; EMP - Embden-Meyerhof-Parnas (EMP) glycolytic pathway.

As reported for the closely related organism *S. subterraneum* (Hügler et al., 2007) and as most members of the *Aquificales* order (Gupta et al., 2013), *S. azorense* Az-Fu1 seems to fix CO_2 through the rTCA and produce all biomass components (Figure 3.2). The presence of three specific enzymes characterizes this fixation pathway: fumarate reductase (SULAZ_0630/R00408), ferredoxin-dependent 2-oxoglutarate synthase (SULAZ_0641/R01197), and ATP-citrate lyase (SULAZ_0527/R00352), all present in the *S. azorense* Az-Fu1 genome. Other CO_2 fixation pathways have been screened. However, several vital enzymes were missing from the organism's genome. Moreover, given the fact that *S. azorense* Az-Fu1 is a thermophilic organism that grows under microaerobic conditions, which is an energy-limiting environment, it makes sense that the rTCA cycle is the preferred route of CO_2 fixation for this organism. Compared to the Calvin-Benson-Bassham cycle, the rTCA cycle requires significantly less energy as ATP and fewer reducing equivalents to synthesize a three-carbon unit (Dahl et al., 2008). Surprisingly, *S. azorense* has a reduced central carbon metabolism, lacking all the oxidative part of the Pentose Phosphate Pathway (PPP) and the enzyme fructose-bisphosphate aldolase, originating an incomplete Embden-Meyerhof-Parnas (EMP) glycolytic pathway (Figure 3.2).

Initially described as an obligate chemolithoautotroph (Aguiar et al., 2004), able to use a variety of electron donors and acceptors, *S. azorense* Az-Fu1 was later considered a bacterium with the ability to grow heterotrophically in yeast extract, bactopeptone, trypticase peptone, and casamino acids (Nakagawa et al., 2005). In fact, amino acid transporters were predicted to be present in *S. azorense* Az-Fu1's genome, suggesting these as the only carbon source besides CO_2. Moreover, no sugar transporters other than cellulose were predicted in this work. As mentioned in section 3.2.2.1, *S. azorense* Az-Fu1 has the main cellulose biosynthesis operon subunits and respective regulators in its genome, suggesting that cellulose production seems metabolically plausible.

3.3.4.2 Sulfur metabolism

As a sulfur-oxidizing bacterium, the study of *S. azorense* Az-Fu1 sulfur metabolism is of significant interest. Sulfur chemolithotrophy is regarded as the earliest self-sustaining metabolism, as it is found in several extremophiles of the deepest phylogenetic branches of Archaea and Bacteria (Ghosh et al., 2009). In fact, sulfur occurs in nature in oxidation states ranging from +2 to -6, and is thus used by prokaryotes to build cell constituents and as an energy source (Dahl et al., 2008). Over the last years, several studies have attempted to identify the sulfur-oxidizing pathways. Neutrophilic sulfur-oxidizing bacteria, such as *S.*

azorense, seem to use two pathways for sulfur oxidation: the multienzyme complex catalyzing the complete oxidation of reduced sulfur compound to sulfate (Sox pathway) and a pathway having sulfite and sulfur as important intermediates (via sulfide:quinone reductase) (Friedrich et al., 2001). According to the annotation, *S. azorense* Az-Fu1 seems to have the Sox enzyme system. This system is present in the periplasm and is responsible for the oxidation of thiosulfate to sulfate. When other sulfur sources are used a first step of conversion into thiosulfate is needed. The Sox gene cluster soxXYZABCD encodes four periplasmic proteins, SoxXA, SoxYZ, SoxB, and Sox(CD)$_2$ (sulfur dehydrogenase). In *S. azorense* Az-Fu1 the protein Sox(CD)$_2$ is absent, indicating that this bacterium is likely to have a truncated Sox system, similar to the closely related organism *Aquifex aeolicus* (Friedrich et al., 2005).

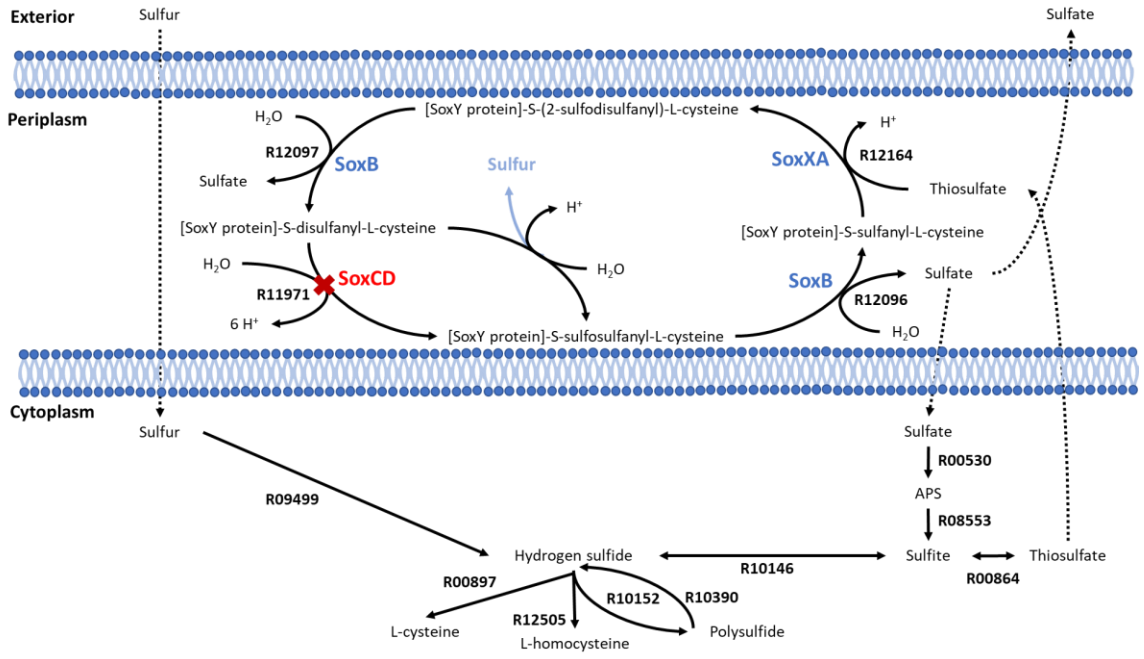


Figure 3.3. *Sulfurihydrogenibium azorense* Az-Fu1 proposed sulfur metabolism through a truncated sulfur-oxidizing (Sox) system. Here elemental sulfur is being used as the main sulfur source. The proposed process is similar when hydrogen sulfide, sulfite, or thiosulfate are used as a sulfur source.

Studies in the purple sulfur bacterium *Allochromatium vinosum*, which also has a truncated Sox system lacking Sox(CD)$_2$, indicate that elemental sulfur is produced from thiosulfate throughout the sulfur-oxidizing process and transferred to the cytoplasm (Figure 3.3) or growing sulfur globules. The mechanism for this process is still unresolved, although it may involve a rhodanese-like enzyme (SoxL) (Dahl & Friedrich, 2008). The truncated Sox enzyme system is less energy efficient once it yields only two

mol electrons per mol of thiosulfate while eight mol are yielded in the complete Sox system (Figure 3.3). As there is no report in literature corroborating the production or accumulation of sulfur globules by *S. azorense* Az-Fu1, the elemental sulfur produced in the periplasm by the truncated Sox system was modeled as being transferred to the cytoplasm.

3.3.5 Model Validation

3.3.5.1 Environmental conditions

The validation of the model involved using a minimal medium for each condition: chemolithoautotrophic (Aguiar et al., 2004) and heterotrophic growth (Nakagawa et al., 2005), both under microaerophilic and anaerobic conditions (Table 3.7).

Table 3.7. Minimal medium composition for each condition tested: chemolithoautotrophic and heterotrophic growth. Oxygen and Ferrous iron (highlighted in grey) were only supplied under microaerophilic conditions.

Chemolithoautotrophic		**Heterotrophic**	
Component	**Uptake value** (mmol g_{DW}^{-1} h^{-1})	**Component**	**Uptake value** (mmol g_{DW}^{-1} h^{-1})
CO_2	12	**Casamino Acids**	0.6
Sulfur	1000	**Sulfur**	1000
Fe^{2+}	1000	**Fe^{2+}**	1000
Ammonia	1000	**Ammonia**	1000
O_2	2	**O_2**	2
$H_3PO_4^-$	1000	**$H_3PO_4^-$**	1000
H_2O	1000	**H_2O**	1000

Heterotrophic growth was tested for casamino acids, the only chemically defined carbon source described in the literature. All medium components were allowed to enter the system unconstrained, except for carbon sources and oxygen, at microaerobic conditions. Two different biomass equations were included in the model to better evaluate growth under microaerobic and anaerobic conditions, as mentioned above. These reactions, "R_Biomass__cytop" and "R_Biomass_anaerobic__cytop", differ in

the presence or absence of heme in the Cofactor composition respectively. Oxygen (O_2) and ferrous iron (Fe^{2+}) were supplied under microaerophilic conditions. Growth using different electron donors and acceptors was also tested. Unless otherwise stated, the uptake rate value for these compounds was set to the maximum of 5 mmol g_{DW}^{-1} h^{-1}.

3.3.5.2 Modeling simulations

The validation process of the *i*SS352 model involved comparing simulation results to the works of (Aguiar et al., 2004) and (Nakagawa et al., 2005). Due to the lack of quantitative data, most assessments were performed qualitatively (Table 3.8).

Table 3.8. *i*SS352 model validation against experimental conditions from literature.

Growth condition	Predicted growth rate (h^{-1})	Observed growth rate (h^{-1})
Carbon source		
CO_2 (microaerophilic)	0.2651	0.28[1]
Casamino acids (microaerophilic)	0.6734	Growth
Energetic metabolism		
CO_2 + Thiosulfate + Fe^{3+}	0.1266	Growth
CO_2 + Fe^{2+} + Oxygen	No growth	Growth
CO_2 + Hydrogen + Sulfur	0.2532	Growth
CO_2 + Hydrogen + Sulfite	0.06336	Growth
CO_2 + Sulfite + Oxygen	0.06147	Growth

[1] Converted from doubling time [h^{-1}] (Aguiar et al., 2004).

The specific growth rate of 0.28 h^{-1} for *S. azorense* Az-Fu1 under chemolithoautotrophic and microaerophilic conditions was measured by Aguiar (2004). Simulations using the *i*SS352 model yielded 0.26 h^{-1} and predicted the production of sulfate when elemental sulfur was added to the medium, as reported by Nakagawa (2005) (Supplementary Table 3.2). No other byproducts were predicted, which was expected as no information was found in the literature on the production of other metabolites. The

model agrees with the report from Nakagawa (2005), which determines that *S. azorense* Az-Fu1 can grow using casamino acids as carbon sources under microaerophilic and anaerobic conditions (Supplementary Table 3.3).

The energetic metabolism was also evaluated using the different electron donors and acceptors reported in the literature (Aguiar et al., 2004). Simulations predicted that elemental sulfur enhances growth under hydrogen oxidizing conditions, as reported for the *Hydrogenobacter* genus (Bonjour et al., 1986) and low growth rates were obtained when the pairs hydrogen + sulfite and sulfite + oxygen were used as electron donors and acceptors (Table 3.8). Growth was not predicted using ferrous iron and oxygen as electron donors and acceptors, respectively, while, under these conditions, the observed experimental growth was acknowledged to be very low (Aguiar et al., 2004).

Analysis of the potential for EPS production

As already stated, *S. azorense* Az-Fu1 can produce chemolithoautotrophically EPS under stress conditions (Lalonde et al., 2005). Although no information about EPS composition was found in the literature, genome annotation analysis suggests that this EPS can have in its composition cellulose (see section 3.3.1). Bacterial cellulose operons and production have been extensively studied (Römling, 2002; Kawano et al., 2011; Romling et al., 2013; Römling et al., 2015; Cacicedo et al., 2016; Hernández-Arriaga et al., 2019; Blanco et al., 2020) due to its unique characteristics and applicability. However, its production by a thermophilic organism has not been yet described. For this reason, allied to the fact that this production can be made chemolithoautotrophically, this discovery is of significant interest.

In order to better understand the metabolic capabilities of cellulose production under chemolithoautotrophic conditions, a metabolic engineering optimization supported by evolutionary algorithms was performed. The optimization process did not return any reaction or gene knock-out solution robust enough to increase cellulose production. An FVA analysis was then performed to understand the model solution space regarding cellulose production in different stress conditions (nitrogen, sulfur, and iron limitation conditions) (Table 3.9).

For each condition, the total consumption of the carbon source was imposed, and the specific growth rate was set to at least 10% of the specific growth rate obtained with pFBA simulation under chemolithoautotrophic conditions. Results showed that the highest value of cellulose production was

achieved at nitrogen limitation conditions. However, none of the tested conditions revealed a mandatory cellulose production coupled with biomass.

Table 3.9. FVA analysis of *i*SS352 model cellulose production capabilities. Total consumption of the carbon source was imposed, and the specific growth rate was set to at least 10% of the specific growth rate obtained with pFBA simulation under chemolithoautotrophic conditions. Uptake values are represented as negative and production values are presented as positive values.

Compound	Carbon Limitation	Nitrogen Limitation	Sulfur Limitation	Iron Limitation
CO_2 (mmol g_{DW}^{-1} h^{-1})	-12.00	-12.00	-12.00	-12.00
Cellulose (mmol g_{DW}^{-1} h^{-1})	**[0.0, 0.4500]**	**[0.0, 0.4950]**	**[0.0, 0.4189]**	**[0.0, 0.4500]**
μ (10%) (h^{-1})	0.2650	0.02650	0.02650	0.02650

When nitrogen limitation conditions were imposed on a pFBA analysis, the byproducts of the simulation predicted the secretion of acetate and/or bicarbonate (Table 3.10). Since no information was found in the literature on the production of side carbon components, the secretion of these metabolites was constrained. Cellulose was then predicted to be excreted in nitrogen-limited conditions when the flux of bicarbonate was restricted and almost reached the maximum production value when no bicarbonate was allowed to be produced. These results show the metabolic capability of *S. azorense* Az-Fu1 to produce cellulose under nitrogen-limiting conditions (stress). Nevertheless, these results should be experimentally validated.

Table 3.10. iSS352 prediction of byproduct production under N-limiting conditions and restriction of bicarbonate production.

Byproduct Production under N-limiting Conditions (mmol g_{DW}^{-1} h^{-1})			
HCO_3^- restriction	**100%**	**50%**	**0%**
Bicarbonate	0.000	5.400	10.80
Acetate	0.001983	0.001983	0.000
Cellulose	0.4499	0.2249	0.000

3.4 Conclusions

The main objective of this study was to reconstruct the first GSMM of the chemolithoautotrophic organism *S. azorense* Az-Fu1 to have insights into its metabolism, genetic adaptation to extreme environments, and to be capable of predicting and optimizing the production of relevant compounds with industrial interest. *S. azorense* Az-Fu1 genome annotation and metabolic analysis revealed the CO_2 fixation route being the rTCA, as for the majority of Aquificales members (Hügler & Sievert, 2010), and also incomplete PPP and EMP pathways. Sulfur metabolism was analyzed as well and the *S. azorense* Az-Fu1 genome showed the presence of a truncated Sox system, able to oxidize elemental sulfur, thiosulfate, and sulfite to sulfate, as indicated in literature (Aguiar et al., 2004).

*i*SS352 performed well on all studies to validate carbon source utilization under microaerophilic and anaerobic conditions, as well as on electron donor and acceptor utilization. The capability of *S. azorense* Az-Fu1 to use different electron donors and acceptors was validated by GSMM simulations and highlighted its metabolic versability as being capable of adapting to highly dynamic environmental conditions, such as extreme environments.

Further analysis of *S. azorense* Az-Fu1 genome functional annotation also revealed the presence of the main subunits of the bacterial cellulose operon and their regulators and the GSMM simulations showed the organism's metabolic capability to produce cellulose under nitrogen-limiting conditions. Some studies (Lalonde et al., 2005) have shown that *S. azorense* Az-Fu1 produces sufficient amounts of exopolysaccharides under stress conditions, although experimental validation must be performed to confirm whether cellulose production is naturally viable.

The lack of experimental quantitative data limits the spectrum of application of this model. However, given that *S. azorense* Az-Fu1 was identified within a natural microbial community (Sahm et al., 2013), and the fact that *i*SS352 is the first curated model for this species, it will be pivotal to study the organism's metabolic role in the microbial community, as well as using the huge potential of microbial community design in industrial biotechnology and discovery of new extremophilic enzymes.

3.5 Supplementary Material

Additional file in SBML format: /SS352.xml

Link: DesignOptimizationMicrobialCommunities/models/

Additional file in Excel format: Chapter3_Supplementary_Material.xlsx

Link: DesignOptimizationMicrobialCommunities/Data

Table S1 Annotation of the genes present in *S. azorense* Az-Fu1 model

Table S2 Biomass composition in mmol of molecules per gram of biomass. Molecular weight in green background cells was calculated using the fatty-acyl Coa as the R group in lipids. Amino acid's molecular weight does not include a water molecule. Nucleotides' molecular weight does not include diphosphate molecules.

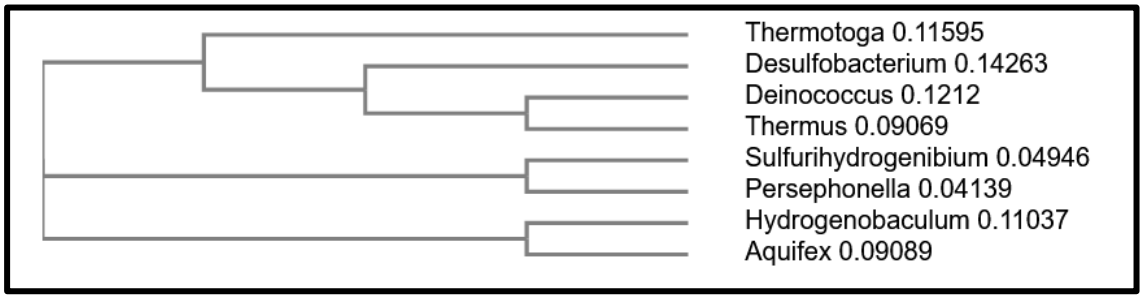
Table S3 Average lipid and fatty acid compositions.

Table S4 Cell wall composition.

Table S5 Genes included in the model

Table S6 Reactions included in the model, including Gene-Protein-Reaction associations.

Table S7 Metabolites included in the model



Supplementary Figure 3.1. Phylogenetic tree of *Sulfurihydrogenibium azorense* related genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbour-joining method.

Supplementary Table 3.1. Cellulose operon units for each main type and respective gene accession numbers.

Gene	**Type I** *Komagataeibacter xylinus*	**Type II** *Escherichia coli*	**Type III** *Methylobacterium extorquens*
bcsA / YhjO	AAA21884.1	NP_417990.4	WP_080518407.1
bscB / YhjN	AAA21885.1	NP_417989.1	WP_012252997.1
bcsC/ YhjL	AAA21886.1	YP_026226.4	———
bcsD	AAA21887.1	———	———
bcsQ/ YhjQ	———	WP_011310329.1	———
bcsZ / YhjM	———	NP_417988.1	WP_012252998.1
bcsE/ YhjS	———	NP_417993.1	———
bcsF/ YhjT	———	NP_417994.2	———
bcsG/ YhjU	———	NP_417995.1	———
bcsR/ YhjR	———	NP_417992.1	———
bcsH/ CcpA	AAA16970.1	———	———
bcsK	———	———	WP_012252999.1
bcsN	———	———	WP_012253000.1
bcsS	———	———	WP_003606498.1

Supplementary Table 3.2. Simulation results under chemolithoautotrophic and microaerophilic conditions.

Simulation Results			
Specific growth rate (μ): 0.2636 h^{-1}			
Consumption		**Production**	
Metabolite	**Uptake Value** ($mmol\ g_{DW}^{-1}\ h^{-1}$)	**Metabolite**	**Production Value** ($mmol\ g_{DW}^{-1}\ h^{-1}$)
CO_2	12.00	**Sulfate**	12.23
H_2O	4.286		
Sulfur	12.27		
Fe^{2+}	0.004437		
Ammonia	2.674		
O_2	9.518		
$H_3PO_4^-$	18.90		

Supplementary Table 3.3. Simulation results under mixotrophic and anaerobic conditions.

Simulation Results			
Specific growth rate (μ): 0.1069 h^{-1}			
Consumption		**Production**	
Metabolite	**Uptake Value** ($mmol\ g_{DW}^{-1}\ h^{-1}$)	**Metabolite**	**Production Value** ($mmol\ g_{DW}^{-1}\ h^{-1}$)
Amino Acids	< 0.1	**H^+**	4.190
Sulfur	0.001790	**H_2O**	1.660
Fe^{2+}	0.001790	**Ammonia**	0.4677
$H_3PO_4^-$	0.1106		

Chapter 4

Microbial Community Simulation Methods

"The lone wolf dies but the pack survives."

George R. R. Martin

Microbial communities directly affect surrounding environments and are an important part of all biological processes. Thus, the study of microbial communities' behavior and composition has been proven to be useful in areas such as biotechnology, environmental, and human health. However, the overall understanding of microbial communities' interactions and dynamics remains a challenge.

Synergies between computational methods and genome-scale metabolic models have been explored in the last years, as a way to unravel community interactions and behavior, as demonstrated by the numerous simulation methods developed for application in the context of microbial communities.

Here, different steady-state simulation methods applied to microbial communities have been used to model the well-established nitrification bioprocess catalyzed by *Nitrosomonas euroapaea* and *Nitrobacter vulgaris*. The different methods' performances were compared to assess which method(s) should be used in a specific community-level context.

The available simulation methods, with application to microbial communities, revealed good phenotypic behavior predictions. Each of the simulation methods exhibited strengths and weaknesses. Hence, to better predict the communities' behavior, it is recommended the use of various simulation methods, whenever possible.

4.1 Introduction

During the last decades, genome-scale metabolic models (GSMMs) have proven to be a valuable tool in Systems Biology, with outstanding biotechnological applications (Gu et al., 2019), predicting cellular behavior under different genetic and environmental conditions (Woolston et al., 2013), and reducing time and costs implicated in experimental tasks. In the last decade, the development of high-throughput techniques and the scientific desire to understand microbial communities' behavior and capabilities, through computational and mathematical modeling approaches, turned communities' GSMMs into the new frontier in the field. The emergence of numerous simulation methods developed to be applied in a microbial community context (Colarusso et al., 2021), with proven applicability in industrial, environmental biotechnology, and human health environments (García-Jiménez et al., 2021) support this claim. The currently available simulation methods, developed to perform simulations with microbial communities GSMMs, can be classified as steady-state, time-course, and spatial-temporal (Figure 4.1).

Steady-state methods allow for predicting individual growth rates and cross-feeding interactions in a situation of community equilibrium. These methods are based on the well-established Flux Balance Analysis (FBA) (Varma et al., 1994) method for individual species. This constraint-based approach is usually applied to biochemical networks, allowing the prediction of biological phenotypes. The steady-state reaction flux distribution in a network is determined by Linear optimization of biologically and ecologically meaningful objective(s) function(s), subject to a set of underlying constraints (Raman et al., 2009). However, these methods are characterized by a large uncertainty in the space of optimal solutions.

Time-course methods, based on dynamic flux balance analysis (dFBA) (Mahadevan et al., 2002), account for the temporal variation in metabolite concentrations and species abundance, allowing for simulation of the response to changes over initial conditions and other environmental perturbations. However, this approach requires the characterization of the substrate uptake kinetics for consumed metabolites (either from the growth medium or through cross-feeding). The impractical nature of such a massive *in vitro* characterization results in the adoption of default parameter values for all species and compounds, which limits the predictive ability of these methods.

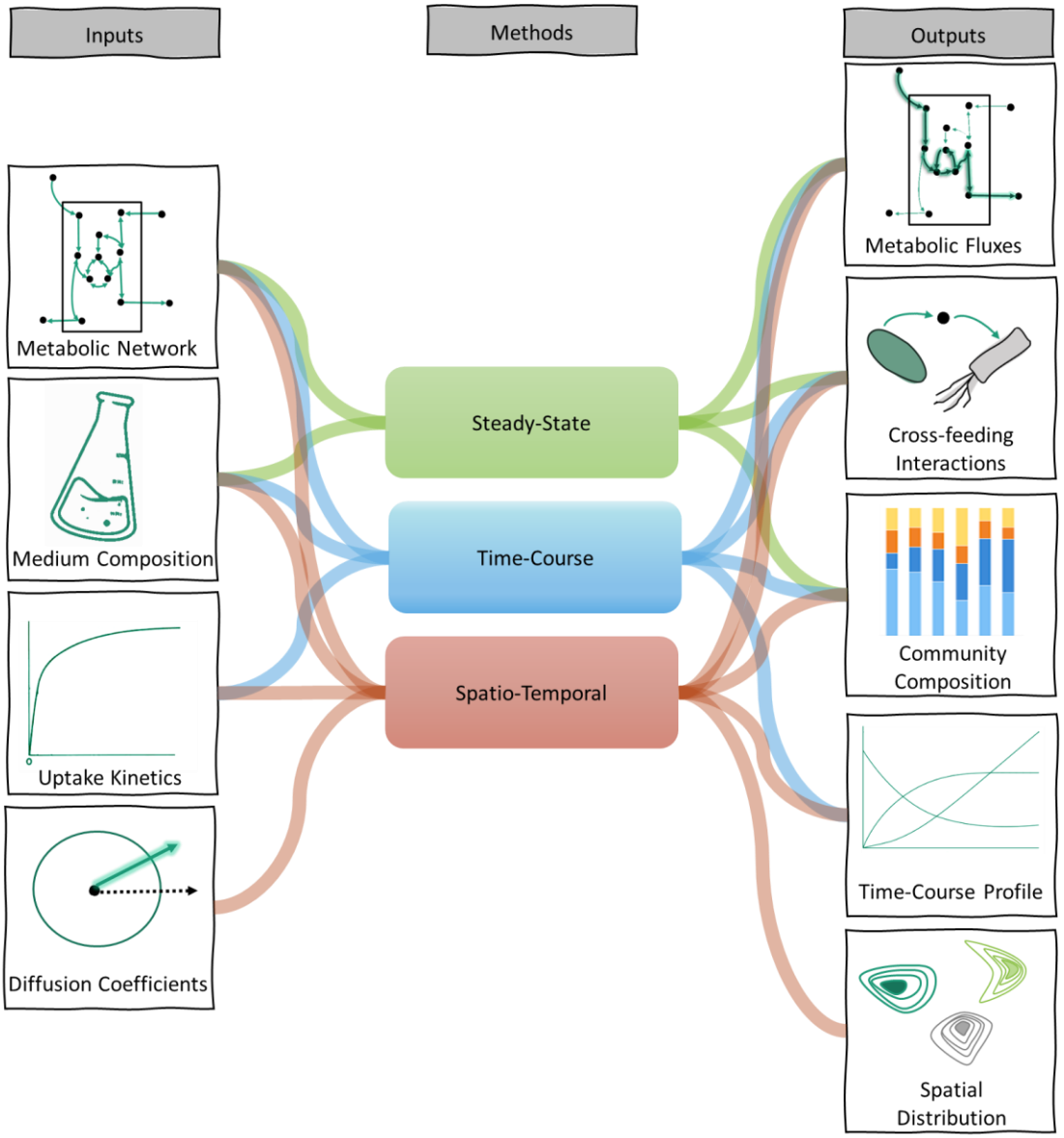


Figure 4.1 Summary of the main features of the current simulation methods with application to microbial communities.

Spatiotemporal methods add a spatial component to the temporal one. These methods are especially relevant in situations of non-heterogeneous or spatially segregated environments such as biofilm colonies or the tract along the human gut, where nutrient diffusion and access to resources play a role in community assembly and potential for cross-feeding interactions. In addition to the parameters on uptake kinetics mentioned before, these methods also require the specification of diffusion rates for cells and metabolites.

Besides having common grounds, as all are based on the FBA method for single organisms, these methods differ in the number and kind of inputs needed and have different levels of simulation detail and

objective functions. Hence, each method can produce a variety of output information, enabling the handling of microbial communities' GSMMs under a wide range of different applications. However, the increasing level of simulation detail simultaneously requires more experimental data for model setup and, naturally, more computational power, which is only possible for small communities composed of well-characterized microorganisms.

Steady-state simulation methods benefit from not requiring parameterization, thus being scalable for large microbial communities. Also, they are compatible with flux variability analysis (FVA) (Gudmundsson et al., 2010), flux coupling analysis (David et al., 2011), and random flux sampling (Schellenberger et al., 2009). Hence, developing steady-state methods for the analysis of microbial communities (Figure 4.2) (Colarusso et al., 2021) has been the main focus of research. Having a great variety of methods, the most difficult task is understanding which method(s) should be applied in a specific context, considering available experimental data, microbial community complexity, and particular final purpose(s). Understanding the differentiating features allows selecting the best strategy for a specific microbial community.

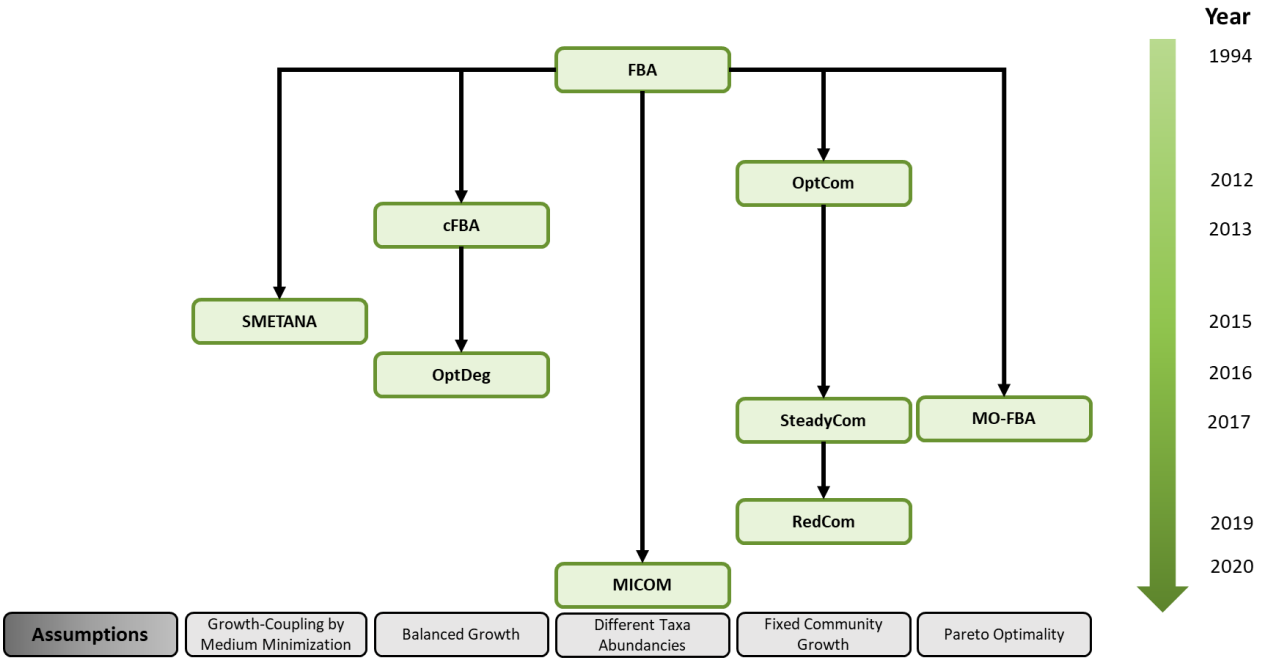


Figure 4.2. Lineage of steady-state simulation methods with application to microbial communities. Each branch of the tree represents different simulation method assumptions.

Here, a systematic analysis of the performance of the most used steady-state simulation tools for predicting microbial communities' behavior is presented. This analysis will cover key features that allow properly selecting the best method, that fits not only the purpose of the analysis but also available experimental data. For that, the most promising simulation methods have been installed and applied to the well-known case of the biological nitrification process catalyzed by the symbiotic interaction between *Nitrosomonas europaea* and *Nitrobacter vulgaris* (Ilgrande et al., 2018). The GSMMs for these bacteria have been reconstructed and manually curated with experimental data within our research group (Cruz, 2018; Raposo, 2017).

4.1.1 Current state of steady-state simulation methods with application to microbial communities

A brief description of the current steady-state simulation methods used in a microbial community context is presented hereafter, highlighting the main features of each one.

Flux Balance Analysis

Although its formulation was developed for predicting the phenotypic behavior of single organisms, Stolyar et al pioneered the use of FBA to simulate the behavior of microbial communities, namely in studying a community of two methanogenic species (Stolyar et al., 2007). The steady-state model must be created before the FBA simulation by merging the stoichiometric matrices of any number of individual organisms. FBA can determine steady-state flux distribution throughout the community network, restricted to a defined medium and corresponding uptake rates of the exchanged metabolites. Hence, by analyzing flux distribution across the network, community, and individual growth rates, as well as cross-feeding interactions in the community can be easily established. Although the definition of an objective function for microbial communities is still ambiguous, FBA and the variant parsimonious enzyme-usage flux balance analysis (pFBA) (Lewis et al., 2010) have been successfully used using growth maximization of the entire microbial community as the objective function. Examples of applications include understanding microbe-microbe and host-microbe interactions (Bordbar et al., 2010; Heinken et al., 2013) and the production of biocompounds in diverse settings, such as the production of vitamin C (Ye et al., 2014) or 1,3-propanediol (Bizukojc et al., 2010).

FBA can be used under diverse platforms, such as MATLAB® through the COBRA toolbox (Heirendt et al., 2019), the python packages COBRApy (Ebrahim et al., 2013) and REFRAMED, or user-friendly frameworks such as Optflux (I. Rocha et al., 2010) or KBase (Arkin et al., 2018).

OptCom

OptCom (Zomorrodi et al., 2012) is an FBA framework that relies on a bi-level optimization structure. The inner level maximizes the growth of individual organisms, while the outer level maximizes total community fitness. The underlying assumption is that the community will optimize the usage of available resources while not compromising the individualistic objectives of its members. If more constraints are provided, OptCom is also able to compute cases where organisms do not present their maximum individual growth, exhibiting cooperative behavior instead. This approach can use any number of GSMMs and returns strain ratio, substrate uptakes, and secretion rates showing applicability in inferring the syntrophic association through hydrogen between *Desulfovibrio vulgaris* and *Methanococcus maripaludis* (Zomorrodi et al., 2012) and the production of butyrate for cancer prevention (El-Semman et al., 2014). OptCom has been implemented in the COBRA toolbox (Heirendt et al., 2019) and integrated into the MICOM Python framework (Diener et al., 2020).

cFBA

Community Flux Balance Analysis (cFBA) (Khandelwal et al., 2013) is a computational method whose primary focus is elucidating the metabolic capabilities of a community and understanding metabolic interactions. This approach is a direct translation of FBA to microbial communities requiring balanced growth and postulating a single objective. Balanced growth implies that the entire community is in steady-state and thereby all metabolites (intra- and extracellular) are at a steady-state level as well.

Although the problem definition of cFBA is non-linear, when biomass fractions are considered a variable, by fixing the individual biomass values, a linear programming problem arises, thus providing a result that identifies the optimal specific flux values. Notice that the community growth rate may not match the maximum growth rate of any individual organism. However, when considering a mutualistic interaction between growing organisms, to preserve the cross-feeding metabolites production and consumption rates equilibrium, organisms must grow at an equal pace. Along with the abundance of all

species present in the community, cFBA predicts fluxes distribution, growth rates, and exchange fluxes between the microbes and the environment. Hence, the GSMMs of the organisms present in the community must be merged before simulation. The cFBA method is applied essentially to microbial communities exposed to constant environmental conditions, such as the ones involved in bioremediation, wastewater treatment, or in laboratory settings.

SMETANA

Species Metabolic Interaction Analysis (SMETANA) (Zelezniak et al., 2015), a Python-based command line tool, uses mixed-integer linear programming to estimate the interaction potential of the species in a microbial community and returns the probability of inter-species metabolite exchange capacity. SMETANA can be applied with few microbial community information, such as species identity and genome sequence. GSMMs for each species are automatically reconstructed and assembled into a community model. This step can be ignored if GSMMs of all species in the community are available.

SMETANA calculates two scores to predict the species' interaction potential: the Metabolic Resource Overlap (MRO) and the Metabolic Interaction Potential (MIP). The MRO is an intrinsic property of any community and is defined by the maximum possible overlap between the minimal nutritional requirements of all member species. MIP represents the tendency of community members to exchange metabolites and is given by the maximum number of essential nutritional components that a community can provide for itself, through the interspecies metabolic exchange. Thus, the higher its value, the higher the probability of community members benefiting from metabolite production from other members. MRO and MIP represent opposite circumstances. While the first indicates the predisposition to competition between the organisms, as both require the same metabolite(s) from the environment, MIP represents the tendency of the community's organisms to rely on each other, not being able to grow on their own.

SMETANA calculates three other scores to predict potential inter-species interactions, Species Coupling Score (SCS), Metabolic Uptake Score (MUS), and Metabolite Production Score (MPS). The SCS score measures the growing dependence of species A in the presence of species B, whereas the MUS score measures the growing dependence of species A in the presence of metabolite *m*, provided by the other members of the community. MPS is a binary score that reflects the ability of a given species to produce metabolite *m*. Finally, the SMETANA score that evaluates the growth dependency of species A on metabolite *m* produced by species B is calculated as a product of the scores SCS, MUS, and MPS.

SMETANA has been successfully applied to communities composed of a high number of species, spanning habitats as diverse as soil, water, and human gut (Machado et al., 2021; Zelezniak et al., 2015; Zorrilla et al., 2021) as well as in a three-species stable synthetic community, with nitrogen overflow capabilities (Ponomarova et al., 2017).

OptDeg

OptDeg (Koch et al., 2016), uses as input a community model for predicting growth and flux rates. This approach applies the previously described concept of balanced growth, in which all organisms in the community grow at the same growth rate. For predicting the community composition, a hierarchical optimization approach that consists of two objective functions is employed. Firstly, the maximization of the community's growth rate is optimized, followed by the maximization of each organism's growth rate under optimal utilization of substrates, which is defined by the authors, as biomass yield. For the quantification of the overall biomass yield, an optimality degree constant (OptDeg) was introduced. This constant is calculated as the quotient of the maximal community growth rate and the individual minimal expected growth rate if the organism uses its substrates optimally. Thus, for OptDeg = 1, all the community species grow at the maximum specific growth rate and maximum biomass yields. OptDeg has been implemented in *CellNetAnalyzer*, a MATLAB package for structural and functional analysis of metabolic and signaling networks (Klamt et al., 2007).

SteadyCom

SteadyCom (Chan et al., 2017) is an FBA-based method that predicts the metabolic flux distribution and relative abundance of each species in a community, assuring that a steady state is imposed. The novelty of this implementation relies on the imposition of a time-average constant growth rate to ensure the co-existence and stability of all members of the community. For instance, when a single organism is considered, the biomass flux is normalized by the organism's rates of consumption or production. However, when multiple organisms are growing, there is not a constant growth rate for all microbes. Therefore, the fastest-growing organism can outgrow the rest of the community members. To avoid such situations, SteadyCom imposes a steady-state condition, that includes a restriction to force zero flux through an organism with zero abundance. As the sub-problems to be solved are independent of

the number of organisms in the community, SteadyCom is highly scalable for large microbial communities, for instance predicting the varying gut microbiota composition (Chan et al., 2019). This algorithm is available in the COBRA toolbox (Heirendt et al., 2019) and the REFRAMED Python package.

MO-FBA

Multi-objective flux balance analysis (MO-FBA) and multi-objective flux variability analysis (MO-FVA) (Budinich et al., 2017), analyze microbial communities by merging individual GSMMs in a compartmentalized community GSMM. Unlike OptCom, which relies on the maximization of the community growth or cFBA that assumes individual balanced growth, MO-FBA describes all optimal solutions, for each strain in the ecosystem, in the sense of the Pareto optimality (Nagrath et al., 2007). Therefore, all solutions can be comprised of system objectives for microbial communities, not requiring complementary restrictions, and maintaining the set of constraints linear, allowing the study of large natural ecosystems. Hence, optimal total growth can be achieved when all members of a community grow below their theoretical maximum individual rates, as when one of the microbial community members decreases its growth rate, more resources are available for other members. The use of the newly available resources can increase the value of the objective functions of the other guilds, increasing the global growth rate.

This tool has a mixed MATLAB and Python implementation and uses BENSOLVE (Löhne & Weißing, 2017) solver to compute a set of directions and points describing the image of the efficient points.

RedCOM

RedCom (Koch et al., 2019), such as SteadyCom, imposes a fixed community growth rate to a known constant value, ensuring that the optimization process is linear. The uniqueness of this approach relies on creating a reduced model of the microbial community, performed by calculating elementary flux vectors (EFVs), which are mandatory to consider constraints such as flux bounds and maintenance coefficients. Individual GSMMs are reduced by eliminating reactions irrelevant to the individual organism's growth. If the single organisms' exchange fluxes in the community models overcome an imposed minimality criterion, the solutions are discarded, and the respective reactions are removed. As the created

models are smaller, smaller ranges of feasible community compositions and exchange fluxes are calculated, ensuring that unrealistic solutions in the community model are excluded.

MICOM

MICOM (*mi*crobial *com*munities) (Diener et al., 2020) has been developed as a COBRApy Python package (Ebrahim et al., 2013). MICOM's mathematical formulation considers two different classifications for growth rates: individual growth rates, which estimate the growth rate of a single organism, and community growth rates, which represent the growth of the entire community. Unlike OptCom and SteadyCom, MICOM does not assume that all taxa in a community grow at the same rate, requiring taxa abundances as input. To represent the community, a particular abundance for each organism (in g_{DW}) is considered and each organism is allocated to an external compartment that represents the community environment. Given a particular abundance of an organism, a sub-model in the whole community, MICOM scales the whole community's internal exchange fluxes to the respective organism abundance.

4.1.2 Case study: Nitrification bioprocess by *Nitrosomonas europaea* and *Nitrobacter vulgaris*

The two-step nitrification process (Figure 4.3) can be accomplished by chemolithoautotrophic bacteria. In the first moment ammonia (NH_3) is oxidized to nitrite (NO_2^-) by ammonia-oxidizing bacteria, such as Nitrosomonas europaea, and NO_2^- is subsequently oxidized to NO_3^- by nitrite-oxidizing bacteria (NOB) (Fowler et al., 2013).

The known AOB and NOB representatives are elements of Nitrosomonas spp. and Nitrobacter spp., respectively (Bagchi, Biswas, & Nandy, 2012). Indeed, some experimental (Grunditz et al., 2001) and modeling (Mellbye et al., 2018; Perez-Garcia et al., 2016) approaches have been pursued to understand the interaction of AOB and NOB organisms and their role in the nitrification process. As the role of each organism is well-established and, to a certain extent, validated, the GSMMs of *N. europaea* and *N. vulgaris* will be used to assess the abovementioned simulation methods with application to microbial communities.

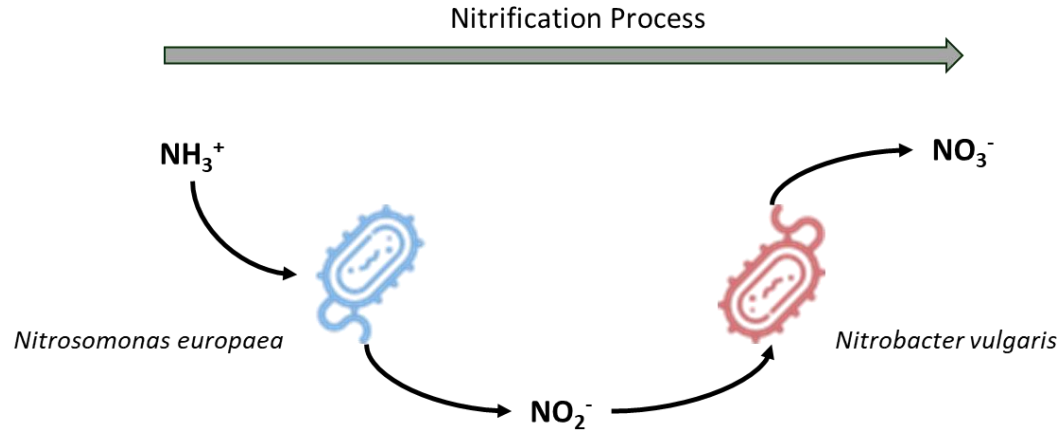


Figure 4.3. Know interactions during the nitrification process catalyzed by the bacteria *Nitrosomonas europaea and Nitrobacter vulgaris*. In the first step, *Nitrosomonas europaea* consumes NH_3^+ and excretes NO_2^- that is, in the second step consumed by *Nitrobacter vulgaris* which excretes nitrogen in the form of NO_3^-.

4.2 Methods

4.2.1 Genome-scale metabolic models

GSMMs for both *N. europaea* and *N. vulgaris* were previously reconstructed and manually curated using experimental data (R. Cruz, 2018; Raposo et al., 2017). Steady-state simulation methods with application to microbial communities were used to predict growth and flux rates, community composition, and predict known and possible interactions.

4.2.2 Simulations

All steady-state simulation methods used in this assessment, as well as related information and main features, are summarized in Table 4.1.

Most methods are implemented in various platforms, such as Matlab® or Python. For consistency, all simulation methods were applied using their Python (version 3.6) implementation and CPLEX v12.8 (academic license) as the solver, under the PyCharm (Educational License) integrated development environment. Although FBA, and its variant pFBA, were not originally developed to simulate microbial communities, its successful application to various community case studies motivated its inclusion in this study.

Table 4.1. List of tools used for analyzing microbial community models and their main features. All tools were run under their Python implementation.

Method	Framework	Output	Method Implementation	Assumptions	References
FBA (pFBA)	REFRAMED	Flux and Growth rates	FBA with maximization of the total community rate	Equal individual growth rate	(Varma et al., 1994)
OptCom	MICOM	Flux and Growth rates	Bilevel FBA with community-level multi-objective function	Fixed community growth rate	(Zomorrodi et al., 2012)
SMETANA	command line	Probability of Interactions	Iterated mixed integer linear problem with alternative solutions	Growth coupling induced by medium minimization	(Zelezniak et al., 2015)
SteadyCom	REFRAMED	Flux rates and Biomass ratios	FBA iterated with an outer loop to find the maximum growth rate	Fixed community growth rate	(Chan et al., 2017)
MICOM	MICOM	Flux and Growth rates; Interactions; Community Composition	Maximization of the community growth rate and minimization of the sum of squares of individual growth rates at a fraction of the maximum community growth rate	Different individual growth rate (taxa abundancies required as input)	(Diener et al., 2020)

A few of the tools described above exhibited problems associated with installation or execution rendering their evaluation infeasible. Although having been published, OptDeg and RedCom were not available for download. Likewise, OptCom's Matlab® implementation was also not available, thus the Python implementation in MICOM's framework was used. The cFBA method, which requires Python 2, exhibited compatibility issues and was consequently excluded from the assessment. Finally, the MO-FVA's reliance on the BENSOLVE solver, which returned errors during installation, was also excluded from this assessment.

All Python scripts, input/output auxiliary files, and GSMMs used in this work are available on GitHub at SophiaSantos/DesignOptimizationMicrobialCommunities.

4.3 Results and Discussion

A qualitative evaluation was performed to assess tools able to simulate the behavior of microbial communities. A list of features considered relevant for assessing the method's performance was created and a score was assigned to each method (1: unsatisfactory, 5: outstanding). These features are related to software performance, ease of use, and adherence to common data standards. Other features analyzed include the predicting capabilities when compared with available experimental data and the output information for each method, which is important for selecting the best for the desired community-specific application. The full discussion of this assessment is available in section 4.3.4 (Figure 4.4).

The analysis of the nitrification process catalyzed by the bacteria *N. europaea* and *N. vulgaris* was selected for the case study, due to the availability of GSMMs manually curated and validated with experimental data (Cruz, 2018; Raposo, 2017). Additionally, the interactions associated with the nitrification process between these bacteria are well known (Figure 4.3), and therefore the prediction outputs from the simulation methods applied can be compared to experimental information. Basic validation of each GSMM was performed before applying the microbial community to these methods, using pFBA.

4.3.1 Environmental conditions

The validation of the single organism GSMMs was performed using a minimal medium, described elsewhere (Cruz, 2018; Raposo, 2017), and is presented in Table 4.2.

Table 4.2. Minimal medium composition used to run single organism pFBA simulations using GSMMs of *N. europaea* and *N. vulgaris*, respectively. Uptake rates are shown in $mmolg_{DW}^{-1}h^{-1}$.

Minimal Medium Constraints			
Nitrosomonas europaea		***Nitrobacter vulgaris***	
Metabolite	**Lower bound**	**Metabolite**	**Lower bound**
CO_2	0.3000	**CO_2**	0.01000
Orthophosphate	1000	**Orthophosphate**	1000
SO_4^{2-}	1000	**SO_4^{2-}**	1000
Ammonia	1000	**NO_2^-**	1000
O_2	1000	**O_2**	1000
Fe^{2+}	1000	**Fe^{2+}**	1000
		Ethanol	0.1000

All medium components were allowed to enter the system unconstrained, except for carbon sources. In the case of *N. vulgaris* growth experiments, CO_2 was used as a carbon source and ethanol as a carbon source and energy source (R. Cruz, 2018), and therefore that medium was replicated for these simulations. *N. vulgaris* can grow using only CO_2 or ethanol, however, biomass production is higher in the presence of ethanol (R. Cruz, 2018). The simulations were performed under aerobic conditions, thus O_2 and ferrous iron were supplemented in both cases.

The minimal medium composition required for the microbial community simulations (Table 4.3) was adapted from the minimal medium of each organism. As for single organism simulations, all medium components were allowed to enter the system unconstrained, except for the carbon source, which in the case of the co-culture where the uptake rate value for CO_2 was set to the sum of each organism's carbon source uptake rate, -0.4210 $mmolg_{DW}^{-1}h^{-1}$, (*N. europaea* - 0.31 $mmolg_{DW}^{-1}h^{-1}$ of CO_2, *N. vulgaris* - 0.011 $mmolg_{DW}^{-1}h^{-1}$ of CO_2 and 0.1 $mmolg_{DW}^{-1}h^{-1}$ of ethanol), once *N. vulgaris* is able to grow using only CO_2.

Table 4.3. Minimal medium composition used to run microbial community simulations. Uptake rates are shown in $mmolg_{DW}^{-1}h^{-1}$.

Minimal Medium Constraints	
Metabolite	**Lower bound**
CO_2	-0.4210
Orthophosphate	-1000
SO_4^{2-}	-1000
Ammonia	-1000
O_2	-1000
Fe^{2+}	-1000

4.3.2 Steady-state simulations of single-organisms

The validation process of the GSMMs of *N. europaea* and *N. vulgaris* involved comparing simulation results to previous studies (Cruz, 2018; Raposo, 2017), to understand the nitrogen uptake and excretion processes.

The validation step revealed an inaccurate flux distribution throughout the oxidative phosphorylation pathway of the *N. vulgaris* GSMM, which is a key pathway under aerobic conditions. After the manual curation of the pathway (Supplementary Table S1), the simulation results (Table 4.5) were in agreement

with the experimental data reported by Cruz (2018). Results of the simulation of *N. europaea* GSMMs (Table 4.4) were in agreement with the information available in the literature (Raposo, 2017).

Table 4.4. *N. europaea* model validation using pFBA as simulation method. Specific growth rate (h^{-1}), uptake (Consumption), and export (Production) rates ($mmolg_{DW}^{-1}h^{-1}$) are presented.

Nitrosomonas europaea			
Specific growth rate: 0.007528 h^{-1}			
Consumption		**Production**	
CO_2	0.3000	**H^+**	0.6654
Orthophosphate	0.002598	**H_2O**	1.182
SO_4^{2-}	0.001556	**NO_2^-**	0.6606
Ammonia	0.7265	**Urea**	0.001338
O_2	0.6624		
Fe^{2+}	0.00001359		

Table 4.5. *N. vulgaris* model validation using pFBA as simulation method, after manual curation of the oxidative phosphorylation pathway. Specific growth rate (h^{-1}), uptake (Consumption), and export (Production) ($mmolg_{DW}^{-1}h^{-1}$) rates are presented.

Nitrobacter vulgaris			
Specific growth rate: 0.03179 h^{-1}			
Consumption		**Production**	
CO_2	0.0100	**H^+**	0.8133
Orthophosphate	0.0005715	**NO_3^-**	0.2612
SO_4^{2-}	0.0004815		
NO_2^-	0.3134		
O_2	0.06714		
Fe^{2+}	0.0002978		
Ethanol	0.250		

N. europaea was able to use ammonia and almost totally convert it into NO_2^- (91% of nitrogen conversion) as reported in the literature, whereas *N. vulgaris* showed a slightly lower proportion of conversion of NO_2^- into NO_3^- (83% of nitrogen conversion), yet consistent with experimental results (Cruz, 2018) when no ammonia is supplemented to the medium and NO_2^- is also used as a nitrogen source for growth.

4.3.3 Steady-state simulations of community

Each method's features were analyzed according to method-specific outputs. These outputs include 1) the organism's individual growth rate within the community; 2) carbon source and ammonia consumption; and 3) cross-feeding metabolites, namely NH_3^+, NO_2^-, and NO_3^-. A summary of all simulations is provided in Table 4.6.

4.3.3.1 FBA

Initially, a community model was created using the REFRAMED Python package. When applying REFRAMED's *Community* function, providing the individual GSMMs for each organism as input, a compartmentalized model is returned. Given the fact that the biomass equation on the community model is defined as the sum of the individual biomass equations at the same proportion when applying the FBA method to microbial communities, all organisms contribute equally to the community biomass. Thus, to maintain the steady-state balance, all specific growth rates are equal. In this specific case (Table 4.6), the results show that the specific growth rate of each organism in the co-culture (0.0077 h^{-1}) mimics the slower-growing organism, *N. europaea* (0.007528 h^{-1}).

Table 4.6. Microbial community, composed of *N. europaea* and *N. vulgaris*, simulation using pFBA, SMETANA, SteadyCom, MICOM, and OptCom as simulation methods. When available, specific growth rate (h^{-1}), uptake (Consumption), export (Production) ($mmolgDW^{-1}h^{-1}$), and interaction rates are presented. Consumed metabolite rates are represented as negative and produced metabolite rates are represented as positive. N.vu. – *Nitrobacter vulgaris*, N.eu. – *Nitrosomonas europaea*, Com – Community. n.a. – Data not available. *Abundancy value used as input for the simulation method.

		Experimental			**pFBA**			**SMETANA**			**SteadyCom**			**MICOM**			**OptCom**		
		N.vu.	**N.eu.**	**Com**	**N.vu.**	**N.eu.**	**Com**	**N.vu.**	**N.eu.**	**Com**	**N.vu.**	**N.eu.**	**Com**	**N.vu.**	**N.eu.**	**Com**	**N.vu.**	**N.eu.**	**Com**
Specific Growth Rate		n.a.	n.a.	0.0147	0.0077	0.0077	0.0154	n.a.	n.a.	n.a.	0.0243	0.000199	0.0244	0.0150	0.0075	0.0151	0.0138	0.0075	0.0139
Metabolite Consumption/Production	CO_2	n.a.	n.a.	n.a.	-0.124	-0.297	-0.421			n.a.	-0.403	-0.0177	n.a.	-0.195	-0.0441	-0.245	n.a.	n.a.	-0.239
	NH_4	n.a.	n.a.	n.a.	-0.0308	- 1.22	-1.25	Producing	Consuming	n.a.	-1.701	-0.100	n.a.	-0.824	-0.236	-1.06	n.a.	n.a.	-1.06
	NO_2^-	Consuming	Producing	n.a.	-1.15	1.15	0	Consuming	Producing	n.a.	1.700	-1.700	n.a.	0	0.999	0.999	n.a.	n.a.	0.999
	Urea	n.a.	n.a.	n.a.	0	0.00128	0.00128			n.a.	2.05×10^{-5}	0	n.a.	0.145	0	0.145	n.a.	n.a.	0.143
	NO_3^-	n.a.	n.a.	n.a.	1.15	0	1.15			n.a.	0	1.700	n.a.	0	0	0	n.a.	n.a.	0
Abundancy		n.a.	n.a.		0.5*	0.5*		n.a.	n.a.	n.a.	0.99	0.01		0.99*	0.01*		0.99*	0.01*	

The analysis of the simulation results for the exchange fluxes demonstrates that, as reported in the literature, the microbial community converts NH_{3}^{+} into NO_{3}^{-}. Moreover, the proportion of conversion is similar to the individual organisms (92% of nitrogen conversion), using CO_2 as a unique carbon source.

Furthermore, the interacting reactions' fluxes (Table 4.6), show that the uptake of CO_2 is analogous to the uptake rate of carbon source for each organism, where *N. europaea* is consuming 0.31 $mmolg_{DW}^{-1}h^{-1}$ of CO_2 growing in isolation (Table 4.4) and -0.297 $mmolg_{DW}^{-1}h^{-1}$ of CO_2 growing in co-culture with *N. vulgaris* (Table 4.6). Although the total CO_2 uptake is similar, the growth rate of *N. vulgaris* within the community is 10-fold smaller, which was already reported in the literature (R. Cruz, 2018) which states that *N. vulgaris* has smaller growth rates when CO_2 is the sole carbon source.

As shown in Table 4.6 and the literature (Cruz et al., 2018; Grunditz et al., 2001; Mellbye et al., 2018), when NH_{3}^{+} is used by *N. vulgaris* as nitrogen source, all NO_{2}^{-} excreted by *N. europaea* is consumed by *N. vulgaris*, demonstrating the biological efficiency of the nitrification process catalyzed by these organisms.

4.3.3.2 SMETANA

SMETANA estimates the interaction potential of the species in a microbial community and returns the probability of inter-species metabolite exchange. No growth and flux rates are directly obtained through SMETANA.

For the case study of the community composed of *N. europaea* and *N. vulgaris*, a pre-simulation step was required. The individual GSMMs were reconstructed using *merlin* (Dias et al., 2018), which relies on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2016), for metabolites and reactions identifiers. However, SMETANA requires BiGG metabolite identifiers, thus those exchange metabolites identifiers were converted. The SMETANA potential inter-species interactions are available in Supplementary Tables S2 and S3. Only inter-species interactions with values greater than 0.4 were considered.

While analyzing possible inter-species interactions, as expected, and reported by literature, *N. europaea* acts within the community as a donor of nitrogen sources. Surprisingly, NH_{3}^{+} seems to have a high probability of being the nitrogen source shared by the organisms. Indeed, when simulating individual growth, when NH_{3}^{+} is added to the medium, *N. vulgaris* rather uses it as a nitrogen source to produce

biomass components (R. Cruz, 2018). SMETANA also predicted a high dependence on glycine, with a SMETANA score of 0.68 (data not shown), which can act as a carbon source and has a relevant role in assimilating one-carbon compounds through the reductive glycine pathway of chemolithoautotrophic organisms, as discussed in previous studies (Sánchez-Andrea et al., 2020; Yishai et al., 2018). However, experimental validation is required to confirm this possibility.

When analyzing the species interaction potential, SMETANA's results show no growth dependence on each other, indicating that both organisms can grow independently (MIP score = N/A). However, the resulting MRO score of 0.6 indicates the tendency to competition between the organisms for the main elemental sources, namely the nitrogen sources.

4.3.3.3 SteadyCom

SteadyCom predicts the metabolic flux distribution and relative abundance of each species in a community. Using the same environmental conditions as before (Table 4.3), SteadyCom predicted the growth of the two organisms, with a community growth rate of 0.02441 h^{-1}. As shown in Table 4.6, despite small numerical differences, the metabolites' interaction profile and respective ratios are quite similar to the experimental data. Importantly, this prediction shows, as predicted by the pFBA method and as reported in the literature (Cruz et al., 2018; Grunditz et al., 2001; Mellbye et al., 2018), that all NO_2^- excreted by *N. europaea* is consumed by *N. vulgaris*.

SteadyCom also calculates relative abundances for all organisms in the microbial community. Regarding this case study, SteadyCom's results are corroborated by individual growth experiments, which show that *N. europaea* grows at a much lower rate than *N. vulgaris* resulting in relative abundancies of 0.01 and 0.99, respectively.

The predicted growth rates using this method are in the same range for *N. vulgaris* when compared with isolated growth, and for *N. europaea* 10-fold lower when compared to isolated growth under the same conditions. Within a community, microorganisms' growth rates' perturbations are expected when organism abundances are significantly distinct, as seems to be the case for this community. However, these results cannot be assessed as additional experimental trials are required.

4.3.3.4 MICOM

MICOM mathematical formulation also considers two different classifications for growth rates: individual growth rates and community growth rates. However, MICOM's implementation does not assume a fixed community growth, thus requiring the community 's taxa abundance.

The first analysis was performed using the same environmental conditions as before and setting identical abundancies for both organisms, such as defined to run the pFBA simulation. MICOM's simulation results showed that in these conditions *N. europaea* is unable to grow (data not shown).

MICOM predicts a given microbial community's growth by fixing the minimum for all community members (equal or different for each community member). Therefore, when fixing the minimum growth rate to the slowest-growing organism in this community (*N. europaea* - a growth rate of 0.0075 h^{-1}), MICOM predicts the growth of both organisms (data not shown). However, the growth rates, as well as uptake and withdrawal rates, are 10-fold and 100-fold greater than the ones predicted with the previous methods. These results are probably associated with the fact that MICOM implementation scales the whole community's internal exchange fluxes to the respective organism's abundance.

To prove this hypothesis, the organism's input abundances were adjusted to the abundance values predicted by SteadyCom (0.01 for *N. europaea* and 0.99 for *N. vulgaris*). The results corroborate the suggested hypothesis, as the growth rates and uptake/production rates are in the same range of the pFBA and SteadyCom methods' predictions (Table 4.6)

However, in both predictions, the nitrogen flow through the networks (Supplementary Table S4) does not match the previous predictions, nor the literature on the nitrification process. Although NH_3^+ is consumed and NO_2^- is produced by *N. europaea*, *N. vulgaris* does not consume the produced NO_2^- and consequently no NO_3^- conversion is achieved by the community.

4.3.3.5 OptCom

As mentioned in section 4.2.2 the Python implementation within the MICOM framework was used for assessing OptCom. Hence, taxa abundancies were a mandatory input for simulations. The organism abundances predicted by SteadyCom were used (0.01 for *N. europaea* and 0.99 for *N. vulgaris*) and a minimal growth rate of 0.0075 h^{-1} for both organisms was also fixed.

Growth and uptake/production rates differ slightly from the ones obtained using the MICOM simulation method, although are in the same range of values. Likewise, the nitrogen flow prediction does not match the previous predictions, showing that NH_3^+ is consumed and NO_2^- is produced by *N. europaea*, but no conversion of NO_2^- to NO_3^- is achieved by the community. No interspecies metabolite interactions were obtained.

4.3.4 General assessment overview

All simulation methods assessed showed different capabilities, exhibiting strengths and weaknesses, as shown in Figure 4.4. For instance, on one hand, SteadyCom allowed the prediction of *Organisms Abundancies*, while such abundancies are not predicted on any other method, being instead an obligatory input parameter for all other methods.

Feature	Simulation Methods				
	FBA	SMETANA	SteadyCom	MICOM	OptCom (MICOM)
Software availability					
Scalable for large communities					
Customizable					
Fast					
Individual Growth Rate					
Community Growth Rate					
Flux Rates					
Species Interactions					
Organisms Abundances					

Unsatisfactory	Poor	Satisfactory	Good	Outstanding

Figure 4.4. Qualitative assessment of the studied steady-state simulation methods, with application to microbial communities. We evaluated each method from an unsatisfactory (red) to an outstanding performance (dark green).

The evaluation of poor and satisfactory in the FBA and MICOM methods, respectively, are associated with the option of adjusting initial taxa abundances (altering the SBML file or as required input, respectively) to perform simulations.

FBA, SteadyCom, and MICOM are all methods that allowed to inspect *Individual* and *Community Growth* and *Flux Rates.*

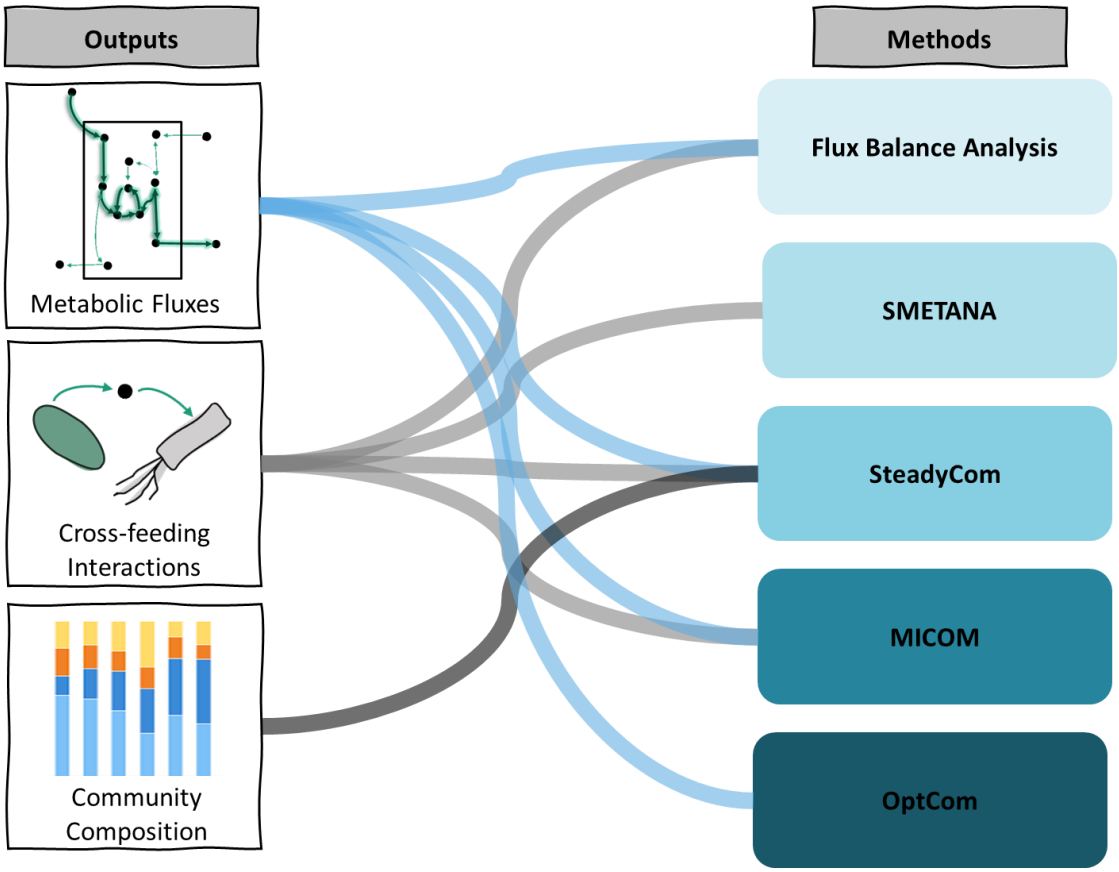


Figure 4.5. Summary of the main output capabilities of each simulation method with application to microbial communities.

SMETANA performed unsatisfactorily in *Individual Growth Rate*, *Community Growth Rate* and *Flux Rates* features and poor on the *Customizable* feature because of its interface's inability to adjust organism abundances and to provide access to growth or flux rates. On the other hand, SMETANA performed as outstanding when checking the ability to predict metabolite *Species Interactions* between organisms. Although FBA and SteadyCom allow inspecting intra-organisms metabolic flux rates easily, and correctly

predict the nitrogen flow throughout the network, no tool other than SMETANA provided insights on potential metabolite exchanges. MICOM does not allow to easily obtain exchange flux distributions. The retrieval of internal *Flux Rates* is not straightforward, needing a conversion step taking into account species abundancies, resulting in an evaluation of this feature of good.

OptCom's original Matlab® version was not available for download (thus evaluated as poor). However, all Python methods evaluated are freely available for download, with straightforward information for installation and scalable for large communities.

As a summary of the main capabilities of simulation methods with application to microbial communities (Figure 4.5) and taking into account the strengths and weaknesses of the respective method, SMETANA should be used in case of screening a microbial community for pairwise interspecies interactions and species interaction potential, and SteadyCom should be used in case of the determination of species abundancies in a community, although in this case, no experimental data is available to validate the abundance predictions obtained. Exception made for SMETANA, all other methods can predict, to some extent, flux and growth rates. According to the results using the case study of the interaction of *N. europaea* and *N. vulgaris*, FBA was the method with prediction values closer to the experimental ones reported in the literature.

4.4 Conclusions

In this work, the performances of different steady-state simulation methods with application to microbial communities were analyzed. A quantitative comparison between simulation methods was not performed due to differences in method-specific definitions of objective functions, which by itself contributes to differences in flux distributions (Santos et al., 2016; Schuetz et al., 2007).

Comparing the simulation methods' performance to predict the nitrification process catalyzed by the bacteria *N. europaea* and *N. vulgaris*, pFBA, and SteadyCom seem to provide the best results. SteadyCom only overcame pFBA by providing more information on the optimum organism abundances, regarding the specific environmental conditions defined. Thus, although its formulation has been developed for predicting single-organism phenotypic behavior, pFBA seems relevant for understanding microbial community interactions.

SMETANA proved to be outstanding at identifying known interactions that are characteristic of the nitrification process and potential interactions between both organisms. However, no information on growth and uptake/withdrawal rates is obtained.

MICOM, and the MICOM implementation version of OptCom, revealed limitations regarding the correct prediction of interactions between the organisms in the nitrification process, as well as the range of flux distribution when an equal organism abundance was assumed. However, when more accurate organism abundances are provided, the obtained fluxes' distribution is comparable to the pFBA and SteadyCom methods.

To summarize, the available simulation methods with application to microbial communities could, to some extent, predict the phenotypic behavior that characterizes the nitrification process catalyzed by *N. europaea* and *N. vulgaris*. Each one of the simulation methods showed strengths and weaknesses and the success of either approach depends on the microbial community composition and complexity. Consequently, the use of more than one simulation method is recommended whenever possible as these showed to complement and validate each other.

Systems biology approaches allow to understand how different species interact and affect their environment. However, these methods are limited to Matlab or Python, which restricts their manipulation to bioinformaticians or other expert researchers. Hence, the implementation of these simulation methods in frameworks with user-friendly interfaces such as Kbase (Arkin et al., 2018) or Optflux (I. Rocha et al.,

2010) should shorten the gap between research in community-based modeling algorithms and wet lab experiments.

4.5 Supplementary Material

Additional file in SBML format: neuropaea.xml

Additional file in SBML format: nvulgaris.xml

Link: DesignOptimizationMicrobialCommunities/models/models_nitro

Additional file in Excel format: Chapter4_Supplementary_Material.xlsx

Link: DesignOptimizationMicrobialCommunities/Data

Table S1 Manual Curation of Oxidative Phosphorylation Pathway - *N. europaea* GSMM

Table S2 SMETANA detailed results

Table S3 SMETANA global results

Table S4 MICOM_fluxes

Chapter 5

Designing Microbial Communities with MEWpy

"We are only as strong as we are united, as weak as we are divided"

J. K. Rowling

In the last years, microbial communities have gained interest for their huge potential to replace engineered single strains throughout the biotechnology, pharmaceutical, and chemical trades. Large amounts of metagenomic high-throughput data are becoming available every day, while Genome-Scale Metabolic Models of microbial communities are also becoming common. However, few studies are available in terms of optimization tools capable of predicting optimal potential genetic modifications at the community level.

Here MEWpy is presented as an integrative Python workbench for metabolic engineering, with methods to explore constraint-based models of microbial communities, allowing the optimization of microbial communities using Evolutionary Algorithms. In specific for microbial community optimization, MEWpy presents multi-objective methods and evaluation functions for the optimization of species cross-feeding interactions, determination of minimal medium composition, as well as untargeted and targeted (directed to specific species in the community) genes/reactions for optimal modifications.

MEWpy can be installed from PyPi (pip install mewpy), the source code being available at https://github.com/BioSystemsUM/mewpy under the GPL license.

Sophia Santos[1], Vítor Pereira[1], Miguel Rocha[1], Oscar Dias[1], Isabel Rocha[2]

[1] Centre of Biological Engineering of University of Minho, Braga, Portugal

[2] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa (ITQB-NOVA)

Authors' contributions

Sophia Santos conceived the study, carried out the simulation and optimization using genome-scale metabolic models of the microbial communities, analyzed the results, and drafted this chapter. Vitor Pereira conceived the study and implemented the optimization methods. Oscar Dias, Miguel Rocha, and Isabel Rocha conceived the study, participated in its design, and coordination, and helped to draft this chapter.

5.1 Introduction

Microbial communities' biological functions are hugely affected by small perturbations linked to the organisms present within them and resource availability (Johns et al., 2016). The manipulation of microbiomes has been successfully implemented (Löffler et al., 2006; McCarty et al., 2011; O'Connell et al., 1996), suggesting that this can be an alternative to improve rational design of single organisms to produce target compounds (Sgobba et al., 2020; Wang et al., 2020).

The possibility of controlling and engineering natural and synthetic microbial communities is one of the main reasons to study microbial communities (García-Jiménez et al., 2021) and, given the success of computational strain optimization methods to rationally design single organisms using GSMMs (Jullesson et al., 2015), it is reasonable to assume that the same strategy can also lead to successful cases using microbial communities. A target product of a microbial community will be a cooperative effort from a structured group of organisms already used to carry out complex interactions and synthesize complex molecules (Bosi et al., 2017; García-Jiménez et al., 2021). However, few optimization tools capable of predicting potential genetic modifications at the community level to unveil the full potential of microbial communities (Eng et al., 2019) are currently available. Most of these tools use strategies to optimize medium composition (Klitgord et al., 2010; Pacheco et al., 2021), intra-species interactions (Thommes et al., 2019), or community configuration (García-Jiménez et al., 2018) for a given objective.

The challenge of performing optimization of microbial communities is the urgent need for the development of more sophisticated and integrative platforms that support different levels of community optimization, focusing not only on optimizing exchange reactions (species cross-feeding interactions, medium composition) but also on the identification of untargeted and targeted (directed to specific species in the community) genes/reactions for deletion/up-/down-regulation, as well as identifying the best community composition for a given objective, using top-down and bottom-up approaches.

In the case of microbial communities, where the complexity of the metabolic network increases per organism included in the community, the use of computational strain optimization metaheuristics, such as Evolutionary Algorithms (EAs), can be a breakthrough (M. Rocha et al., 2008) due to their scalability and flexibility in the definition of objective functions. EAs are stochastic algorithms inspired by nature. Mimicking the Darwinian evolutionary principles, they aim to find sets of modifications whose phenotype best addresses the optimization problem. At each generation, mating and mutation operators

produce a new solution set, from which the fittest are selected to integrate the next population. EAs can be applied using a single optimization objective or using multiple optimization objectives. The use of a single optimization objective, despite being frequently applied in metabolic engineering, is more likely to get caught in a local optimum, evidencing premature convergence, as they manifest more difficulty in preserving high diversity within the populations (Pandey et al., 2014). On the other hand, multi-objective EAs can deliver in a single optimization run a set of solutions with different trade-offs between more than one objective (for example, product rate, growth rate, biomass product coupled yield, number of modifications). Using this approach, a broader set of possible perturbations for analysis can be achieved.

Among the platforms that offer the possibility of the use of EAs for strain optimization, MEWpy (Pereira et al., 2021) offers a practical interface to several strain optimization heuristics, including a set of multi-objective methods driving the optimization towards the best set of enzymes, genes, or reactions, to under/over-express or delete to maximize the production of a target compound on GSMMs defining gene–protein-reaction associations. Moreover, MEWpy allows for phenotype simulation using constraint-based methods provided by COBRApy (Ebrahim et al., 2013) and REFRAMED libraries, such as SteadyCom (Chan et al., 2017), specific for simulating microbial communities.

Here we propose methods for the optimization of microbial communities, using EAs under the MEWPy framework. These methods allow the optimization of species cross-feeding interactions, minimal medium composition as well as untargeted and targeted (directed to specific species in the community) genes/reactions.

5.2 Implementation

MEWpy is fully implemented in the Python programming language, which is being increasingly used by the scientific community. The conceptual architecture of MEWpy comprises three layers: problem definition, phenotype simulation, and optimization. The basis of the MEWpy architecture was fully maintained including only some specific features related to the handling of microbial communities.

In the problem definition layer, during the definition of modification targets, besides the selection of reactions, genes, proteins, or regulatory variables present in the model, it is also possible to target the compartment where those are assigned to, being now possible to select to perform modifications on selected organisms in a microbial community. This layer also includes the selection of the modeling

framework, the modification strategy (deletion, under/overexpression) as well as the target product and the environmental conditions.

In the phenotype simulation layer all the constraint-based methods already used to evaluate wild-type and mutant strains such as Flux Balance Analysis (FBA) (Varma & Palsson, 1994) or Minimization of Metabolic Adjustment (MOMA) (Segrè et al., 2002), as well as Flux Variability Analysis (FVA), can be selected. The different phenotype simulation methods are provided by COBRApy (Ebrahim et al., 2013) and REFRAMED libraries.

In the optimization layer, all the optimization heuristics used for strain design and respective objective functions can be found. The candidate solution fitness of each optimization iteration is evaluated by running the respective phenotype simulation required by the objective function. Here an objective function was included to minimize the molecular weight of the compounds of exchange reactions when using the minimal medium optimization. The EAs in MEWpy are implemented by the Inspyred (Tonda, 2020) and JMetalPy (Benítez-Hidalgo et al., 2019) Python libraries. MEWpy requires a compatible linear programming solver (CPLEX, GUROBI, or GLPK), with installed Python dependencies.

Currently, in specific for microbial communities, MEWpy allows the optimization of 4 different scenarios from a community perspective (Figure 5.1):

(i) Minimal medium optimization, including the objective function of minimizing the molecular weights (*MolecularWeight* fitness evaluation function) of the exchange reactions compounds (possible when chemical formulas of compounds are included in GSMMs);

(ii) Interactions of metabolites among species;

(iii) Identification of reactions/enzymes/genes in the community to manipulate targeting the production of desired compounds, and

(iv) Identification of organism-specific reactions/enzymes/genes to manipulate targeting the production of desired compounds.

MEWpy can be installed from PyPi (*pip install mewpy*) and the source code is available at https://github.com/BioSystemsUM/mewpy under the GPL license.

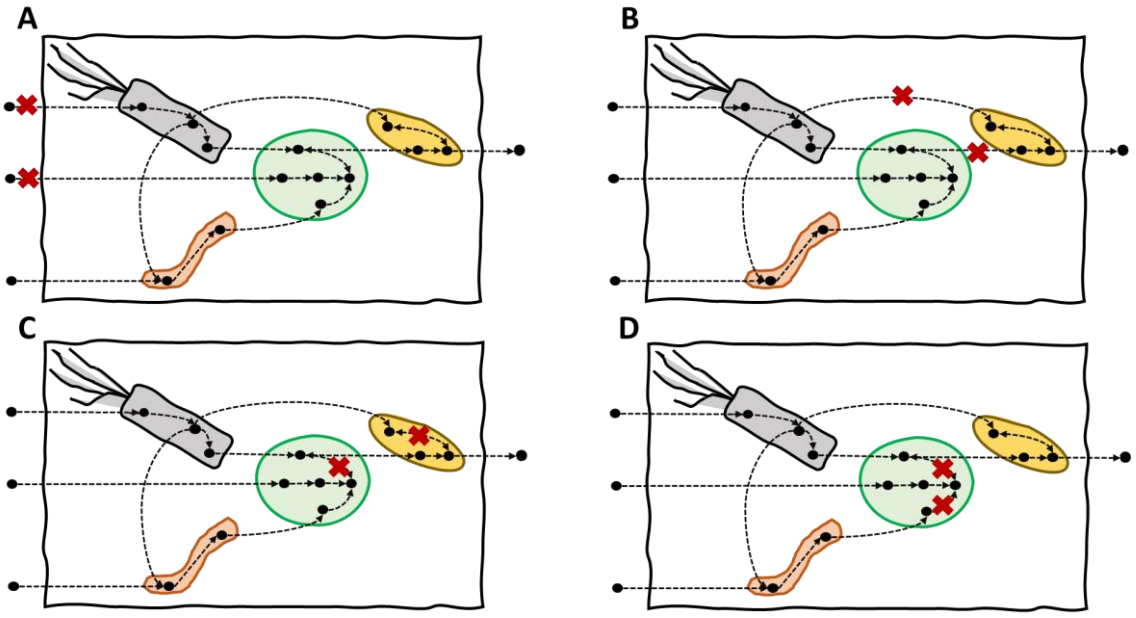


Figure 5.1 Summary of the different strategies to optimize and design microbial communities using MEWpy. A – Medium optimization. B – Intraspecies metabolite exchange optimization. C – Community reaction/gene optimization. D – Organism-specific reaction/gene optimization.

5.3 Usage examples

To illustrate the applicability of the evolutionary algorithms to GSMMs of microbial communities, only two specific cases were tested, namely regarding Minimal Medium Optimization and Organism Specific Reaction Optimization due to the difficulty of finding suitable case studies for the other optimization scenarios. For each scenario, different case studies were evaluated. All simulations were performed using CPLEX 12.8.0 as a solver, through the PyCharm Integrated Development Environment (IDE).

All scripts, input/output auxiliary files, and GSMMs used in this work are available on GitHub at SophiaSantos/DesignOptimizationMicrobialCommunities.

5.3.1 Optimization of target compound production through reaction manipulation

5.3.1.1 Case study

The selected case study shown here considers a synergistic co-culture of model organisms *Escherichia coli* (*E. coli*) and *Saccharomyces cerevisiae* (*S. cerevisiae*) developed for the heterologous

production of the high-value flavonoid naringenin from xylose, by performing further genetic manipulation in only one of the co-culture organisms, in this case, *E. coli* (Zhang et al., 2017). The genetic manipulation performed used gene deletions (*Δ pyk* and *Δ pheA*) and gene over-expressions (*tktA, ppsA aroG*^{fbr}*, aroE,* and *tyrA*^{fbr}) as represented in Figure 5.2. Naringenin is widely produced by citrus plants, especially in grapefruit and bitter orange. The heterologous route to produce naringenin was introduced in *S. cerevisiae*, which is more suitable than prokaryotic organisms to express compounds produced by plants.

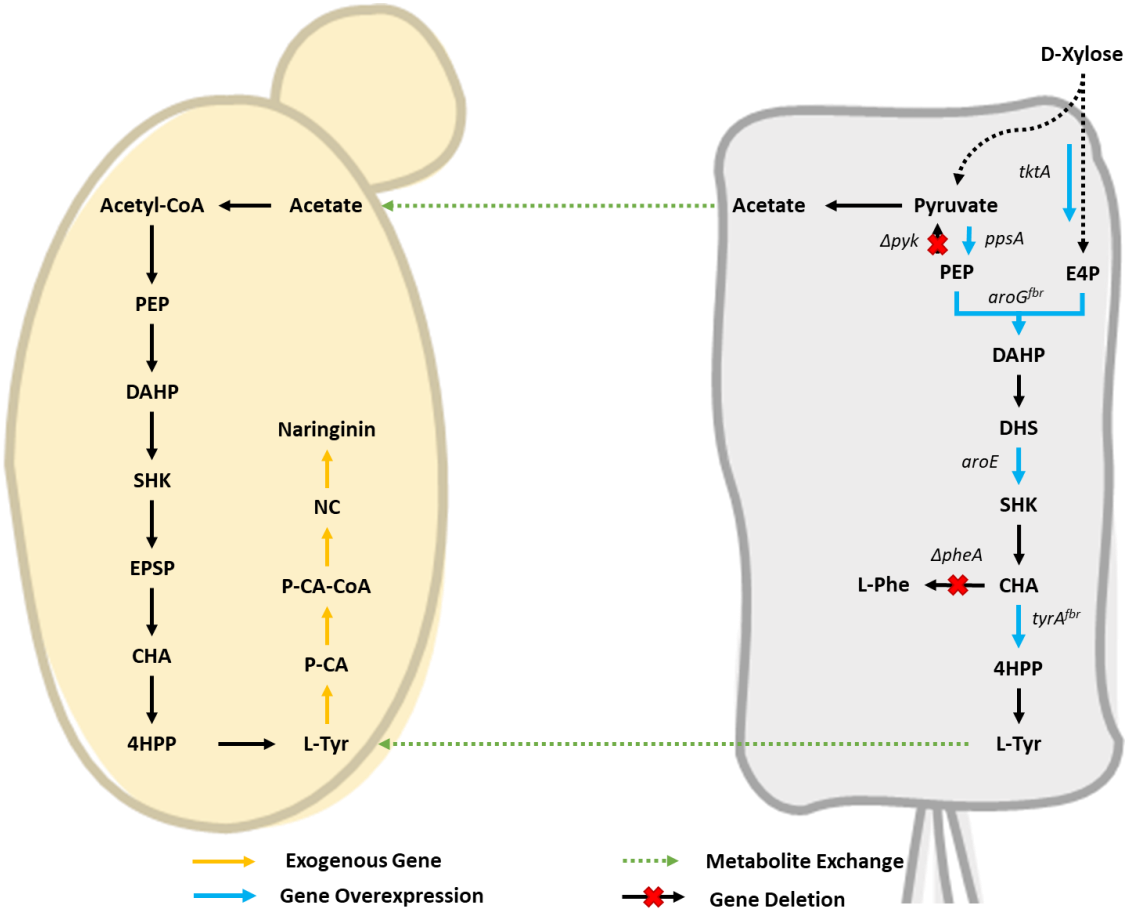


Figure 5.2. Schematic representation of the metabolic pathway for biosynthesis of naringenin via the co-culture of *S. cerevisiae* and *E. coli* with the experimental metabolic engineering strategy used. Gene deletions of *Δpyk* and *ΔpheA*; Gene over-expression of *tktA, ppsA aroG*^{fbr}*, aroE,* and *tyrA*^{fbr}. Adapted from (Zhang et al., 2017). The heterologous pathway for the synthesis of naringenin from L-tyrosine is composed of four enzymes (in yellow). Abbreviations: PEP - phosphoenolpyruvate, E4P - erythrose-4-phosphate, DAHP - 3-deoxy-D-arabino-heptulosonate-7-phosphate, DHS - 3-dehydroshikimic acid, SHK - shikimic acid, CHA - chorismic acid, 4HPP - 4-hydroxyphenylpyruvic acid, L-Phe - L-phenylalanine, L-Tyr - L- tyrosine, EPSP - 5-enolpyruvylshikimate-3-phosphate, p-CA - p-coumaric acid, p-CA-CoA - p-coumaroyl-CoA, NC - naringenin chalcone.

5.3.1.2 GSMMs and parameter definition

For the validation of Reaction Optimization by MEWpy, the GSMMs for *E. coli* (iAF1260) developed by (Feist et al., 2007) and for *S. cerevisiae* (iMM904) developed by (Mo et al., 2009) were used. Both models were retrieved from the BiGG database (Schellenberger et al., 2010), having the same metabolite identifiers, which is required for the construction of the community model. Biomass equations identifiers of both GSMMs were switched to match the same identifier ('R_BIOMASS') for community model construction purposes. The defined medium was inferred from information in the literature (Zhang et al., 2017). Xylose was set as *E. coli's* carbon source and acetate produced by *E. coli* was set as *S. cerevisiae's* carbon source using reaction community identifiers. As shown in Figure 5.2 the metabolic engineering strategy in this case study includes overexpression and deletion of *E. coli* genes. Both strategies were tested for reactions (ROUProblem and RKOProblem), using MEWpy. Target reactions for the deletion or under-over expression optimization were defined as the reactions assigned with iAF1260 (*E. coli* GSMM identifier), removing from those all the essential and exchange reactions. Although the case study objective is the production of naringenin by *S. cerevisiae*, all the metabolic engineering performed in *E. coli* had the objective to produce L – tyrosine, and for that reason, the objective of all optimizations will be the production of L – tyrosine by *E. coli*. All optimizations were run using 100 as the maximum number of generations and the candidate maximum size of two was used for the RKOProblem (maximum number of reaction knock-outs) and five for the ROUProblem (maximum number of over/under-expression of reactions), as shown in the case study culture (Zhang et al., 2017). A Jupyter Notebook with the whole process is available in Supplementary Material – naringenin_optimization.ipynb.

Table 5.1 FVA analysis of the co-culture composed by *E. coli* and *S. cerevisiae* L-tyrosine production capabilities. The specific growth rate was set to at least 10% of the specific growth rate obtained with the pFBA simulation.

	L – tyrosine Production (mmol g_{DW}^{-1} h^{-1})
E. coli	[-0.0572, 0.463]
S. cerevisiae	[-0.463, 0.0572]

An FVA was performed before the optimization process to try to verify the community model's ability to produce L-tyrosine (Table 5.1). The FVA analysis showed that under the environmental conditions defined, L–tyrosine is being exchanged between the two organisms. Moreover, higher production of L–tyrosine is obtained by *E. coli* (0.463 mmol g_{DW}^{-1} h^{-1}).

5.3.1.3 Reaction Deletion Optimization (RKOProblem)

When performing the reaction deletion optimization (RKOProblem), although a maximum of two modifications was set, the best solutions included only 1 modification, *Δpyk* which is one of the two deletions tested in the case study (Figure 5.3).

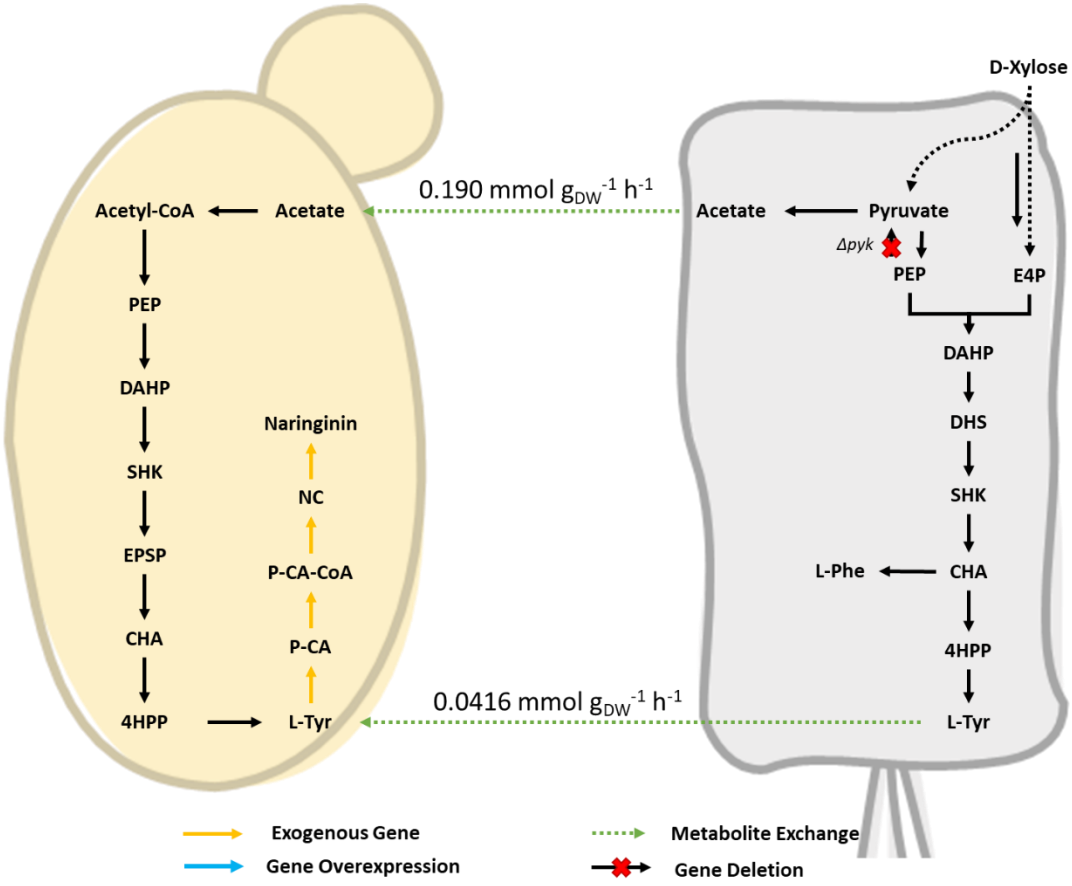


Figure 5.3. Schematic representation of the metabolic pathway for biosynthesis of naringenin via the co-culture of *S. cerevisiae* and *E. coli* with the results of MEWpy reaction knock-out optimization and the pFBA prediction values of the effect of deletion of the *Δpyk* corresponding reaction. Adapted from (Zhang et al., 2017). The heterologous pathway for the synthesis of naringenin from L-tyrosine is composed of four enzymes (in yellow). Abbreviations: PEP - phosphoenolpyruvate, E4P - erythrose-4-phosphate, DAHP - 3-deoxy-D-arabino-heptulosonate-7-phosphate, DHS - 3-dehydroshikimic acid, SHK - shikimic acid, CHA - chorismic acid, 4HPP - 4-hydroxyphenylpyruvic acid, L-Phe - L-phenylalanine, L-Tyr - L- tyrosine, EPSP - 5-enolpyruvylshikimate-3-phosphate, p-CA - p-coumaric acid, p-CA-CoA - p-coumaroyl-CoA, NC - naringenin chalcone.

Even in solutions with two modifications, *Δpyk* is included in all of them. *ΔpheA* appears in solutions when the number of modifications is higher than 2 jointly with the deletion of reactions in the

L–serine biosynthesis and in mannose metabolism (Supplementary Material – naringenin_optimization.ipynb) but never together with *Δ pyk*. When performing a pFBA simulation using the knock-out of *pyk* and *pheA,* the predicted value for the production of L-tyrosine by *E. coli* and consequent consumption by *S. cerevisiae* is very similar to the sole deletion of *pyk* (Table 5.2). In summary, the results of the reaction deletion optimization suggest that indeed *Δ pyk* seems to have a relevant role *in silico* for the *E. coli* production of L-tyrosine under the co-culture with *S. cerevisiae*, showing also the MEWpy applicability on reaction deletion optimization in the context of microbial communities.

5.3.1.4 Reaction Over/Under expression Optimization (ROUProblem)

When performing the reaction over/under expression optimization (ROUProblem), although a maximum of five modifications was set, the best solutions included only one modification, a 2-fold over-expression of *tktA* which is one of the five over-expression genes used in the case study (Figure 5.2). Even in solutions with more modifications, *tktA* is included in all of them jointly with over-expressions in reactions on sugar metabolism (Supplementary Material – naringenin_optimization.ipynb).

Table 5.2. pFBA analysis of the co-culture composed by *E. coli* and *S. cerevisiae* for L-tyrosine production using the MEWpy deletion and/or overexpression reactions predictions. Consumed metabolite rates are represented as negative and produced metabolite rates are represented as positive.

***E. coli* L–tyrosine Production in co-culture** (mmol g_{DW}^{-1} h^{-1})	
Wild-type	-0.0571
Δ pyk	0.0416
Δ pyk and *Δ pheA*	0.0415
tktA 2-fold overexpression	0.0419
Δ pyk and *tktA* 2-fold overexpression	0.0415

Indeed, when performing a pFBA simulation constraining the *tktA* reaction to a 2-fold flux value, L-tyrosine is predicted to be produced by *E. coli* (Table 5.2) in the same range as with *pyk* knock-out. However, when adding to the pFBA simulation constraints the knock-out of *Δ pyk*, there is no effect on the production of L-tyrosine. Nevertheless, MEWpy over/under expression optimization showed significant

information on which reactions to try to over/under express for this co-culture showing once more, its applicability in optimizing microbial communities.

5.3.2 Minimal Medium Optimization

For the validation of the Minimal Medium Optimization by MEWpy, eight published and manually curated GSMMs of single prokaryotic and eukaryotic organisms were used that are part of one or more of the five selected natural and synthetic microbial communities (Table 5.3).

Table 5.3. Microbial community case studies used for the Minimal Medium Optimization and information on the published and manually curated GSMMs used for each community case study.

Genome-scale Metabolic Models			**Microbial Community**	
ID	**Organism**	**Reference**	**ID**	**Reference**
iAO358	*Lactococcus lactis*	(Oliveira et al., 2005)	1	(Ponomarova et al., 2017)
iBT721	*Lactobacillus plantarum*	(Teusink et al., 2006)		
iMM904	*Saccharomyces cerevisiae*	(Mo et al., 2009)		
iBif452	*Bifidobacterium adolescentis*	(El-Semman et al., 2014)	2	(El-Semman et al., 2014)
iFap484	*Faecalibacterium prausnitzii*	(El-Semman et al., 2014)		
iJN1411	*Pseudomonas putida*	(Nogales et al., 2017)	3	(Sgobba et al., 2018)
iAF1260	*Escherichia coli*	(Feist et al., 2007)		
iYS432	*Corynebacterium glutamicum*	(Shinfuku et al., 2009)	4	(Sgobba et al., 2018)
iAF1260	*Escherichia coli*	(Feist et al., 2007)		
iAF1260	*Escherichia coli*	(Feist et al., 2007)	5	(Zhang et al., 2017)
iMM904	*Saccharomyces cerevisiae*	(Mo et al., 2009)		

GSMMs were screened for the same external metabolite identifiers, and the BiGG (Schellenberger et al., 2010) annotation was adopted when different annotations were found. Biomass equations

identifiers of all GSMMs were switched to match the same identifier ('R_BIOMASS') for community model construction purposes, and compound chemical formulas were included in all GSMMs that lacked that information. The experimentally defined medium of each microbial community was determined by information in the literature (Table 5.3) - Supplementary Material – Table S1.

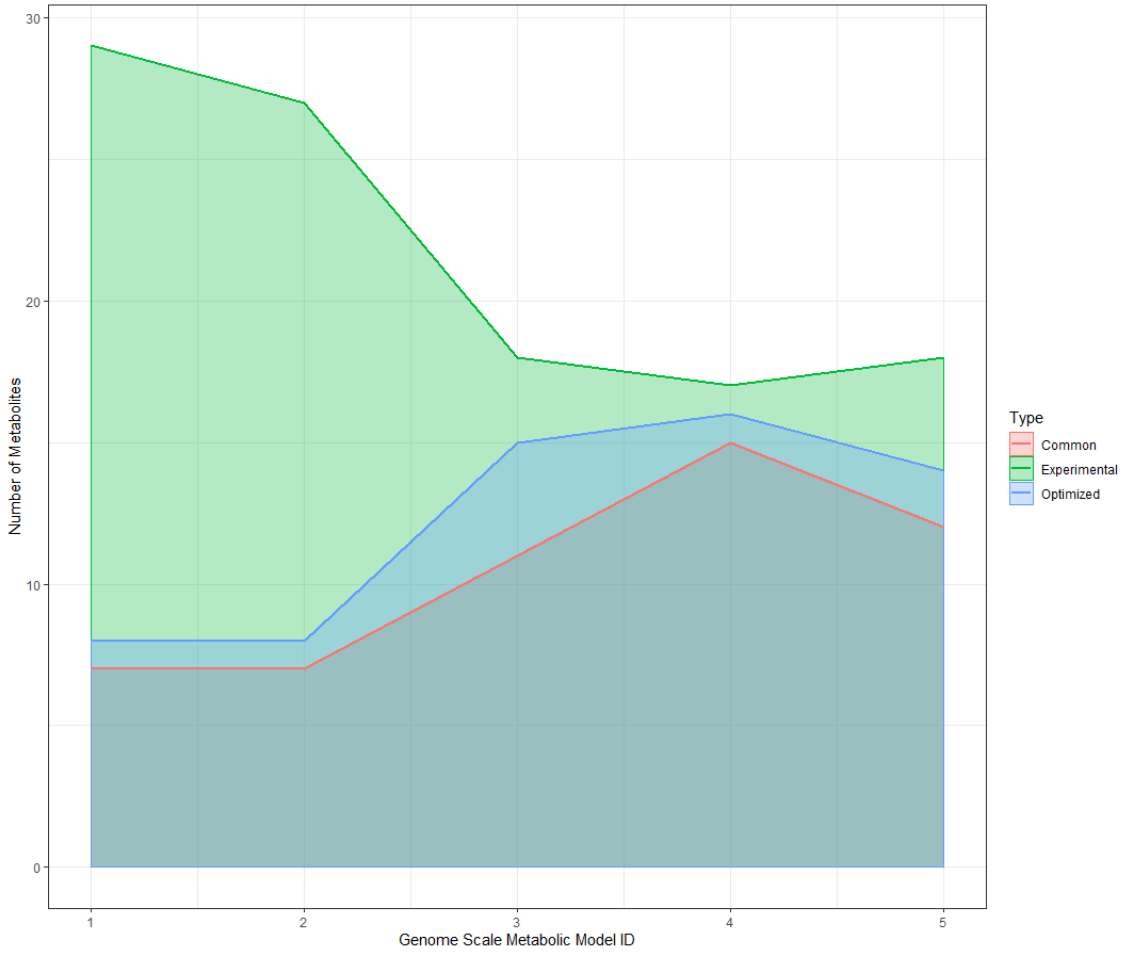


Figure 5.4. Number of exchange reactions defined on the Experimental medium (green line), on the optimized medium by MEWpy (blue line) without using Molecular Weight minimization, and the number of exchange reactions that are common in both mediums (pink line).

For the optimization of the minimal medium for each microbial community, a carbon source, defined in the literature, was set as a constraint, as well as all exchange reactions that are essential for each community GSMM have growth. Minimal Medium Optimization was tested with and without minimizing exchange compound Molecular Weight. The best results were compared and validated with

experimental data from literature (Figure 5.4) - Supplementary Material – Table S2, Table S3, Table S4, Table S5, Table S6.

When analyzing the obtained results, major differences are exhibited in communities ID 1 (co-culture *Lactococcus lactis*, *Lactobacillus plantarum,* and *S. cerevisiae*) and ID 2 (co-culture *Bifidobacterium adolescentis* and *Faecalibacterium prausnitzii*). Indeed, in these two cases, a rich medium is used in the experimental setup, which resulted in an *Experimental defined medium* with a high number of metabolites. In community ID 1 most of the *Experimental medium* components are ions, vitamins, and cofactors, as well as amino acids which are essential to *Lactococcus lactis.* Whereas in community ID 2 *Experimental medium* components are mainly amino acids, which are an important part of the gut ecosystem. In both cases, the majority of the amino acids in the *Experimental medium* were not included in the *Optimized medium*, indicating that individual amino acid auxotrophies are being bridged by other organisms within the community.

Importantly, a large part of the *Optimized medium* components is similar to the *Experimental medium*. Differences between experimental and predicted mediums are higher in the community ID 3 composed of *Pseudomonas putida* and *E. coli* (Sgobba et al., 2018), where the predictions revealed that vitamins (for instance, riboflavin, thymidine, and pantothenate) are not required for the community growth. These differences cannot be directly translated for the definition of a new experimental minimal medium because most of them are associated with the lack of metabolite requirement information included in the GSMM. In all other cases, the differences happen because MEWpy often selects complex molecules that can, at the same time, be nitrogen, phosphate, and/or sulfur sources (e. g. guanosine 3-phosphate, adenosine 3, 5–biphosphate or L-cysteine), as the objective is to determine a minimal number of exchange reactions that allow the growth of a specific community. Also, the best solutions generally differed among them on the nitrogen, phosphate, sulfur, or iron sources when the GSMMs included more than one suitable replacement (for instance, sulfate, thiosulfate, hydrogen sulfide, or L-cysteine). These results suggest that in cases where the exchange metabolite chemical formula is available in the GSMM, the minimization of the sum of the Molecular Weight of the metabolites in the solution would be helpful to eliminate mainly complex molecules that can function as nitrogen, phosphate, and/or sulfur sources. This hypothesis was corroborated in part using the minimization of exchange metabolite Molecular Weight, which eliminated the best solutions that had complex molecules, such as the replacement of

sulfate (molecular weight of 96.06 g mol^{-1}) for hydrogen sulfide (molecular weight of 34.08 g mol^{-1}). However, in none of the case studies the number of metabolites in the best solutions changed.

Overall, the results obtained for Minimal Medium Optimization using MEWpy showed biological significance, correctly predicting a minimal medium for each community. However, results should be carefully analyzed considering differences between the information included in each organism GSMM within the community and the experimental data available.

5.4 Conclusions

Here methods using EAs were developed and implemented in MEWpy for the *in silico* design of microbial communities using GSMMs, allowing, depending on the individual GSMMs detail, (i) minimal medium optimization, (ii) species metabolite interactions, (iii) community reaction/gene/enzyme optimization (iv) organism specific reaction/gene/enzyme optimization for a given objective (maximize/minimize growth or production of target compounds). MEWpy already offers a practical interface to use strain optimization metaheuristics, including multi-objective methods that are suitable for application in GSMMs of microbial communities. Results on organism-specific reaction optimization and minimal medium optimization showed good approximations to experimental designs, showing the applicability of the implemented methods for the design of microbial communities.

MEWpy also allows simulating microbial communities, as it offers all phenotype simulation methods from the COBRApy (Ebrahim et al., 2013) and REFRAMED libraries, this last one with implemented methods specific for the application in microbial communities, such as the construction of community models from individual GSMMs or the SteadyCom (Chan et al., 2017) simulation method.

In conclusion, MEWpy aims at being a reference tool for the metabolic engineering community trying to make available a diverse catalog of simulation and optimization heuristics and functions appliable not only to GSMMs of individual organisms and microbial communities but also GECKO (Sánchez et al., 2017) and Metabolism Expression and Thermodynamics Flux (ETFL) (Salvy & Hatzimanikatis, 2020) models as well as kinetic models. However, there is still room for new implementations on MEWpy concerning microbial communities' optimization. Current efforts are focused on optimizing the best community composition for a given objective, using top-down and bottom-up approaches.

5.5 Supplementary Material

Additional file in IPYNB format: naringenin_optimization.ipynb

Link: DesignOptimizationMicrobialCommunities/JupyterNootbooks/

Additional files in SBML format: iAO358.xml, iBT721.xml, iMM904.xml, iBif452.xml, iFap484.xml, iJN1411.xml, iAF1260.xml, iYS432.xml.

Link: DesignOptimizationMicrobialCommunities/models/models/

Additional files in SBML format: community_bif_fab.xml, community_eco_ppu.xml, community_eco_sce.xml, community_eco_cgl.xml, community_sce_lla_lpl.xml.

Link: DesignOptimizationMicrobialCommunities/models/community_models/

Additional file in Excel format: Chapter5_Supplementary_Material.xlsx

Link: DesignOptimizationMicrobialCommunities/Data

Table S1 Experimental Medium defined for each Microbial Community

Table S2 Minimal Medium Optimization results for Community ID 1

Table S3 Minimal Medium Optimization results for Community ID 2

Table S4 Minimal Medium Optimization results for Community ID 3

Table S5 Minimal Medium Optimization results for Community ID 4

Table S6 Minimal Medium Optimization results for Community ID 5

Chapter 6

Modeling and Design of Microbial Communities from Extremophilic Environments in the Azores

"Alone we can do little, together we can do so much."

Helen Keller

Genome-scale metabolic models of nine extremophilic organisms expected to be present in the hydrothermal sites in São Miguel, Azores, are here presented. The metabolic models can be used independently or taken together to investigate the extraordinary microbial metabolism and co-metabolism in extreme environments. These reconstructions, together with simulation tools used in a community context, allowed us to elucidate the roles of specific organisms. In particular, *Pyrobaculum aerophilum* was predicted to have an important role in producing amino acids in the presence of other extremophile organisms. Furthermore, *S. azorense* seems to have all the required metabolic traits to produce cellulose in the presence of *T. adornatus*.

Untargeted and targeted (directed to specific species in the community) co-culture optimization was performed using the MEWpy framework, to evaluate *S. azorense's* cellulose production capabilities. Results demonstrated acetate's important role in the metabolic cellulose production route by *S. azorense*.

6.1 Introduction

Microorganisms rarely grow in isolation but rather in complex and diverse communities, interacting with other microorganisms competing for available resources, or cooperating through metabolite exchanges (Zelezniak et al., 2015). The capability of microorganisms to adapt to different ecosystems and available resources is remarkable, especially in extreme environments where conditions impose a lack of nutrients and energy (Ando et al., 2021; Sarmiento et al., 2015). Often, these adaptations are a consequence of key changes in the organism enzymes' amino acid sequences, which are translated into variations in the structure, flexibility, charge, and/or hydrophobicity (Sarmiento et al., 2015). Also, biocompounds produced by extremophiles are usually more stable under a wide range of temperatures, pH, and saline conditions. Thus, the possibility of controlling and engineering extremophile communities is of huge interest due to their exceptional metabolic capabilities and the exciting industrial opportunities that their extremozymes and added-value products can offer (Van den Burg, 2003).

Mathematical modeling of these complex biological systems, through the use of genome-scale metabolic models (GSMMs), is nowadays an indispensable tool, allowing faster and more systematic predictions of the phenotypic behavior, under different environmental conditions (Chan et al., 2017; García-Jiménez et al., 2018; Zelezniak et al., 2015; Zorrilla et al., 2021). The list of applications of GSMMs in a microbial community context is increasing every day, with emphasis on the prediction of microbe-microbe/host-microbe interactions (Almut Heinken et al., 2020; Zelezniak et al., 2015) as well as the design and engineering of microbial communities (García-Jiménez et al., 2018; Pacheco & Segrè, 2021), which is being used as an alternative to improve limitations of rational design of pure cultures (Sgobba et al., 2020; Wang et al., 2020).

High-quality GSMMs' offer more reliable predictions of the phenotypic behavior (Lieven et al., 2020). However, such GSMMs must be manually curated and validated with experimental data, a process that, even with tools that automate most of the reconstruction tasks, is still laborious and time-consuming (Thiele et al., 2010). This issue worsens for microbial communities, which often include dozens of organisms. Workflows that semi-automatically construct a GSMM of a prokaryotic organism in a matter of minutes and decrease the time spent in these reconstructions have been developed (Arkin et al., 2018; Heinken et al., 2020; Machado et al., 2018). Moreover, these models have already shown their reliability when predicting growth, response to nutrients, and gene essentiality in single organisms and even microbial communities (Chng et al., 2020; Machado et al., 2021; Nayfach et al., 2020). However, in the

case of extremophiles (which have particular metabolic pathways), and from a rational design perspective, these models should be used with caution, as these tools are based on model organisms' templates (e.g., *Escherichia coli*) and the returned models may be inaccurate (Lieven et al., 2020), and need further manual curation.

Microbial communities in high-temperature environments are generally less diverse (Inskeep et al., 2013) making hydrothermal habitats an ideal model system to study principles of community structure and function (Sahm et al., 2013) using GSMMs. Hydrothermal sites have generally a high number of members of the Aquificales (Strazzulli et al., 2017), which have recently received scientific interest for being believed to be the earliest lineage within the domain Bacteria (Takacs-vesbach et al., 2013). The hydrothermal sites in São Miguel, Azores, is an example of an Aquificales' highly populated environment as detailed in Chapter 2. The most abundant Aquificales member expected to be present in samples from the analyzed Azorean hydrothermal sites is *Sulfurihydrogenibium azorense* Az-Fu1, which curiously was first isolated in January 2001 from terrestrial hot springs at Furnas, in the same Azorean island (Aguiar et al., 2004). The GSMM reconstruction of *S. azorense* Az-Fu1, analyzed in Chapter 3, revealed the metabolic potential for this organism to produce cellulose under nitrogen-limiting conditions. However, the metabolic engineering optimization, supported by evolutionary algorithms, performed for this work, did not return a solution robust enough to increase cellulose production. *S. azorense* Az-Fu1 was projected to be part of two microbial communities' samples in the hydrothermal area explored. Thus, the organism's metabolic role within each microbial community and whether the co-occurrence of different organisms influences *S. azorense*'s ability to produce cellulose must be studied.

In this chapter, we present the GSMMs of nine extremophile microorganisms likely to be present in the hydrothermal sites in São Miguel, Azores. Community-based computational strain simulations and optimizations were performed to investigate whether *S. azorense* Az-Fu1's ability to produce cellulose can be indeed enhanced by a microbial community.

Lastly, these reconstructions can be used independently or taken together to investigate the extraordinary microbial metabolism and co-metabolism in extreme environments.

6.2 Methods

6.2.1 Selection of the extremophile microorganisms and retrieval of whole-genome sequences

Five hydrothermal samples were analyzed to determine prokaryotic community composition, using metagenomic approaches. Taxonomic profiling of all samples was performed using assembly-based and read-based analysis algorithms as recommended (Quince et al., 2017). Based on the taxonomic profiling methods applied, 12 organisms were predicted to be part of one or multiple of the analyzed samples. Results (see Chapter 2) showed a high presence of *Aquificales* and *Crenarchaeota* members, regularly present in hydrothermal sites (Strazzulli et al., 2017). Based on literature and single-organism growth information, nine of these organisms were selected for the genome-scale metabolic reconstruction step. *Sulfurihydrogenibium azorense* Az-Fu1 was isolated in January 2001 from terrestrial hot springs at Furnas, São Miguel Island, Azores, Portugal (Aguiar et al., 2004) and was predicted to be one of the most abundant organisms present in two of the five samples analyzed. Hence, a manually curated GSMM was reconstructed and presented in Chapter 3. *Acidimicrobium ferrooxidans* DSM 10331 and *Acidithiobacillus caldus* SM-1 have been reportedly used in bioleaching experiments (Oyama et al., 2018), and their interaction was evaluated using GSMMs of both organisms, which were reconstructed as part of a Master Thesis (Nunes, 2021) developed at our research group. The remaining GSMMs were reconstructed in this work.

6.2.2 Online Databases

Different online databases were used in each stage of this work, most of them through *merlin*'s framework. The National Center for Biotechnology Information (NCBI) (NCBI Resource Coordinators, 2018), was used to retrieve the genome sequences and all genome files (Table 6.1) were imported by *merlin*, Universal Protein Resource Knowledgebase (UniProtKB) (Consortium, 2021) and BRENDA (A. Chang et al., 2021) were used to obtain enzyme functional information through *merlin*'s re-annotation pipeline, Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al., 2016) was used to obtain reactions based on the enzyme commission numbers (EC numbers) of the annotated genome into *merlin*, and also to perform pathway analysis, MetaCyc (Caspi et al., 2014) and BiGG models (King et al., 2016) were used for network curation, ModelSEED (Seaver et al., 2021) was used for *merlin*'s correct

reversibility of reactions workflow, PSORTb 3.0 (Yu et al., 2010) was used to predict reactions compartments, the Transporters Classification Database (TCDB) (Saier et al., 2021) was used to predict the model transport reactions through *merlin*'s plug-in TranSyT.

Table 6.1. Genome's information used to import genome files into *merlin*.

Organism	Taxonomy ID	Assembly Accession Number
Desulfurococcus amylolyticus DSM 16532	NCBI: txid768672	ASM23101v3
Pyrobaculum aerophilum str. IM2	NCBI: txid178306	ASM722v1
Thermodesulfovibrio yellowstonii DSM 11347	NCBI: txid289376	ASM2098v1
Thermofilum adornatus 1910b	NCBI: txid697581	ASM81324v1
Thermoplasma acidophilum DSM 1728	NCBI: txid273075	ASM19591v1
Thermus scotoductus SA-01	NCBI: txid743525	ASM18700v1

6.2.3 Metabolic Models Reconstruction

merlin (Capela et al., 2021) is a user-friendly framework that allows performing several steps of the reconstruction process semi-automatically, downloading relevant information from several databases, and was used to assist in the reconstruction of all GSMMs. Moreover, it has a graphical interface that facilitates GSMM information reviewing and manual curation. The main steps of each GSMM model reconstruction process are hereafter described.

6.2.3.1 Genome Annotation

merlin allows performing the functional annotation of a genome, using as similarity search engines the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990) and Diamond (Buchfink et al., 2021) against databases that contain reviewed (such as UniProt/Swiss-Prot) and unreviewed (such as UniProt/TrEMBL) enzyme information. The EC numbers and enzymatic functions assigned to each gene

are scored based on the taxonomy and frequency of similar sequences, as described elsewhere (Dias et al., 2018). The genome annotation is a crucial step of the GSMM reconstruction process, as incorrect EC numbers and enzymatic function assignments can significantly impact the model performance. Once the similarity search is complete, *merlin*'s automatic annotation workflow feature (Capela et al., 2021) can prioritize the obtained gene products and EC numbers annotations. This operation considers a list of organisms ordered by phylogenetic similarity provided by the user and defines a confidence level (A to I) for each gene annotation when a match is found. For all organisms within this study, due to the lack of available information on closely related organisms and to maximize retrieval of reviewed information, the automatic workflow feature was configured to use genus instead of species as input. The ranked list of the closely related phylogenetic genus of all organisms and the associated phylogenetic trees (Supplementary Material) were constructed from 16S RNA sequences reference organisms of each genus using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool (Sievers et al., 2011).

6.2.3.2 Assembling the Metabolic Network

The assembly of a metabolic network starts with gathering all reactions present in the organism. *merlin* retrieves reactions by importing them from KEGG, based on the annotated EC numbers from the previous step and spontaneous reactions. Reactions should also be balanced (generic and metabolites without formula must be curated), and reversibility must be confirmed to avoid mispredictions of the model. To assist these steps, *merlin* checks whether a reaction is balanced and also corrects the reactions' reversibility. Although these steps are automated in *merlin*, we performed manual curation based on curated information from literature or databases, such as MetaCyc (Caspi et al., 2014), as this step is strongly encouraged.

Compartmentalization

The compartmentalization of each model was based on results obtained from PSORTb 3.0 (Yu et al., 2010) choosing the specific organism type and gram stain before online submission. The "Long format" report generated was imported, and *merlin*'s compartments feature assigned each reaction to its specific compartment (Capela et al., 2021).

Transport Reactions

After compartmentalizing the model, we defined transport reactions between compartments and also with the exterior. *merlin*'s TranSyT (Lagoa et al., 2021) was used to generate system-specific transport reactions associated with Gene-Protein-Reactions (GPRs) rules that are automatically integrated into each model.

Genes, proteins, and reactions

A high-quality GSMM requires GPRs' rules to predict genetic modifications accurately. These associations are usually defined according to databases and literature (Rocha et al., 2008; Thiele et al., 2010). *merlin*'s *"Gene-Protein-Reaction rules"* feature automatically adds GPRs rules to the model. The algorithm used by *merlin* to implement these rules is described elsewhere (Dias et al., 2015).

6.2.3.3 Converting the Metabolic Network to a Stoichiometric Model

Biomass equation

Biomass composition must be experimentally determined in cells growing in the log phase before being included in the model. However, in the absence of experimental information organism-specific, data from genome information (particularly nucleotides, deoxynucleotides, and amino acids) can be used or adapted from phylogenetically related organisms. The importance of an accurate biomass composition determination has been reported (Santos et al., 2016), showing that even minor differences in biomass component coefficients can significantly impact simulations' predictions. Therefore, a well-defined biomass equation is crucial for the GSMM reconstruction process.

Information regarding the biomass composition of the organisms present in this study is scarce in the literature. Hence, for organisms of the domain Bacteria, the macromolecular composition was adapted from the gram-negative bacterium *Escherichia coli* (*E. coli*) (Feist et al., 2007) or the gram-positive bacterium *Bacillus subtilis* (*B. subtilis*) (Dauner & Sauer, 2001). For organisms for the domain Archaea, the macromolecular composition was adapted from the archaeon *Methanosarcina fusaro* (*M. fusaro*) (Goyal et al., 2014). The composition of amino acids, nucleotides, and deoxynucleotides was estimated from genome information through *merlin*'s "e-Biomass Equation" feature. This feature also

automatically includes cofactor composition based on a study of universal essential cofactors in prokaryotes (Xavier et al., 2017). The fatty acids composition was taken from experimental data from the literature for all organisms. The lipid, carbohydrates, and cell wall composition (when required) compositions were adapted from the information of reference organisms considering enzyme annotation to include or exclude specific elements. When required, new coefficients were calculated, maintaining the relative abundances of the original data.

Alternative biomass equations were defined for simulations under aerobic and anaerobic conditions. Since in the latter environmental conditions, Heme is not required, the compound was removed from the Cofactor composition, and all other coefficients were recalculated as mentioned before.

Growth and maintenance ATP requirements

No information was found for any of these organisms for growth and ATP maintenance. Therefore, such data was adapted from experimental data for *E. coli* (Feist et al., 2007; Neidhardt et al., 1990).

6.2.3.4 Metabolic Model Curation and Validation

The model curation is an iterative process that stops when simulation results match experimental data in the literature. *merlin*'s interface allows the user to efficiently perform re-annotations, correct reactions stoichiometric balance and directionality, include or exclude reactions from the model, and finally, export models in Systems Biology Markup Language (SMBL) format to be used in simulation using platforms such as Optflux (I. Rocha et al., 2010), Matlab®'s COBRA toolbox (Heirendt et al., 2019), COBRApy (Ebrahim et al., 2013) or REFRAMED.

Gap-filling

Before being ready for simulations, the metabolic network must be screened for possible gaps. The presence of gaps can deeply compromise the synthesis of biomass components and other relevant compounds. To assist in this process, *merlin*'s "BioISO" (Cruz et al., 2021) helps trace back the network to identify gaps that can be originated from errors in genome annotation, absence of enzymatic, transport, or exchange reactions, or incorrect reaction irreversibility or direction. Other features included in *merlin*,

such "find Blocked reactions", which identifies reactions that contain dead-end metabolites, and "Draw in Browser" which opens on a web browser a selected KEGG pathway, showing specifically highlighted enzymes and reactions present in the model, facilitate the detection of gaps in the network (Capela et al., 2021). Literature and databases (KEGG, MetaCyc, BRENDA, for instance) should be used to assist the gap-filling process. This is an iterative process, repeated until all biomass precursors, and other essential compounds, can be synthesized, and a feasible model is obtained.

Aerobic and anaerobic metabolism and Carbon Source utilization

Experimental data from the literature was consulted to identify the growth conditions for each organism. A defined minimal medium was favored when available. Whenever an organism was reported to have aerobic and anaerobic growth, both conditions were tested. Growth under different carbon sources was also tested and the presence of the specific transport reactions was screened.

6.2.3.5 Simulations and Strain Optimizations

Software

SMETANA (Zelezniak et al., 2015) was used to predict pairwise interactions. All other simulations and community strain optimizations were performed using the Python package MEWpy (Pereira et al., 2021) version 0.1.28, CPLEX 12.8.0 as a solver, through the PyCharm Integrated Development Environment (IDE). MEWpy allows for phenotype simulation using constraint-based methods provided by COBRApy (Ebrahim et al., 2013) and REFRAMED libraries, such as Flux Balance Analysis (FBA) (Varma & Palsson, 1994), or SteadyCom (Chan et al., 2017), specific for simulating microbial communities.

Quantitative evaluation of the model was performed using Flux Variability Analysis (FVA) (Mahadevan et al., 2003) to determine *S. azorense's* cellulose production capabilities in co-culture. The analysis included setting the specific growth rate to at least 10% of the specific growth rate obtained with parsimonious FBA (pFBA) (Lewis et al., 2010) in the respective reference flux distribution.

All Python scripts, input/output auxiliary files, and GSMMs used in this work are available on GitHub at SophiaSantos/DesignOptimizationMicrobialCommunities.

6.3 Results and Discussion

6.3.1 Samples Characterization

Taxonomic profiling of all samples from hydrothermal vents in São Miguel, Azores was performed using assembly-based and read-based analysis algorithms (Chapter 2). Joint information from both approaches revealed the potential presence of 12 different microorganisms. Several of these organisms were predicted to be present in various samples. Physiological and metabolic characteristics of all identified organisms were screened to determine if their presence in such environments is known or viable, based on each sample site's physiochemical properties. Two of the five samples, FCRG and PCRG, were excluded from this characterization, as both are composed of only two organisms present in the CV sample and therefore their interactions will be screened in that sample.

From all organisms identified through the profiling algorithms nine were selected for the GSMM reconstruction process:

***Acidithiobacillus caldus* SM-1**

Acidithiobacillus caldus SM-1 (*A. caldus*) is a gram-negative, moderately thermophilic, acidophilic (pH range 1.0 - 3.5), chemolithoautotrophic bacterium. Growth temperature ranges between 32 and 52°C. Mixotrophic growth is obtained with tetrathionate and glucose or yeast extract (Hallberg et al., 1994). *A. caldus* is capable of oxidizing elemental sulfur and other reduced inorganic sulfur compounds (Hallberg et al., 1996) and therefore is one of the dominant sulfur-oxidizing bacteria in bioleaching reactors together with iron-oxidizing bacteria (L. Chen et al., 2012; Watkin et al., 2009; Watling et al., 2014). The GSMM of this organism was reconstructed as part of the master thesis entitled "Genome-scale metabolic modeling of an extremophile microbial community" by Rui Barros Nunes.

***Acidimicrobium ferrooxidans* DSM 10331**

Acidimicrobium ferrooxidans DSM 10331 (*A. ferrooxidans*) is a gram-positive, moderately thermophilic, acidophilic bacterium. Growth has been observed between temperatures of 45 and 50°C of temperature and 1 and 3.5 pH values. Autotrophic growth is possible on ferrous iron and heterotrophic growth is possible on yeast extract (Cleaver et al., 2007). *A. ferrooxidans* is usually found in mixed cultures of thermophilic microorganisms in bioleaching processes, to enhance metal extraction in heaps (Watling

et al., 2014). The GSMM of this organism was reconstructed as part of the master thesis entitled "Genome-scale metabolic modeling of an extremophile microbial community" by Rui Barros Nunes.

***Desulfurococcus amylolyticus* DSM 16532**

Desulfurococcus amylolyticus DSM 16532 (*D. amylolyticus*) is an obligately anaerobic and hyperthermophilic archaeon. Growth is optimal at pH values between 6.0 and 6.5 and temperature values between 80 and 92°C (Perevalova et al., 2016). Can grow on a broad range of carbon sources (sugars, polysaccharides, and amino acids) but is one of the few Crenarchaeota able to grow in cellulose (Reischl et al., 2018).

***Pyrobaculum aerophilum* str. IM2**

Pyrobaculum aerophilum str. IM2 (*P. aerophilum*), in contrast with most species of the genus *Pyrobaculum,* is an aerobic, hyperthermophilic archaeon. Organic and inorganic compounds serve as substrates under both aerobic and anaerobic respiration. Growth temperatures range between 75 and 105°C and pH values between 5.8 and 9.0. Autotrophic growth is achieved by the oxidation of hydrogen or thiosulfate (Volkl et al., 1993).

***Sulfurihydrogenibium azorense* Az-Fu1**

Sulfurihydrogenibium azorense Az-Fu1 (*S. azorense*) is a gram-negative, thermophilic, chemolithoautotrophic, and microaerophilic bacterium (Lalonde et al., 2005). The bacterium grows optimally at 68°C, pH 6, and at low concentrations of NaCl and can also grow heterotrophically (Nakagawa et al., 2005) and use elemental sulfur, thiosulfate, hydrogen, and ferrous iron as energy sources (Aguiar et al., 2004). As an *Aquificales* member, *S. azorense* fixes CO_2 via reductive Tricarboxylic Acid Cycle (rTCA) (Hügler et al., 2007) to generate acetyl-CoA as an end product (Gupta et al., 2013). The GSMM of this organism was reconstructed in Chapter 3 of this thesis.

Thermodesulfovibrio yellowstonii DSM 11347

Thermodesulfovibrio yellowstonii DSM 11347 (*T. yellowstonii*) is a thermophilic sulfate-reducing, strictly anaerobic gram-negative bacterium. Growth was observed for temperatures between 40°C and 70°C and pH between 5.5 and 8.5. *T. yellowstonii* can use sulfate, thiosulfate, and sulfite terminal electron acceptors (Bhatnagar et al., 2015). In the presence of sulfate, growth was only observed with lactate, pyruvate, hydrogen plus acetate, or formate plus acetate. Pyruvate and lactate are oxidized to acetate (Henry et al., 1994).

Thermofilum adornatus strain 1910b

Thermofilum adornatus strain 1910b (*T. adornatus*) is a strict anaerobe archaeon. As moderately acidophilic and hyperthermophilic crenarchaea, *T. adornatus* requires various components of cells or culture broths of other Crenarchaeota, specifically culture broth filtrate of *Desulfurococcus* and *Pyrobaculum*. Growth was observed for temperatures between 50°C and 95°C, and pH between 5.3 and 8.5. Peptone, yeast extract, cellulose, starch, glucose, lactose, mannose, and pyruvate can be used as carbon sources. However, *T. adornatus* can utilize cellulose as the sole carbon and energy source (Zayulina et al., 2020).

Thermoplasma acidophilum DSM 1728

Thermoplasma acidophilum DSM 1728 (*T. acidophilum*) is a thermophilic, acidophilic archaeon lacking a cell wall. Growth takes place over a range of temperatures from 45°C to 62°C and a pH between 1 and 3.5 (Darland et al., 1970). Anaerobic growth is highly supported by the addition of sulfur which results in the accumulation of thiosulfate. Under fully aerobic conditions, sulfur compounds show no influence on growth. Yeast extract is required for growth in all conditions (Segerer et al., 1988).

Thermus scotoductus SA-01

Thermus scotoductus SA-01 (*T. scotoductus*) is a gram-negative aerobic thermophilic bacterium, which can grow at temperatures between 42°C and 73°C and pH values between 5.0 and 11. Is a

facultative anaerobe, reducing nitrate to nitrite anaerobically. Growth was observed in proline, glutamate, pyruvate, and sucrose (Kristjánsson et al., 1994).

6.3.1.1 Caldeira Velha (CV)

The metagenomic approaches used to determine the microbial composition predicted that nine microorganisms were present in this sample. However, there was experimental evidence of the presence of the genus (Sahm et al., 2013) of only seven of such microorganisms, which were therefore selected to proceed to the GSMM reconstruction process (Table 6.2). Nevertheless, it should be noted that inconsistencies in terms of optimal growth temperatures and pH values can be spotted concerning the sample global physicochemical properties for some of the predicted organisms.

Table 6.2. Caldeira Velha microbial community composition and organism physicochemical properties. Organisms' abundances were recalculated maintaining the relative abundances of the original data.

Organism	Domain	Temperature (ºC)	pH	Abundance
Ambiental conditions		**98**	**2.23**	
S. azorense Az-Fu1	Bacteria (Gram -)	50 - 73°C	5.8 - 9.0	61.4%
T. adornatus 1910b	Archaea	70 - 90°C	5.5 - 7.0	10.4%
T. yellowstonii DSM 11347	Bacteria (Gram -)	40 - 70°C	5.5 - 8.5	7.6%
A. caldus SM-1	Bacteria (Gram -)	32 - 52°C	1.0 - 3.5	7.3%
A. ferrooxidans DSM 10331	Bacteria (Gram +)	45 - 50°C	1.0 - 3.5	5.1%
T. acidophilum DSM 1728	Archaea	45 - 63°C	0.8 - 4.0	5.0%
P. aerophilum str. IM2	Archaea	75 - 104°C	5.0 - 7.0	3.1%

The taxonomic profiling algorithms varied only on the assignment of different members of the Pyrobaculum genus (*Pyrobaculum aerophlilum* str. IM2 and *Pyrobaculum islandicus* DSM 4184). These differences can be justified by the different reference genome databases in each approach. However, *Pyrobaculum aerophlilum* str. IM2, besides having more growth and metabolic information available (188

versus 64 hits on the Web of Science database), has more predicted genes (2518 versus 2043) and therefore it was selected for the GSMM reconstruction process. *Thermorodis peleae* and *Thiomonas intermedia K12* were excluded from the study, as there is no evidence of neither being present in such environments (Sahm et al., 2013).

6.3.1.2 Nascente Poente (NP)

The taxonomic profiling approaches predicted the presence of five microorganisms for this sample. Among them, are two different *Thermus* strains, *Thermus antranikianii* DSM 12462 and *Thermus scotoductus* SA-01. The two strains are phylogenetically very close (Chung et al., 2000; Lapierre et al., 2006), and therefore only one was selected for the GSMM reconstruction process. *Thermus scotoductus* SA-01 was selected (Table 6.3) to proceed with this study because it is reported to be one of the *Thermus* species isolated in the hot springs of the Azores (Santos et al., 1989).

Table 6.3. Nascente Poente microbial community composition and organism physicochemical properties. Organisms' abundances were recalculated maintaining the relative abundances of the original data.

Organism	Domain	Temperature (ºC)	pH	Abundance
Ambiental conditions		**95.8**	**6.88**	
P. aerophilum str. IM2	Archaea	75 - 104°C	5.0 - 7.0	79.9%
T. scotoductus SA-01	Bacteria (Gram -)	65 - 80°C	5.0 - 11	17.7%
D. amylolyticus DSM 16532	Archaea	68 - 97°C	5.7 - 7.5	1.28%
T. adornatus 1910b	Archaea	70 - 90°C	5.5 - 7.0	1.14%

Note that *T. adornatus* is characterized by a growth dependence on various components of cells or culture broths of other Crenarchaeota, specifically, culture broth filtrate of *Desulfurococcus* and *Pyrobaculum* species (Zayulina et al., 2020), which curiously were also predicted as present in this samples.

6.3.1.3 Esguicho de Maio (ESG)

Taxonomic profiling algorithms have predicted the presence of three microorganisms in this sample (Table 6.4). As stated for the CV sample, *Pyrobaculum aerophlilum* str. IM2, predicted by the read-based metagenomic approach was selected for the GSMM reconstruction process instead of *Pyrobaculum islandicus* DSM 4184, predicted by the assembly-based metagenomic approach.

Table 6.4. Esguicho de Maio microbial community composition and organism physicochemical properties. Organisms' abundances were recalculated maintaining the relative abundances of the original data.

Organism	Domain	Temperature (ºC)	pH	Abundance
Ambiental conditions		**98**	**7.29**	
P. aerophilum str. IM2	Archaea	75-104°C	5.0-7.0	97%
S. azorense Az-Fu1	Bacteria (Gram -)	50-73°C	5.8-9.0	1.7%
D. amylolyticus DSM 16532	Archaea	68-97°C	5.7-7.5	1.3%

6.3.2 Genome Annotation

All GSMM reconstructions started with a genome re-annotation step, supported by the phylogenic trees developed for each organism (Supplementary Figure 6.1 - 6.6) and *merlin*'s feature automatic workflow (Supplementary Table 6.1 - 6.6). The complete list of genes reviewed for each organism is available in Supplementary Table S1. Manual curation was performed and the number of genes in each model decreased in general by about 20% of the total metabolic genes due, for instance, to the removal of pseudo and truncated genes, incomplete EC numbers, and blocked reactions with encoded genes.

6.3.3 Biomass composition

The biomass macromolecular composition according to the domain and gram staining of each organism is presented in Table 6.5. The detailed biomass composition for each organism is available in Supplementary Table S2. When no experimental data were available for a given organism, the biomass

composition of the reference organisms *Escherichia coli* (*E. coli*) (Feist et al., 2007) for gram-negative bacteria, *Bacillus subtilis* (*B. subtilis*) (Dauner & Sauer, 2001) for gram-positive bacteria or *Methanosarcina fusaro* (*M. fusaro*) (Goyal et al., 2014) for archaea, was adapted. If the organism did not exhibit the enzymes responsible for producing one or more of the respective adapted compositions, the compounds were excluded from the biomass equation. The coefficients of the remaining compounds were recalculated, maintaining the relative abundances of the original data.

Table 6.5. Biomass macromolecular composition according to the organism domain and gram staining.

Biomass Composition (g gDW^{-1} (%))			
	Bacteria		Archaea
	Gram-negative	Gram-positive	
Protein	53.3	52.8	61.0
DNA	2.7	2.6	3.4
RNA	13.6	6.6	27.7
Lipids	2.9	7.6	5.3
Carbohydrates	10.7	3.1	1.0
Cell Wall	6.8	22.4	—
Cofactors	10	4.9	6.6
Total	100	100	100
Reference	(Feist et al., 2007)	(Dauner & Sauer, 2001)	(Goyal et al., 2014)

Amino acid, deoxynucleotide, and nucleotide composition were calculated based on the genome's information for each organism using the e-Biomass feature in *merlin*.

The lipid and carbohydrate compositions were adapted from the gram-negative bacterium *Escherichia coli* (*E. coli*) (Feist et al., 2007), the gram-negative bacterium *Bacillus subtilis* (*B. subtilis*) (Dauner & Sauer, 2001) or the archaeon *Methanosarcina fusaro* (*M. fusaro*) (Goyal et al., 2014) depending on the organism domain and gram staining.

The average fatty acid composition for organisms of the domain bacteria was retrieved from experimental data available for each organism (Supplementary Table S3).

Exception made for the archaeon *Thermoplasma acidophilum,* which has a unique cellular membrane (Smith et al., 1973), cell wall compositions were reconciled between the KEGG reactions assigned through *merlin*'s annotation to its biosynthesis pathway (Supplementary Table S4) and the cell wall compositions of the respective domain and gram staining.

Cofactor composition was based on the study of universal essential cofactors in prokaryotes (Xavier et al., 2017).

The growth-associated energy (GAM) and non-growth-associated energy (NGAM) requirements have not been experimentally determined yet for any of the organisms present in this study. Therefore, for all organisms, the GAM requirements were estimated according to (Thiele et al., 2010) and based on data for *E.coli* (Neidhardt et al., 1990). The NGAM requirements of 8.39 mmol $_{ATP}$ gDW^{-1} h^{-1} were adopted from *E. coli* (Feist et al., 2007).

6.3.4 Metabolic Models

All genome-scale metabolic reconstructions were generated through a bottom-up approach. A summary of the final metabolic reconstructions is presented in Table 6.6.

Table 6.6. Metabolic information of the final Genome-Scale Metabolic Reconstructions. Compartments are divided into c – cytosol, p – periplasm, and e – extracellular. Archaea organisms include a pseudo-compartment to simulate the proton motive force (PMF).

Organism	Genes	Reactions	Metabolites	Subsystems	Compartments	Gene Ratio (%)
D. amylolyticus	188	494	417	81	2 (c, e)	13
P. aerophilum	284	646	508	88	2 (c, e)	12
T. yellowstonii	326	727	656	88	3 (c, p, e)	17
T. adornatus	225	555	436	76	2 (c, e)	12
T. acidophilum	292	691	543	89	2 (c, e)	19
T. scotoductus	412	979	766	96	3 (c, p, e)	17

Some reactions without GPR associations were included in the reconstructions and those comprise spontaneous, exchange, outer membrane transport, and diffusion transport reactions of metabolites, such as CO_2, water, or O_2. Final GSMMs are available as an SBML version 3 file, and all detailed metabolic information is included on the GitHub repository SophiaSantos/DesignOptimizationMicrobialCommunities/models/models_azores/.

6.3.5 Models' validation

6.3.5.1 Environmental Conditions

The validation of each model involved using a defined minimal medium. In the cases of microorganisms that require a rich medium to grow, a minimal defined medium was set according to the physiological and metabolic information (Table 6.7).

When relevant aerobic and anaerobic conditions were tested. All medium components were allowed to enter the system unconstrained, except for carbon sources and amino acids, which were constrained according to literature data for each organism. Carbon sources present in Table 6.7 were all tested. For the microorganisms able to grow aerobically and anaerobically two different biomass equations were included in the model to better evaluate growth under both conditions. These reactions, "R_Biomass__cytop" and "R_Biomass_anaerobic__cytop", differ in the presence or absence of heme in the Cofactor composition respectively. Oxygen (O_2) and ferrous iron (Fe^{2+}) were supplied under aerobic conditions.

For microbial community simulations, the environmental conditions applied for each community were defined by a gathering of the compounds set for individual organisms present in the specific community. No amino acids were included in the defined environmental conditions. When possible glucose was the carbon source preferred. Medium components were allowed to enter the system unconstrained, except for carbon sources, which were constrained according to literature data for each organism.

Table 6.7. Minimal medium composition for each condition tested: chemolithoautotrophic and heterotrophic growth. Oxygen and Ferrous iron (highlighted in grey) were only supplied under aerophilic conditions.

Organism	Defined Medium	Carbon sources	Aerobic/Anaerobic
D. amylolyticus	Ammonia, Sulfur, Orthophosphate, Folate, Nicotinate, and Hydrogen	Cellulose, Glucose, Fructose, and CO_2	Anaerobic
P. aerophilum	Ammonia, Sulfate, Orthophosphate, Folate, Nicotinate, Cyanide ion, and Hydrogen	CO_2, acetate, and Casaminoacids	Oxygen and Fe^{2+}
T. yellowstonii	Ammonia, Sulfate, Orthophosphate, Hydrogen, Cyanide ion	Lactate, Pyruvate, Acetate	Anaerobic
T. adornatus	Ammonia, Sulfate, Orthophosphate, Nicotinate, Thymine, Coenzyme A, Riboflavin, Folate, Pyridoxal Phosphate, GTP, L-Arginine, L-Histidine, L-Isoleucine, L-Leucine, L-Lysine, L-Phenylalanine, L-Tryptophane, L-Tyrosine, L-Valine	Glucose, Cellulose	Anaerobic
T. acidophilum	Ammonia, Sulfur, Orthophosphate, Riboflavin, Nicotinate, L-Histidine, L-Leucine, L-Lysine, L-Valine	Glucose	Oxygen and Fe^{2+}
T. scotoductus	Nitrate, Sulfate, Orthophosphate	Glucose, Proline, Glutamate, Maltose	Oxygen and Fe^{2+}

6.3.5.2 Modeling Simulations

Growth under the different carbon sources reported in the literature was simulated and *in silico* growth was compared to data in the literature. Whenever possible, growth under aerobic and anaerobic conditions was simulated. Simulation under different electron acceptors was not screened.

Overall, the results of the pFBA predictions match the data retrieved from the literature specific for each organism in the different conditions tested, showing, when available, a predicted specific growth rate in the same range as reported in the literature. The organisms *T. acidophilum*, *T. scotoductus,* and *P. aerophilum* are aerobic but able to grow under anaerobic conditions. The predictions also showed the

ability of those organisms to grow in the absence of oxygen. Simulation under different sulfur and nitrogen sources or byproduct predictions were not screened due to the lack of information in the literature for most of the organisms.

6.3.6 Community Simulation

Samples of CV, ESG, and NP were selected to proceed into a community simulation process. SMETANA, FBA/pFBA, and SteadyCom were used to simulate all three samples. These tools have distinct outputs, which together complement and validate each other (as shown in Chapter 4 of this thesis). Thus, when possible, the use of different tools is recommended. SMETANA proved to predict possible pairwise interactions, FBA/pFBA and SteadyCom showed to provide good results on predicting individual and community-specific growth rate, as well as interspecies flux distributions. SteadyCom also adds information on species abundances within the community, which is relevant to the communities studied here.

6.3.6.1 SteadyCom

SteadyCom (Chan et al., 2017) predicts the metabolic flux distribution and relative abundance of each species in a community. However, when multiple organisms are growing, there is not a constant growth rate for all microbes and therefore the fastest-growing organism can outgrow the rest of the community members. Although SteadyCom imposes a steady-state condition, that includes a restriction to force zero flux through an organism with zero abundance to avoid such situations. In these cases, the fast-growing organism achieves the maximum abundance percentage of 1, and fluxes are computed for that organism, while other organisms' abundance is set to 0, and no fluxes are computed.

Table 6.8. Models' validation against experimental conditions from literature.

Organisms	*D. amylolyticus*		*P. aerophilum*		*T. yellowstonii*		*T. adornatus*		*T. acidophilum*		*T. scotoductus*	
Carbon Source	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})	**Observed Growth** (h^{-1})	**Predicted Growth** (h^{-1})
Glucose	0.059	0.064	———	———	———	———	0.139	0.219	0.062	0.069	0.415	0.447
CO_2	0.004	0.005	0.102	0.145	———	———	———	———	———	———	———	———
Cellulose	0.059	0.100	———	———	———	———	Growth	0.487	———	———	———	———
Acetate	———	———	0.187	0.201	Growth	0.019	———	———	———	———	———	———
Fructose	0.038	0.024	———	———	———	———	———	———	———	———	———	———
Maltose	———	———	———	———	———	———	———	———	———	———	Growth	0.894
Lactate	———	———	———	———	0.029	0.031	———	———	———	———	———	———
Pyruvate	———	———	———	———	Growth	0.042	———	———	———	———	———	———
Casaminoacids	———	———	0.258	0.398	———	———	———	———	———	———	———	———
L-Proline	———	———	———	———	———	———	———	———	———	———	Growth	0.373
L-Glutamate	———	———	———	———	———	———	———	———	———	———	Growth	0.373
Aerobic/Anaerobic												
Aerobic	No growth	No growth			No growth	No growth	No growth	No growth				
Anaerobic			Growth	0.058					Growth	0.114	Growth	0.151

For our case studies, the organisms indeed show different growth rate ranges (from 0.059 to 0.415 h-1 (Table 6.8)), which impairs the prediction of species abundancies and consequent flux distribution on these samples using SteadyCom.

6.3.6.2 SMETANA

SMETANA estimates the interaction potential of the species in a microbial community and returns the probability of inter-species metabolite exchange. No growth and flux rates are directly obtained through SMETANA.

SMETANA calculates two scores to predict the species' interaction potential: the Metabolic Resource Overlap (MRO) and the Metabolic Interaction Potential (MIP), and a SMETANA score that evaluates the growth dependency of species A on metabolite m produced by species B.

MRO and MIP represent opposite circumstances. While the first indicates the predisposition to competition between the organisms, as both require the same metabolite(s) from the environment, MIP represents the tendency of the community's organisms to depend on each other, not being able to grow on their own. Results showed that for the three samples, SMETANA's MIP score was not available and the MRO score was about 0.14, both indicating that, for samples, all organisms barely depend on the others to grow.

SMETANA score identifies pairwise interactions, independent of the other organisms in the community. All individual GSMMs were reconstructed using *merlin* (Dias et al., 2018), which relies on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2016), for metabolites and reactions identifiers. However, SMETANA requires BiGG metabolite identifiers, thus these exchange metabolites identifiers were converted. The SMETANA species interaction potential and potential inter-species interactions are available in Table 6.9. Only inter-species interactions with values greater than 0.5 were considered. While analyzing possible inter-species interactions patterns emerged. *A. ferrooxidans* acts as a Fe^{2+} donor while *A. caldus* acts mainly as a sulfur donor. These results compare with information from the literature that reports the roles of *A. caldus* as one of the dominant sulfur-oxidizing bacteria in bioleaching reactors, together with *A. ferrooxidans* as an iron-oxidizing bacterium (Watkin et al., 2009; Watling et al., 2014).

Table 6.9. Pairwise interactions predicted by the SMETANA simulation method. Only inter-species interactions with values greater than 0.5 were considered.

Receiver	Metabolites	Donnor
A. caldus	Fe^{3+}	*A. ferrooxidans*
A. ferrooxidans	Hydrogen Sulfide	*A. caldus*
P. aerophilum	L-Cysteine	*A. caldus*
	CO_2, Ammonia	*T. scotoductus*
D. amylolyticus	L-Cysteine	*P. aerophilum*
	Riboflavin, CO_2, Ammonia	*T. scotoductus*
S. azorense	Fe^{3+}, Thiosulfate	*A. caldus*
	Fe^{3+}	*A. ferrooxidans*
	Hydrogen sulfide	*D. amylolyticus*
T. acidophilum	Sulfur	*A. caldus*
	L-Histidine, L- Isoleucine, L-Leucine, L-Lysine, L-Valine	*P. aerophilum*
T. adornatus	L-Cysteine, Thiosulfate	*A. caldus*
	L-Histidine, L- Isoleucine, L-Leucine, L-Lysine, L-Phenylalanine, L-Tryptophane, L-Tyrosine, L-Valine	*P. aerophilum*
	Cellulose	*S. azorense*
	Riboflavin	*T. scotoductus*
T. yellowstonii	Hydrogen sulfide, Ammonia	*A. caldus*
	Hydrogen sulfide Ammonia	*A. ferrooxidans*

Regardless of no literature evidence, SMETANA's predictions show the relevant role of *P. aerophilum* as an amino acid donor for organisms with amino acid auxotrophies such as *T. adornatus* (Zayulina et al., 2020) and *T. acidophilum* (Segerer et al., 1988). *P. aerophilum* has the metabolic

capability to produce all amino acids, corroborating these findings. Also, results show that *T. scotoductus* is the only organism with a donor character, with a specific riboflavin production role in the presence of the archaeons *T. adornatus* and *D. amylolyticus.* SMETANA predicts *T. adornatus* as a receiver organism, which is supported by experimental evidence demonstrating its growth dependence on culture broth filtrate of *Desulfurococcus* and *Pyrobaculum* (Zayulina et al., 2020).

Notably, the presence of *T. adornatus*, which uses cellulose as a carbon source (Zayulina et al., 2020), seems to trigger *S. azorense* to produce cellulose. These results show a possible route of cellulose production by *S. azorense* that should be further investigated using other simulation and optimization methods.

6.3.6.3 FBA

Initially, a community model was created using the REFRAMED Python package, using the *Community* function, providing the individual GSMMs as input and getting a compartmentalized model as output.

As no experimental data from the literature was found on specific growth rates for community growth, only intraspecies interactions were analyzed. Driven by SMETANA's predictions, samples CV and ESG were essentially screened for the potential production of cellulose by *S. azorense,* while sample NP was screened for the potential production of amino acids by *P. aerophilum*. No amino acids were included in the environmental constraints to test the hypothesis of *P. aerophilum* amino acid production for the other organisms in the community.

ESG's sample is composed of three organisms, *D. amylolyticus*, *P. aerophilum,* and *S. azorense*. Of the three organisms present in this sample, *D. amylolyticus* is the more demanding, being auxotrophic for some amino acids (Perevalova et al., 2016). Moreover, *D. amylolyticus* uses cellulose as a carbon source, which might trigger *S. azorense* to produce cellulose. Analyzing the pFBA results (Figure 6.1), *P. aerophilum* is indeed producing the essential amino acids for *D. amylolyticus'* growth, as well as amino acids for *S. azorense's* growth. However, *D. amylolyticus* uses glucose as a carbon source, and no cellulose is produced by *S. azorense*. Even in nitrogen-limiting conditions, which trigger *S. azorense* to produce cellulose when growing in isolation, no cellulose is predicted to be produced. These results

corroborate SMETANA's predictions showing *P. aerophilum's* capacity to produce amino acids and the inability of *D. amylolyticus* to trigger *S. azorense* to produce cellulose.

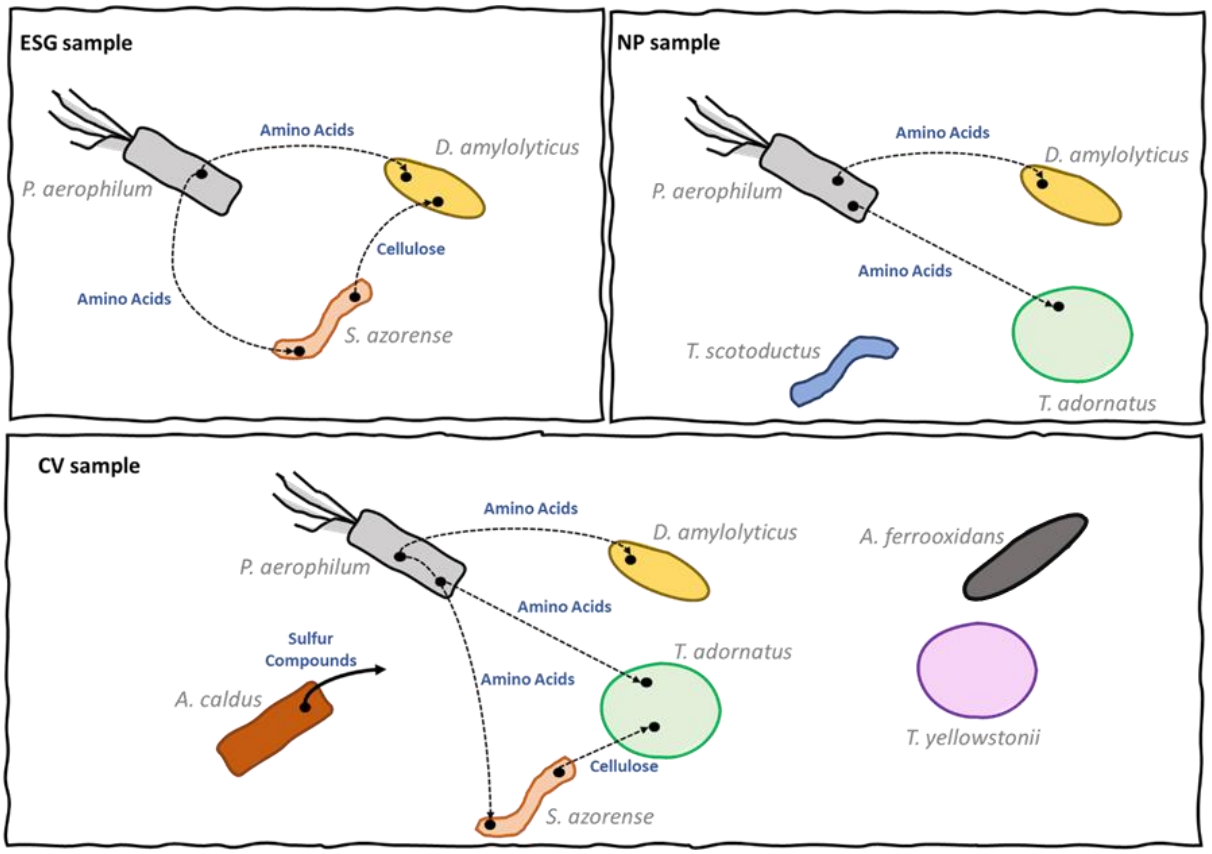


Figure 6.1. Schematic representation of the main metabolite interactions, predicted by the pFBA simulation method, between the organisms present in the three analyzed samples. Dashed lines correspond to known metabolite exchange routes; Solid line correspond to unknown metabolite exchange routes.

NP sample is composed of 4 organisms, *P. aerophilum*, *D. amylolyticus, T. adornatus,* and *T. scotoductus*. The pFBA results once more showed that *P. aerophilum* has an essential role in the production of amino acids when growing in the community, not only acting as an amino acid donor to the organism *D. amylolyticus*, as identified in the ESG sample but also an amino acid donor for *T. adornatus* that is reported to be highly dependent on culture broth filtrate of *Desulfurococcus* and *Pyrobaculum* (Zayulina et al., 2020) species (Figure 6.1). As predicted by SMETANA, the pFBA results show that *T. scotoductus* seems to have a neutral role in the community, not predicting any possible interactions with the rest of the organisms in this sample.

The pFBA interspecies interactions predictions obtained for the CV sample (Figure 6.1), which is composed of *S. azorense*, *T. adornatus*, *T. yellowstonii*, *A. caldus*, *A. ferrooxidans*, *T. acidophilum,* and *P. aerophilum*, follow the same pattern of the predictions for other samples, in which *P. aerophilum* is the main donor, producing amino acids for all other organisms. *A. caldus* is also producing high amounts of sulfur compounds (Watling et al., 2014). However, the consuming route is not easy to follow. Notably, in the pFBA analysis of this sample, cellulose is produced by *S. azorense*. The ability to produce cellulose seems to be triggered by the presence of *T. adornatus*, as it consumes all cellulose produced by *S. azorense*, which also supports SMETANA's predictions.

More importantly, predictions of SMETANA and pFBA methods for these communities showed accordance with the possible path for cellulose production by *S. azorense*, so these results will be screened using community optimization tools.

6.3.7 Community Optimization

In Chapter 3 of this thesis, the individual metabolic capability of *S. azorense* cellulose production was screened using a metabolic engineering optimization supported by evolutionary algorithms. This analysis did not return solutions robust enough to increase cellulose production. Here, with the information gathered from the simulation predictions, a community metabolic engineering optimization using MEWpy was tested. The simulation results showed the potential production of cellulose by *S. azorense* when co-cultured with *T. adornatus*, thus such a hypothesis was further investigated. A second hypothesis was tested using a co-culture of *S. azorense* and *Escherichia coli* K-12 MG1655 (*E. coli*), using the iAF1260 GSMM (Feist et al., 2007), as *E. coli* is one of the most studied organisms and one of the best suitable metabolic engineering platforms (Chen et al., 2013). Therefore, the optimization process in this co-culture case was restricted to *E. coli* reactions.

An FVA analysis was performed, before the optimization, to understand the community models solution space regarding cellulose production under stress conditions (nitrogen limitation conditions) (Table 3.10). Under these conditions, *S. azorense* exhibited cellulose production capability when growing in isolation. For each co-culture, the total consumption of the carbon source was imposed, and the specific growth rate was set to at least 10% of the specific growth rate obtained with pFBA simulation under nitrogen-limiting conditions. FVA results of the individual growth of *S. azorense* showed that none of the

tested co-cultures required mandatory cellulose production. However, a high amount of cellulose might be produced using the *S. azorense* and *E. coli* co-culture.

Table 6.10. FVA analysis of *S. azorense* cellulose production capabilities using a 2 organisms community. Total consumption of the carbon source by *S. azorense* was imposed, and the specific growth rate was set to at least 10% of the specific growth rate obtained with pFBA simulation under nitrogen-limiting conditions.

Cellulose Production under N-limiting Conditions (mmol g_{DW}^{-1} h^{-1})	
Community	
S. azorense* and *T. adornatus	[0.0, 0.013]
S. azorense* and *E. coli	[0.0, 1.119]

MEWpy was used for the validation of both co-culture cellulose optimization. All reactions in the co-culture of *S. azorense* and *T. adornatus* community model were defined as targets. Whereas, for the *S. azorense* and *E. coli* co-culture, iAF1260 reactions (*E. coli* GSMM identifier) were defined as a target. In both cases, essential and exchange reactions were removed. All optimizations were run for a maximum of 100 generations and a maximum size of two for the candidate, through an RKOProblem (maximum number of reaction deletions). Jupyter Notebooks with the whole process are available in Supplementary Material – saz_tac_optimization.ipynb and saz_eco_optimization.ipynb.

When performing the reaction deletion optimization, although a maximum of two modifications was set, the best solutions in the *S. azorense* and *T. adornatus* co-culture included only one modification. All best solutions, in this case, were evaluated by a pFBA simulation, showing that indeed *S. azorense* produces cellulose, but all cellulose produced is consumed by *T. adornatus*, raising doubt about the effectiveness of this co-culture to produce cellulose in high amounts.

In the case of the community formed by *S. azorense* and *E. coli* (Figure 6.2) when reaction deletions are performed in *E. coli*, a set of two modifications was always obtained, although when performing a pFBA evaluation of the modifications cellulose production by *S. azorense* was achieved using only one of the modifications.

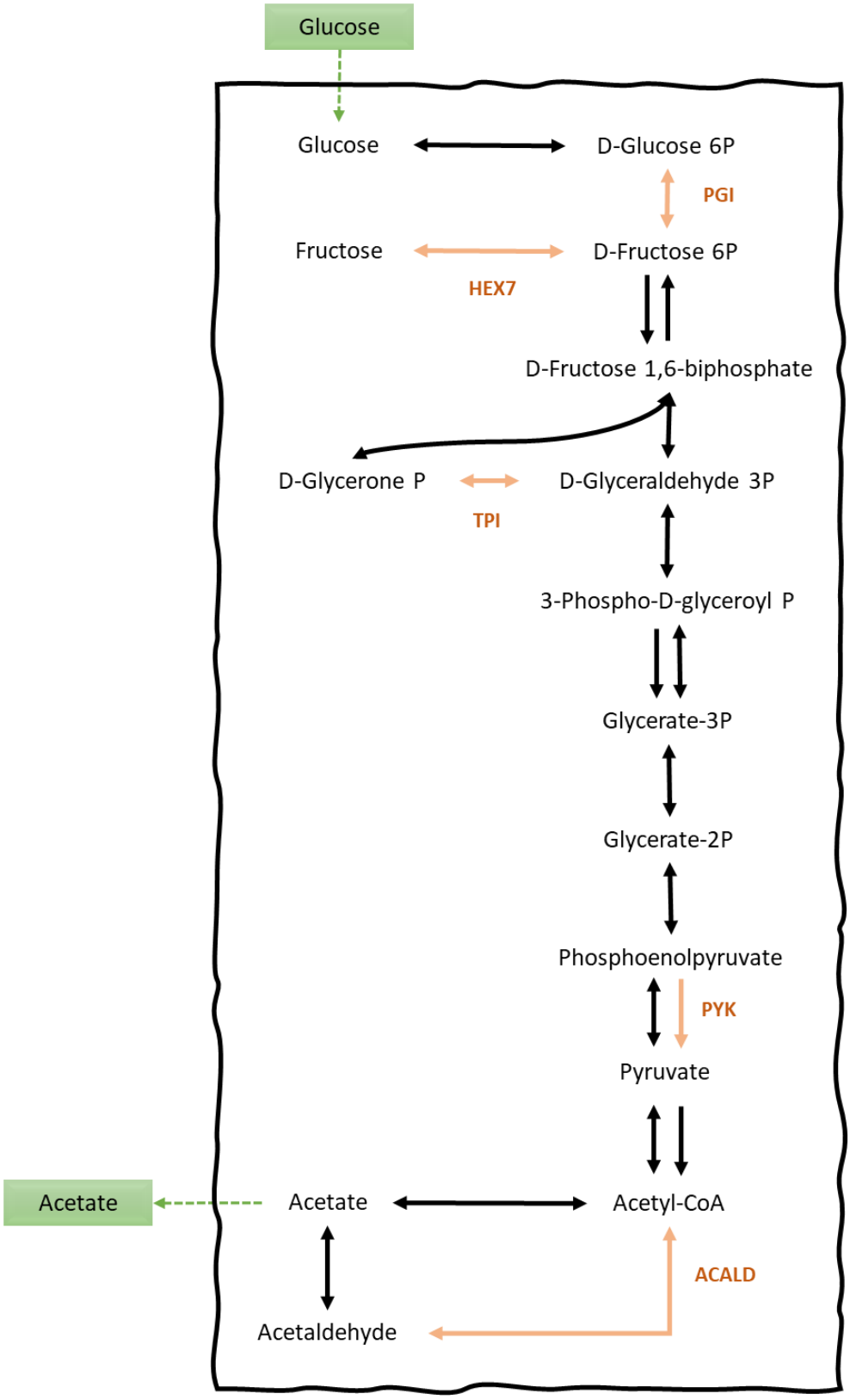


Figure 6.2. Schematic representation of *E. coli* K12 MG1655 Glycolisis Pathway deletions predicted using MEWpy to improve *S. azorense* Az-Fu1 cellulose production when growing in a co-culture system. For each one of the predicted deletions (orange), acetate is produced by *E. coli* K12 MG1655. ACALD - acetaldehyde dehydrogenase (acetylating), HEX7 – hexokinase, PGI - glucose-6-phosphate isomerase, PYK - pyruvate kinase, TPI - triose-phosphate isomerase.

All reaction deletions are predicted to belong to *E. coli's* central carbon metabolism (Table 6.11), which directly influences a higher production of acetate. Acetate is then consumed by *S. azorense*, resulting in cellulose production.

Table 6.11. Reaction KO analysis of the community formed by *S. azorense and E. coli for* cellulose production capabilities. Total consumption of the carbon source by *S. azorense* was imposed, and the specific growth rate was set to at least 10% of the specific growth rate obtained with pFBA simulation under nitrogen-limiting conditions. pFBA simulations were performed under nitrogen-limiting conditions. ACALD - acetaldehyde dehydrogenase (acetylating), HEX7 – hexokinase, PGI - glucose-6-phosphate isomerase, PYK - pyruvate kinase, TPI - triose-phosphate isomerase.

Cellulose Production under N-limiting Conditions (mmol g_{DW}^{-1} h^{-1})	
Reaction Knock-Out	**Cellulose production**
PYK	0.492
ACALD	0.536
PGI	0.492
HEX7	0.561
TPI	0.580

Considering these results, a new pFBA analysis was performed to test the *S. azorense* cellulose production in the presence of acetate. When growing in isolation, *S. azorense* cellulose production capabilities improve when acetate consumption is forced, suggesting that the cellulose production mechanism is highly dependent on unknown factors.

Although the optimization results show cellulose production potential, experimental validation is still required. In this specific case, the validation of these results can face some difficulties as *E. coli* and *S. azorense* have distinct optimal growth temperature ranges (23-40°C and 50-73°C, respectively). Studies have already tried to adapt *E. coli* to high temperatures (Rudolph et al., 2010), which can be a route for the experimental validation of the hypotheses here suggested the optimization of *S. azorense* cellulose production.

6.4 Conclusions

A microbial community study of extreme environments from the hydrothermal sites in São Miguel, Azores, using GSMM reconstructions was performed. Three of the five hydrothermal samples analyzed in Chapter 2 were characterized in terms of predicted organisms and the GSMM for each of the present organisms was reconstructed and validated with data available in the literature. The reconstructed GSMMs can be used independently for investigating the extraordinary microbial metabolism in extreme environments.

Simulation methods with validated applications to microbial communities, such as FBA/pFBA (Stolyar et al., 2007) and SMETANA (Zelezniak et al., 2015) were used to predict the possible interactions between the organisms in such environments. Simulations considered the co-occurrence in the samples ESG, NP, and CV. The predictions have elucidated two main metabolic roles within the specific communities: *P. aerophilum* as an amino acid donor and *T. adornatus* as a trigger for *S. azorense* cellulose production. *S. azorense* has been reported to produce sufficient amounts of exopolysaccharides under stress conditions (Lalonde et al., 2005), and the metabolic capabilities of *S. azorense* production of cellulose presented by its GSMM, under nitrogen-limiting conditions, were described in Chapter 3 of this thesis. Although optimization of cellulose production using evolutionary algorithms did not return robust solutions when *S. azorense* was growing in isolation, results of the simulations in a community context showed some testable hypotheses.

Using MEWpy for the *in silico* design of microbial communities using GSMMs and evolutionary algorithms, the strain optimization of the co-cultures *S. azorense* with *T. adornatus* and *S. azorense* with *E. coli* was performed. The co-culture of *S. azorense* and *E. coli* was tested since *E. coli* is one of the most studied organisms and one of the best suitable metabolic engineering platforms (Chen et al., 2013), and MEWpy allows performing optimization in a specific organism within a community, as described in Chapter 5.

The community optimization results showed that indeed *T. adornatus* triggers *S. azorense* to produce cellulose. However, the cellulose is consumed by *T. adornatus*, raising doubt about the effectiveness of this co-culture to produce cellulose in high amounts. Moreover, when reaction deletions are performed in *E. coli* within its co-culture with *S. azorense,* cellulose production by *S. azorense* is

achieved. The deleted *E. coli* reactions have a specific role in increasing acetate production, which then is consumed by *S. azorense*, revealing a significant part of the metabolic cellulose production route.

Although the optimization results here reported using MEWpy demonstrate its applicability in the design of microbial communities and *S. azorense* as a possible cellulose producer, experimental validation is still required.

6.5 Supplementary Material

Additional file in IPYNB format: saz_tac_optimization.ipynb and saz_eco_optimization.ipynb

Link: DesignOptimizationMicrobialCommunities/JupyterNootbooks/

Additional files in SBML format: acaldus.xml, aferrooxidans.xml, damylolyticus.xml, paerophilum.xml, sazorense.xml, tacidophilum.xml, tadornatus.xml, tscotoductus.xml and tyellowstonii.xml

Link: DesignOptimizationMicrobialCommunities/models/models_azores/

Additional files in SBML format: community_CV.xml, community_NP.xml and community_ESG.xml

Link: DesignOptimizationMicrobialCommunities/models/community_models/

Additional file in Excel format: Chapter6_Supplementary_Material.xlsx

Link: DesignOptimizationMicrobialCommunities/Data

Table S1 Annotation of the genes present in each model

Table S2 Biomass composition in mmol of molecules per gram of biomass. Molecular weight in green background cells was calculated using the fatty-acyl Coa as the R group in lipids. Amino acids' molecular weight does not include a water molecule. Nucleotides' molecular weight does not include diphosphate molecule.

Table S3 Average lipid and fatty acid compositions for each organism.

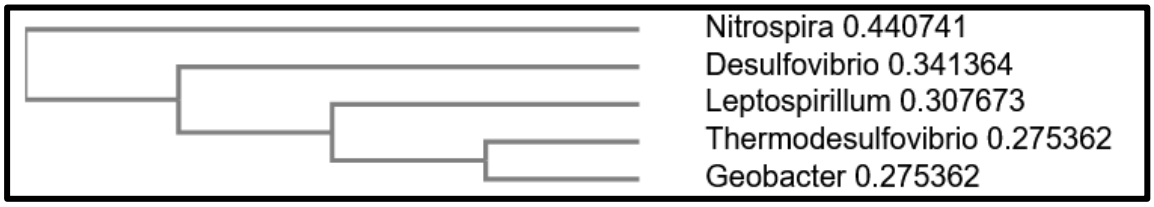
Table S4 Cell wall composition for each organism.

Table S5 Genes included in each model.

Table S6 Reactions included in each model, including Gene-Protein-Reaction associations.

Table S7 Metabolites included in each model

Thermodesulfovibrio yellowstonii

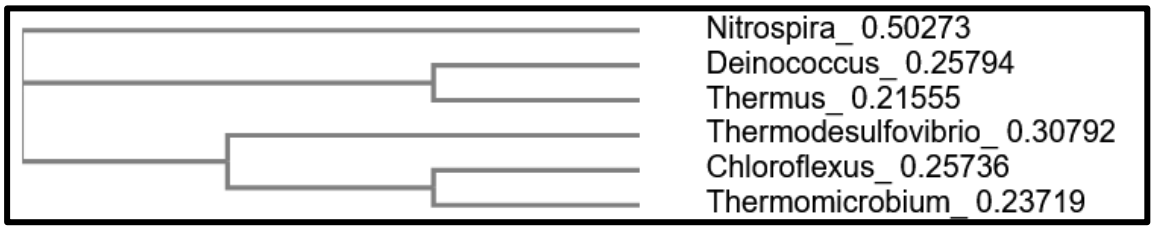


Supplementary Figure 6.1. Phylogenetic tree of *Thermodesulfovibrio yellowstonii* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.1 List of phylogenetic similar organisms/genus to *Thermodesulfovibrio yellowstonii* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Thermodesulfovibrio yellowstonii (strain ATCC 51303 / DSM 11347 / YP87)	A
genus *Thermodesulfovibrio*	B
genus *Geobacter*	C
genus *Leptospirillum*	D
genus *Desulfovibrio*	E
genus *Nitrospira*	F

Thermus scotoductus

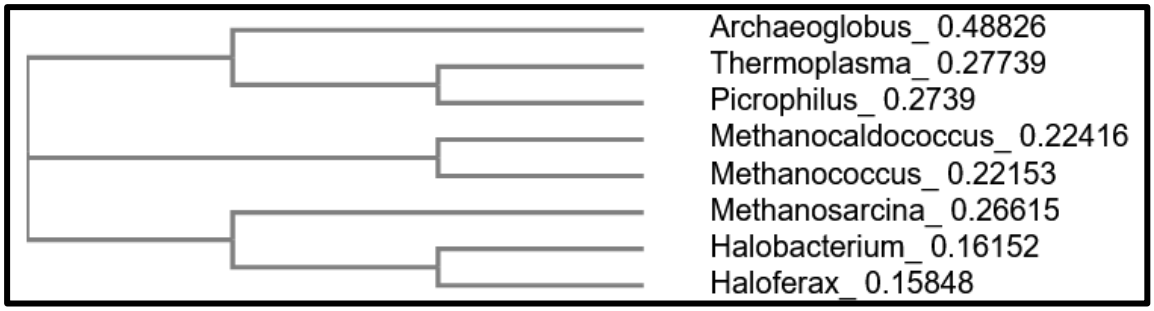


Supplementary Figure 6.2. Phylogenetic tree of *Thermus scotoductus* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.2. List of phylogenetic similar organisms/genus to *Thermus scotoductus* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Thermus scotoductus (strain ATCC 700910 / SA-01)	A
genus *Thermus*	B
genus *Deinococcus*	C
genus *Thermodesulfovibrio*	D
genus *Thermomicrobium*	E
genus *Chloroflexus*	F
genus *Nitrospira*	G

Thermoplasma acidophilum

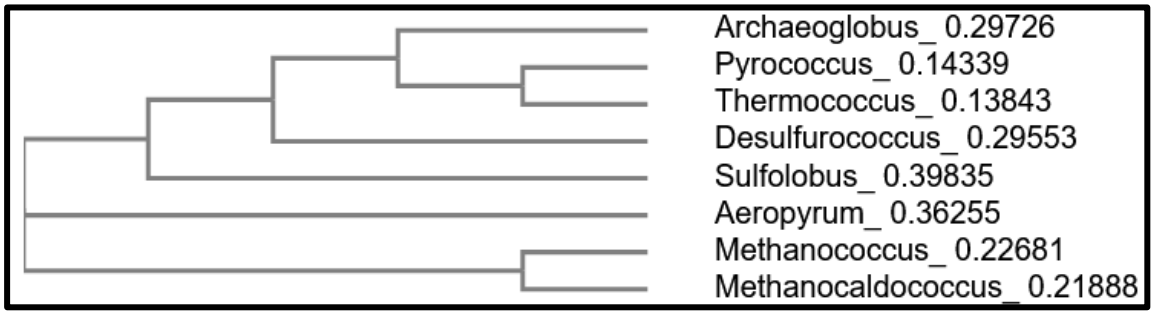


Supplementary Figure 6.3. Phylogenetic tree of *Thermoplasma acidophilum* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.3. List of phylogenetic similar organisms/genus to *Thermoplasma acidophilum* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Thermoplasma acidophilum	A
genus *Thermoplasma*	B
genus *Picrophilus*	C
genus *Archaeoglobus*	D
genus *Methanocaldococcus*	E
genus *Methanococcus*	F
genus *Methanosarcina*	G
genus *Halobacterium*	H
genus *Haloferax*	I

Desulfurococcus amylolyticus

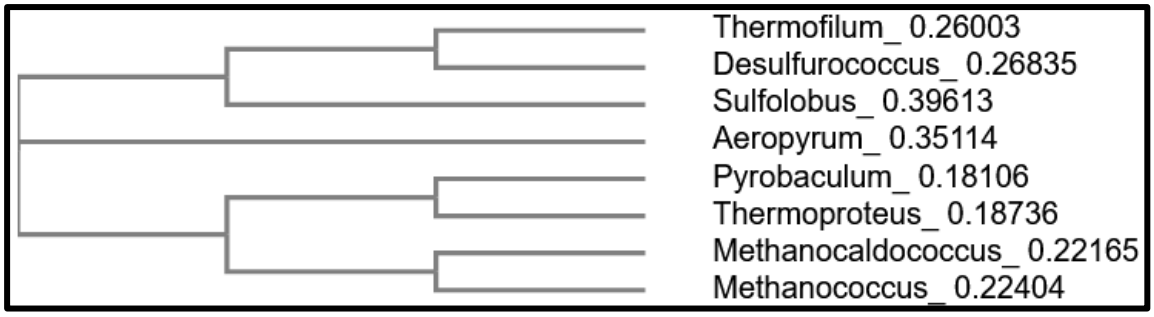


Supplementary Figure 6.4. Phylogenetic tree of *Desulfurococcus amylolyticus* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.4. List of phylogenetic similar organisms/genus to *Desulfurococcus amylolyticus* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Desulfurococcus amylolyticus (strain DSM 18924 / JCM 16383 / VKM B-2413 / 1221n)	A
genus *Desulfurococcus*	B
genus *Thermococcus*	C
genus *Pyrococcus*	D
genus *Archaeoglobus*	E
genus *Sulfolobus*	F
genus *Aeropyrum*	G
genus *Methanococcus*	H
genus *Methanocaldococcus*	I

Pyrobaculum aerophilum

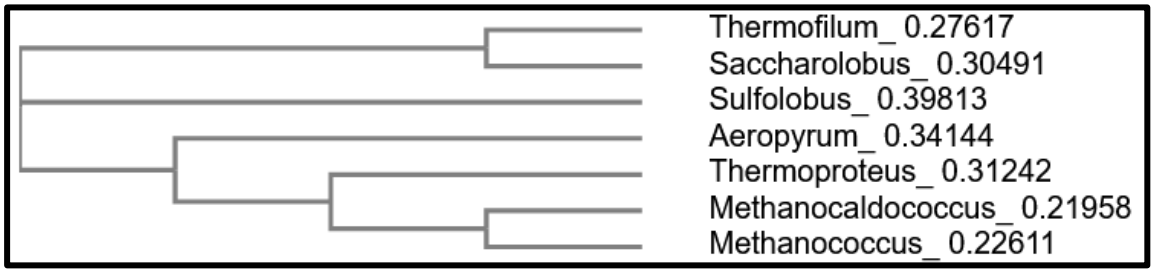


Supplementary Figure 6.5. Phylogenetic tree of *Pyrobaculum aerophilum* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.5. List of phylogenetic similar organisms/genus to *Pyrobaculum aerophilum* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Pyrobaculum aerophilum	A
genus *Pyrobaculum*	B
genus *Thermoproteus*	C
genus *Methanocaldococcus*	D
genus *Methanococcus*	E
genus *Aeropyrum*	F
genus *Sulfolobus*	G
genus *Desulfurococcus*	H
genus *Thermofilum*	I

Thermofilum adornatus



Supplementary Figure 6.6. Phylogenetic tree of *Thermofilum adornatus* relative genus. This tree was built using the EMBL-EBI Clustal OMEGA multiple sequence alignment tool. Numbers in front of each genus represent the branch lengths to each node generated automatically by the tool using the Neighbor-joining method.

Supplementary Table 6.6. List of phylogenetic similar organisms/genus to *Thermofilum adornatus* given to the automatic workflow feature in *merlin*.

Organism	Confidence level
Thermofilum adornatus 1505	A
genus *Thermofilum*	B
genus *Saccharolobus*	C
genus *Sulfolobus*	D
genus *Aeropyrum*	E
genus *Thermoproteus*	F
genus *Methanocaldococcus*	G
genus *Methanococcus*	H

Chapter 7

Conclusions and Future Perspectives

This final chapter includes the overall conclusions achieved by the research conducted in this thesis, as well as some perspectives on future research to address unanswered and new questions raised throughout this work.

7.1 Overall outcomes

This thesis had as a major objective to give insights into the use of computational methods for the rational design of microbial communities, using extremophilic microbial communities from hydrothermal sites in the Azores as a case study. Overall, the main goals were achieved and the main conclusions obtained throughout the developed work are as follows:

- The review of the literature has shown the crucial role of microbial communities throughout Earth' biosphere(Gilbert et al., 2014; Rusch et al., 2007; Turnbaugh et al., 2007), especially the ones in extremophilic environments, due to their remarkable genomic and metabolic attributes (Durvasula et al., 2018) and industrial applications. To support the understanding of functions and interactions within these complex systems, the applications of metabolic modeling throughout the use of GSMMs in a microbial community context are increasing every day, with emphasis on the prediction of microbe-microbe/host-microbe interactions (Almut Heinken et al., 2020; Zelezniak et al., 2015), as well as some studies on the design and engineering of microbial communities (García-Jiménez et al., 2018; Pacheco & Segrè, 2021). However, some limitations and challenges are known, such as the accurate identification of organisms in a microbial community, reconstruction of GSMMs with good phenotypic predictions, and the availability of experimental data for the validation of simulation and optimization methods.
- The prokaryotic diversity of five samples from hydrothermal vents at São Miguel, Azores, was characterized using assembly-based and read-based taxonomic profiling algorithms. Both taxonomic profiling approaches presented very similar results, demonstrating that the two approaches validate and complement each other and indeed both should be applied whenever possible (Quince et al., 2017). Differences spotted on the assignment of different members of the *Pyrobaculum* genus, depending on the profiling algorithm, are justified by the use of different reference genome databases. However, results showed that a significant part of the prokaryotic diversity present in the five samples was not identified due to the still large amount of unculturable organisms without a reference genome available in databases (Zorrilla et al., 2021).
- Taxonomic profiling algorithms predicted a high abundance of *Aquificales* and *Crenarchaeota* members in three of the five samples analyzed (Caldeira Velha, Esguicho de Maio, and Nascente Poente), which are regularly present in hydrothermal sites (Strazzulli et al., 2017). Specifically, *S.*

azorense Az-Fu1 was one of the most abundant organisms predicted. It was first isolated in January 2001 from terrestrial hot springs at Furnas, São Miguel Island, Azores, Portugal (Aguiar et al., 2004).

- The GSMM of the chemolithoautotrophic organism *S. azorense* Az-Fu1 was reconstructed to try to get insights into its metabolism, and genetic adaptation to extreme environments, and investigate its capability to produce compounds with industrial interest. In fact, genome annotation and metabolic analysis revealed important carbon and sulfur metabolism routes. In specific, the CO_2 fixation route appears to be through the rTCA (Hügler & Sievert, 2010), there are incomplete Pentose Phosphate and Embden-Meyerhof-Parnas pathways, and the presence of a truncated SOX system, as indicated in literature (Aguiar et al., 2004), was confirmed.
- Moreover, during the genome re-annotation of *S. azorense* Az-Fu1, the presence of the main subunits of the bacterial cellulose operon and their regulators were found, and model simulations showed the organism's metabolic capability to produce cellulose under nitrogen-limiting conditions. These results are in line with reports from literature (Lalonde et al., 2005) that have shown *S. azorense* Az-Fu1 produces exopolysaccharides under stress conditions. Optimization of cellulose production using evolutionary algorithms was tested; however, no robust enough solutions were returned. Also, experimental validation must be performed to confirm whether cellulose production is naturally viable.
- The lack of experimental data on quantitative matter limits the spectrum of application of *S. azorense* Az-Fu1 GSMM; however, given the fact that *S. azorense* Az-Fu1 was identified in natural microbial communities, the qualitative study of the organism's metabolic role through computational simulation and optimization within a microbial community is of huge importance.
- Different steady-state simulation methods with application to microbial communities were analyzed to try to understand their potential phenotypic behavior prediction performances. The analyzed methods showed, to some extent, to predict the phenotypic behavior that characterizes the nitrification process catalyzed by *N. europaea* and *N. vulgaris*. Each one of the simulation methods showed strengths and weaknesses and the success of either approach depends on the microbial community composition and complexity. Consequently, the use of more than one simulation method is recommended whenever possible as these showed to complement and validate each other.
- To complement existing algorithms for the study and manipulation of microbial communities, several tools were also developed and implemented in MEWpy, aiming to make available a suitable tool for the *in silico* design of microbial communities using GSMMs. This allows, depending on the individual GSMMs detail, to perform (i) minimal medium optimization, (ii) species metabolite interactions, (iii)

untargeted, and (iv) targeted reaction/gene/enzyme optimization for a given objective (maximize/minimize growth or production of target compounds). MEWpy already allowed the simulation of microbial communities, once it includes phenotype simulation methods for individual or community GSMMs from the COBRApy (Ebrahim et al., 2013) and REFRAMED libraries. It also offers a practical interface to strain optimization metaheuristics, such as EAs, including multi-objective methods that are suitable for application in GSMMs of microbial communities.

- A microbial community study of extreme environments from the hydrothermal sites in São Miguel, Azores, using GSMM reconstructions was performed. Three of the five hydrothermal samples analyzed in Chapter 2 were characterized in terms of predicted organisms and the GSMM for each of the present organisms was reconstructed and validated with available data in the literature. Overall, 9 organisms were selected to proceed with the GSMM reconstruction process. The reconstructed GSMMs can also be used independently for investigating the extraordinary microbial metabolism in extreme environments.
- Simulation methods with validated applications to microbial communities were used to predict the possible interactions between the organisms in the sampled environments. Predictions have elucidated two main metabolic roles within the specific communities: *P. aerophilum* str. IM2 acts as an amino acid donor and *T. adornatus* strain 1910b acts as a trigger for *S. azorense* Az-Fu1 cellulose production.
- Untargeted and targeted co-culture optimization was performed using the new features of MEWpy as a microbial community optimization framework to evaluate *S. azorense* Az-Fu1 capabilities of cellulose production.
- Untargeted *in silico* design of the co-culture *S. azorense* Az-Fu1 and *T. adornatus* strain 1910b showed that indeed *S. azorense* Az-Fu1 produces cellulose; however, that production is totally consumed by *T. adornatus* strain 1910b, raising doubts about the effectiveness of this co-culture to produce cellulose in high amounts.
- Targeted *in silico* design co-culture of *S. azorense* Az-Fu1 and *E. coli* K12 MG1655 showed that acetate produced by *E. coli* K12 MG1655 is consumed by *S. azorense* Az-Fu1, being a significant part of the cellulose production route.
- Although the optimization results here reported using MEWpy demonstrate the framework applicability in the design of microbial communities, there is still room for new implementations on MEWpy concerning microbial communities' optimization.

7.2 Future Work

The present thesis contributed to developing an optimized co-culture for the production of cellulose by the extremophile *S. azorense* Az-Fu1 present in samples of hydrothermal vents in São Miguel, Azores, using reconstruction and simulation methods and developing metabolic engineering tools using MEWpy. However, additional research is recommended either to improve and expand the obtained results or to address other questions:

- Although the GSMMs of the extremophilic organisms here presented have some experimental validation, manual curation, and further validation are still needed, essentially on carbon and sulfur metabolisms, as well as under different electron donors and acceptors, due to the adapted nutrient-limiting environments.
- Organism-specific metabolic routes should be investigated, aided by these GSMMs, to find, similarly to what was done for *S. azorense* Az-Fu1, individual genomic, metabolic, or enzymatic traits that can result in a possible production of compounds with industrial interest.
- MEWpy aims to be a reference tool for the metabolic engineering of communities, making available a diverse catalog of simulation and optimization heuristics and functions applicable in a microbial community context. However, there is still room for new implementations on MEWpy concerning microbial communities' optimization. Current efforts are focused on:
    (i) implementing optimization of the best community composition for a given objective, using top-down and bottom-up approaches,
    (ii) implementing additional constraints to improve SteadyCom simulation predictions in cases of the presence of fast-growing organisms. These constraints can be made by trying to restrict organism-specific exchange reactions lower and/or upper bounds, limiting the organism's growth in the same range as others present in the community.
- Although the simulation and optimization methods applied showed possible routes for the production of cellulose by *S. azorense* Az-Fu1 in nitrogen limitation and in a co-culture with *E. coli* K12 MG1655, experimental validation is still needed. Limitations will be found once *E. coli* K12 MG1655 and *S. azorense* Az-Fu1 have distinct optimal growth temperature ranges. Possible alternatives can range from the adaptation of an *E. coli* strain to high temperatures or finding a thermophilic organism with the potential of *E. coli* K12 MG1655 to trigger *S. azorense* Az-Fu1 to produce cellulose.

- In this work, simulation and optimization methods showed a possible path for the optimization of a co-culture product. Questions are raised if these methods using GSMMs are in fact scalable for a number of organisms in the range of dozens, not due to the implementation of the methods *per se* but essentially by the actual possibility of experimental validation.

References

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., ... Huttenhower, C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, *8*(6), e1002358. https://doi.org/10.1371/journal.pcbi.1002358

Ackert, L. (2013). *Sergei Vinogradskii and the Cycle of Life* (Vol. 34). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-5198-9

Aguiar, P., Beveridge, T. J., & Reysenbach, A. L. (2004). *Sulfurihydrogenibium azorense*, sp. nov., a thermophilic hydrogen-oxidizing microaerophile from terrestrial hot springs in the Azores. *International Journal of Systematic and Evolutionary Microbiology*, *54*(1), 33–39. https://doi.org/10.1099/ijs.0.02790-0

Albuquerque, L., Rainey, F. A., Fernanda Nobre, M., & da Costa, M. S. (2012). *Hydrotalea sandarakina* sp. nov., isolated from a hot spring runoff, and emended descriptions of the genus Hydrotalea and the species *Hydrotalea flava*. *International Journal of Systematic and Evolutionary Microbiology*, *62*(Pt 7), 1603–1608. https://doi.org/10.1099/IJS.0.034496-0

Albuquerque, L., Rainey, F. A., Nobre, M. F., & da Costa, M. S. (2010). *Meiothermus granaticius* sp. nov., a new slightly thermophilic red-pigmented species from the Azores. *Systematic and Applied Microbiology*, *33*(5), 243–246. https://doi.org/10.1016/J.SYAPM.2010.04.001

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods 2014 11:11*, *11*(11), 1144–1146. https://doi.org/10.1038/nmeth.3103

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ando, N., Barquera, B., Bartlett, D. H., Boyd, E., Burnim, A. A., Byer, A. S., ... Watkins, M. B. (2021). The Molecular Basis for Life in Extreme Environments. *Annual Review of Biophysics*, *50*(1), 343–372. https://doi.org/10.1146/annurev-biophys-100120-072804

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data – ScienceOpen. Retrieved February 18, 2022, from https://www.scienceopen.com/document?vid=de674375-

ab83-4595-afa9-4c8aa9e4e736

Antonakoudis, A., Barbosa, R., Kotidis, P., & Kontoravdi, C. (2020). The era of big data: Genome-scale modelling meets machine learning. *Computational and Structural Biotechnology Journal*, *18*, 3287–3300. https://doi.org/10.1016/J.CSBJ.2020.10.011

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., ... Yu, D. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology 2018 36:7*, *36*(7), 566–569. https://doi.org/10.1038/nbt.4163

Atalah, J., Cáceres-Moreno, P., Espina, G., & Blamey, J. M. (2019). Thermophiles and the applications of their enzymes as new biocatalysts. *Bioresource Technology*, *280*, 478–488. https://doi.org/10.1016/J.BIORTECH.2019.02.008

Ayling, M., Clark, M. D., & Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, *21*(2), 584–594. https://doi.org/10.1093/BIB/BBZ020

Bagchi, S., Biswas, R., & Nandy, T. (2012). Autotrophic Ammonia Removal Processes: Ecology to Technology. *Critical Reviews in Environmental Science and Technology*, *42*(13), 1353–1418. https://doi.org/10.1080/10643389.2011.556885

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bar-Even, A., Noor, E., Lewis, N. E., & Milo, R. (2010). Design and analysis of synthetic carbon fixation pathways. *Proceedings of the National Academy of Sciences*, *107*(19), 8889–8894. https://doi.org/10.1073/pnas.0907176107

Barnard, D., Casanueva, A., Tuffin, M., & Cowan, D. (2010). Extremophiles in biofuel synthesis. *Environmental Technology*, *31*(8–9), 871–888. https://doi.org/10.1080/09593331003710236

Barrett, A. J. (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *European Journal of Biochemistry*, *250*(1), 1–6. https://doi.org/10.1111/J.1432-1033.1997.001_1.X

Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., & Kaleta, C. (2017). BacArena: Individual-based

metabolic modeling of heterogeneous microbes in complex communities. *PLOS Computational Biology*, *13*(5), e1005544. https://doi.org/10.1371/journal.pcbi.1005544

Bauermeister, A., Moeller, R., Reitz, G., Sommer, S., & Rettberg, P. (2011). Effect of relative humidity on *Deinococcus radiodurans*' resistance to prolonged desiccation, heat, ionizing, germicidal, and environmentally relevant UV radiation. *Microbial Ecology*, *61*(3), 715–722. https://doi.org/10.1007/S00248-010-9785-4

Benítez-Hidalgo, A., Nebro, A. J., García-Nieto, J., Oregi, I., & Del Ser, J. (2019). jMetalPy: A Python framework for multi-objective optimization with metaheuristics. *Swarm and Evolutionary Computation*, *51*, 100598. https://doi.org/10.1016/J.SWEVO.2019.100598

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, *22*(1), 178–193. https://doi.org/10.1093/BIB/BBZ155

Bhatnagar, S., Badger, J. H., Madupu, R., Khouri, H. M., O'Connor, E. M., Robb, F. T., ... Eisen, J. A. (2015). Genome sequence of the sulfate-reducing thermophilic bacterium *Thermodesulfovibrio yellowstonii strain* DSM 11347T (phylum Nitrospirae). *Genome Announcements*, *3*(1), 1994–1995. https://doi.org/10.1128/genomeA.01489-14

Bizukojc, M., Dietz, D., Sun, J., & Zeng, A. P. (2010). Metabolic modelling of syntrophic-like growth of a 1,3-propanediol producer, *Clostridium butyricum*, and a methanogenic archeon, *Methanosarcina mazei*, under anaerobic conditions. *Bioprocess and Biosystems Engineering*, *33*(4), 507–523. https://doi.org/10.1007/s00449-009-0359-0

Blanco Parte, F. G., Santoso, S. P., Chou, C. C., Verma, V., Wang, H. T., Ismadji, S., & Cheng, K. C. (2020). Current progress on the production, modification, and applications of bacterial cellulose. *Critical Reviews in Biotechnology*, *40*(3), 397–414. https://doi.org/10.1080/07388551.2020.1713721

Blevins, S. M., & Bronze, M. S. (2010). Robert Koch and the 'golden age' of bacteriology. *International Journal of Infectious Diseases*, *14*(9), e744–e751. https://doi.org/10.1016/J.IJID.2009.12.003

Bokulich, N. A., Lewis, Z. T., Boundy-Mills, K., & Mills, D. A. (2016). A new perspective on microbial landscapes within food production. *Current Opinion in Biotechnology*, *37*, 182–189. https://doi.org/10.1016/J.COPBIO.2015.12.008

Bonjour, F., & Aragno, M. (1986). Growth of thermophilic, obligatorily chemolithoautotrophic hydrogen-oxidizing bacteria related to *Hydrogenobacter* with thiosulfate and elemental sulfur as electron and energy source. *FEMS Microbiology Letters*, *35*(1), 11–15. https://doi.org/10.1111/J.1574-6968.1986.TB01490.X

Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., & Jamshidi, N. (2010). Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Molecular Systems Biology*, *6*(1), 422. https://doi.org/10.1038/msb.2010.68

Bordbar, A., Monk, J. M., King, Z. A., & Palsson, B. O. (2014). Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics 2014 15:2*, *15*(2), 107–120. https://doi.org/10.1038/nrg3643

Borer, B., Ataman, M., Hatzimanikatis, V., & Or, D. (2019). Modeling metabolic networks of individual bacterial agents in heterogeneous and dynamic soil habitats (IndiMeSH). *PLOS Computational Biology*, *15*(6), e1007127. https://doi.org/10.1371/JOURNAL.PCBI.1007127

Bosi, E., Bacci, G., Mengoni, A., & Fondi, M. (2017). Perspectives and challenges in microbial communities metabolic modeling. *Front. Genet.*, *8*, 88. https://doi.org/10.3389/fgene.2017.00088

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, *20*(4), 1125–1136. https://doi.org/10.1093/BIB/BBX120

Brenner, K., You, L., & Arnold, F. H. (2008). Engineering microbial consortia: a new frontier in synthetic biology. *Trends in Biotechnology*, *26*(9), 483–489. https://doi.org/10.1016/J.TIBTECH.2008.05.004

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods 2021 18:4*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Budinich, M., Bourdon, J., Larhlimi, A., & Eveillard, D. (2017). A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE*, *12*(2). https://doi.org/10.1371/journal.pone.0171744

Burgard, A. P., Pharkya, P., & Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and Bioengineering*, *84*(6), 647–657. https://doi.org/10.1002/BIT.10803

Buszewski, B., Rogowska, A., Pomastowski, P., Złoch, M., & Railean-Plugaru, V. (2017). Identification of Microorganisms by Modern Analytical Techniques. *Journal of AOAC INTERNATIONAL*, *100*(6), 1607–1623. https://doi.org/10.5740/JAOACINT.17-0207

Cacicedo, M. L., Castro, M. C., Servetas, I., Bosnea, L., Boura, K., Tsafrakidou, P., … Castro, G. R. (2016). Progress in bacterial cellulose matrices for biotechnological applications. *Bioresource Technology*, *213*, 172–180. https://doi.org/10.1016/j.biortech.2016.02.071

Campodonico, M. A., Vaisman, D., Castro, J. F., Razmilic, V., Mercado, F., Andrews, B. A., … Asenjo, J. A. (2016). *Acidithiobacillus ferrooxidans* comprehensive model driven analysis of the electron transfer metabolism and synthetic strain design for biomining applications. *Metabolic Engineering Communications*, *3*, 84–96. https://doi.org/10.1016/J.METENO.2016.03.003

Capela, J., Lagoa, D., Rodrigues, R., Cunha, E., Cruz, F., Barbosa, A., … Dias, O. (2021). merlin v4.0: an updated platform for the reconstruction of high-quality genome-scale metabolic models. *BioRxiv*, 2021.02.24.432752. https://doi.org/10.1101/2021.02.24.432752

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., … Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal 2012 6:8*, *6*(8), 1621–1624. https://doi.org/10.1038/ismej.2012.8

Cardoso, J. G. R., Jensen, K., Lieven, C., Hansen, A. S. L., Galkina, S., Beber, M., … Sonnenschein, N. (2018). Cameo: A Python Library for Computer Aided Metabolic Engineering and Optimization of Cell Factories. *A C S Synthetic Biology*, *7*(4), 1163–1166. https://doi.org/10.1021/ACSSYNBIO.7B00423

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., … Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, *42*(D1), D459–D471. https://doi.org/10.1093/NAR/GKT1103

Chan, S H J, Simons, M. N., & Maranas, C. D. (2017). SteadyCom: Predicting microbial abundances while ensuring community stability. *PLoS Comput. Biol.*, *13*(5), e1005539.

https://doi.org/10.1371/journal.pcbi.1005539

Chan, Siu H.J., Friedman, E. S., Wu, G. D., & Maranas, C. D. (2019). Predicting the Longitudinally and Radially Varying Gut Microbiota Composition using Multi-Scale Microbial Metabolic Modeling. *Processes 2019, Vol. 7, Page 394*, *7*(7), 394. https://doi.org/10.3390/PR7070394

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., … Schomburg, D. (2021). BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research*, *49*(D1), D498–D508. https://doi.org/10.1093/NAR/GKAA1025

Chang, A. L., Tuckerman, J. R., Gonzalez, G., Mayer, R., Weinhouse, H., Volman, G., … Gilles-Gonzalez, M. A. (2001). Phosphodiesterase A1, a regulator of cellulose synthesis in *Acetobacter xylinum*, is a heme-based sensor. *Biochemistry*, *40*(12), 3420–3426. https://doi.org/10.1021/bi0100236

Chen, L., Ren, Y., Lin, J., Liu, X., Pang, X., & Lin, J. (2012). *Acidithiobacillus caldus* Sulfur Oxidation Model Based on Transcriptome Analysis between the Wild Type and Sulfur Oxygenase Reductase Defective Mutant. *PLoS ONE*, *7*(9), 1–13. https://doi.org/10.1371/journal.pone.0039470

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, X., Zhou, L., Tian, K., Kumar, A., Singh, S., Prior, B. A., & Wang, Z. (2013). Metabolic engineering of *Escherichia coli*: A sustainable industrial platform for bio-based chemical production. *Biotechnology Advances*, *31*(8), 1200–1223. https://doi.org/10.1016/J.BIOTECHADV.2013.02.009

Chng, K. R., Ghosh, T. S., Tan, Y. H., Nandi, T., Lee, I. R., Ng, A. H. Q., … Nagarajan, N. (2020). Metagenome-wide association analysis identifies microbial determinants of post-antibiotic ecological recovery in the gut. *Nature Ecology & Evolution 2020 4:9*, *4*(9), 1256–1267. https://doi.org/10.1038/s41559-020-1236-0

Chung, A. P., Rainey, F. A., Valente, M., Nobre, M. F., & Da Costa, M. S. (2000). *Thermus igniterrae* sp. nov. and *Thermus antranikianii* sp. nov., two new species from Iceland. *International Journal of Systematic and Evolutionary Microbiology*, *50*(1), 209–217. https://doi.org/10.1099/00207713-50-1-209

Cleaver, A. A., Burton, N. P., & Norris, P. R. (2007). A Novel *Acidimicrobium* Species in Continuous

Cultures of Moderately Thermophilic, Mineral-Sulfide-Oxidizing Acidophiles. *Applied and Environmental Microbiology*, *73*(13), 4294. https://doi.org/10.1128/AEM.02658-06

Colarusso, A. V., Goodchild-Michelman, I., Rayle, M., & Zomorrodi, A. R. (2021). Computational modeling of metabolism in microbial communities on a genome-scale. *Current Opinion in Systems Biology*, *26*, 46–57. https://doi.org/10.1016/J.COISB.2021.04.001

Consortium, T. U., Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., … Teodoro, D. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/NAR/GKAA1100

Counts, J. A., Zeldes, B. M., Lee, L. L., Straub, C. T., Adams, M. W. W., & Kelly, R. M. (2017). Physiological, metabolic and biotechnological features of extremely thermophilic microorganisms. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *9*(3). https://doi.org/10.1002/wsbm.1377

Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C., & Banfield, J. F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature*, *558*(7710), 440–444. https://doi.org/10.1038/S41586-018-0207-Y

Cruz, F., Capela, J., Ferreira, E. C., Rocha, M., & Dias, O. (2021). BioISO: an objective-oriented application for assisting the curation of genome-scale metabolic models. *BioRxiv*, 2021.03.07.434259. https://doi.org/10.1101/2021.03.07.434259

Cruz, R. (2018). Nitrobacter vulgaris*: genome-scale model reconstruction and interactions with* Nitrosomonas europaea. University of Minho.

Dahl, C., & Friedrich, C. G. (2008). *Microbial Sulfur Metabolism*. *Springer-Verlag Berlin Heidelberg New York*.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., … Wagner, M. (2015). Complete nitrification by *Nitrospira* bacteria. *Nature 2015 528:7583*, *528*(7583), 504–509. https://doi.org/10.1038/nature16461

Dalmasso, C., Oger, P., Selva, G., Courtine, D., L'Haridon, S., Garlaschelli, A., … Alain, K. (2016). *Thermococcus piezophilus* sp. nov., a novel hyperthermophilic and piezophilic archaeon with a broad pressure range for growth, isolated from a deepest hydrothermal vent at the Mid-Cayman

Rise. *Systematic and Applied Microbiology*, *39*(7), 440–444. https://doi.org/10.1016/J.SYAPM.2016.08.003

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/GIGASCIENCE/GIAB008

Darland, G., Brock, T. D., Samsonoff, W., & Conti, S. F. (1970). A Thermophilic, Acidophilic Mycoplasma Isolated from a Coal Refuse Pile. *Science*, *170*(3965), 1416–1418. https://doi.org/10.1126/SCIENCE.170.3965.1416

Dauner, M., & Sauer, U. (2001). Stoichiometric growth model for riboflavin-producing *Bacillus subtilis*. *Biotechnol Bioeng*, *76*(2), 132–143. https://doi.org/10.1002/bit.1153

David, L., Marashi, S.-A., Larhlimi, A., Mieth, B., & Bockmayr, A. (2011). FFCA: a feasibility-based method for flux coupling analysis of metabolic networks. *BMC Bioinformatics*, *12*(1), 236. https://doi.org/10.1186/1471-2105-12-236

Deng, Z., & Delwart, E. (2021). ContigExtender: a new approach to improving de novo sequence assembly for viral metagenomics data. *BMC Bioinformatics*, *22*(1), 1–19. https://doi.org/https://doi.org/10.1186/s12859-021-04038-2

Dias, O., Rocha, M., Ferreira, E. C., & Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Research*, *43*(8), 3899–3910. https://doi.org/10.1093/nar/gkv294

Dias, O., Rocha, M., Ferreira, E. C., & Rocha, I. (2018). Reconstructing High-Quality Large-Scale Metabolic Models with merlin. *Methods in Molecular Biology*, *1716*, 1–36. https://doi.org/10.1007/978-1-4939-7528-0_1

Diener, C., Gibbons, S. M., & Resendis-Antonio, O. (2020). MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota. *MSystems*, *5*(1). https://doi.org/10.1128/mSystems.00606-19

Donia, M. S., Cimermancic, P., Schulze, C. J., Wieland Brown, L. C., Martin, J., Mitreva, M., … Fischbach, M. A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell*, *158*(6), 1402–1414. https://doi.org/10.1016/J.CELL.2014.08.032

Durvasula, R., & Rao, D. V. S. (2018). Extremophiles: from Biology to Biotechnology. In *Extremophiles* (1st ed., pp. 1–18). Boca Raton : Taylor & Francis, a CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa plc, 2018.: CRC Press. https://doi.org/10.1201/9781315154695-1

Dworkin, M., & Gutnick, D. (2012). Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. *FEMS Microbiology Reviews*, *36*(2), 364–379. https://doi.org/10.1111/J.1574-6976.2011.00299.X

Ebrahim, A., Lerman, J. A., Palsson, B. O., & Hyduke, D. R. (2013). COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology*, *7*, 74. https://doi.org/10.1186/1752-0509-7-74

Edwards, J. S., & Palsson, B. O. (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *Journal of Biological Chemistry*, *274*(25), 17410–17416. https://doi.org/10.1074/jbc.274.25.17410

El-Semman, I. E., Karlsson, F. H., Shoaie, S., Nookaew, I., Soliman, T. H., & Nielsen, J. (2014). Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction. *BMC Systems Biology*, *8*(1), 41. https://doi.org/10.1186/1752-0509-8-41

Eng, A., & Borenstein, E. (2019). Microbial Community Design:Methods, Applications, and Opportunities. *Current Opinion in Biotechnology*, *58*, 117. https://doi.org/10.1016/J.COPBIO.2019.03.002

Escobar-Zepeda, A., De León, A. V. P., & Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, *6*(DEC), 348. https://doi.org/10.3389/FGENE.2015.00348/BIBTEX

Fang, X., Lloyd, C. J., & Palsson, B. O. (2020). Reconstructing organisms *in silico*: genome-scale models and their emerging applications. *Nature Reviews Microbiology 2020 18:12*, *18*(12), 731–743. https://doi.org/10.1038/s41579-020-00440-4

Faria, J. P., Khazaei, T., Edirisinghe, J. N., Weisenhorn, P., Seaver, S. M. D., Conrad, N., … Henry, C. S. (2017). Constructing and analyzing metabolic flux models of microbial communities. In T. J. McGenity, K. N. Timmis, & B. Nogales (Eds.), *Hydrocarbon and Lipid Microbiology Protocols: Genetic, Genomic and System Analyses of Communities* (pp. 247–273). Berlin, Heidelberg:

Springer Berlin Heidelberg. https://doi.org/10.1007/8623_2016_215

Faust, K. (2018). Microbial consortium design benefits from metabolic modeling. *Trends Biotechnol.* https://doi.org/10.1016/j.tibtech.2018.11.004

Faust, Karoline, & Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology 2012 10:8*, *10*(8), 538–550. https://doi.org/10.1038/nrmicro2832

Faust, Karoline, Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology*, *8*(7). https://doi.org/10.1371/journal.pcbi.1002606

Feinberg, L. F., Srikanth, R., Vachet, R. W., & Holden, J. F. (2008). Constraints on anaerobic respiration in the hyperthermophilic archaea *Pyrobaculum islandicum* and *Pyrobaculum aerophilum*. *Applied and Environmental Microbiology*, *74*(2), 396–402. https://doi.org/10.1128/AEM.02033-07

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., ... Palsson, B. Ø. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, *3*(121), 121. https://doi.org/10.1038/msb4100155

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. (2008). Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiology 2009 7:2*, *7*(2), 129–143. https://doi.org/10.1038/nrmicro1949

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L., & Palsson, B. (2009). Reconstruction of Biochemical Networks in Microbial Organisms. *Nature Reviews. Microbiology*, *7*(2), 129. https://doi.org/10.1038/NRMICRO1949

Feist, A. M., & Palsson, B. O. (2010). The biomass objective function. *Current Opinion in Microbiology*, *13*(3), 344–349. https://doi.org/10.1016/j.mib.2010.03.003

Feist, A. M., Scholten, J. C. M., Palsson, B. Ø., Brockman, F. J., & Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Molecular Systems Biology*, *2*(1), 2006.0004. https://doi.org/10.1038/MSB4100046

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology 2017 15:10*, *15*(10), 579–590. https://doi.org/10.1038/nrmicro.2017.87

Fowler, D., Coyle, M., Skiba, U., Sutton, M. A., Cape, J. N., Reis, S., ... Voss, M. (2013). The global nitrogen cycle in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1621). https://doi.org/10.1098/RSTB.2013.0164

França, L., Rainey, F. A., Nobre, M. F., & da Costa, M. S. (2006). *Tepidicella xavieri* gen. nov., sp. nov., a betaproteobacterium isolated from a hot spring runoff. *International Journal of Systematic and Evolutionary Microbiology*, *56*(Pt 4), 907–912. https://doi.org/10.1099/IJS.0.64193-0

Freeman, M. F., Helf, M. J., Bhushan, A., Morinaka, B. I., & Piel, J. (2017). Seven enzymes create extraordinary molecular complexity in an uncultivated bacterium. *Nature Chemistry*, *9*(4), 387–395. https://doi.org/10.1038/NCHEM.2666

Friedrich, C. G., Bardischewsky, F., Rother, D., Quentmeier, A., & Fischer, J. (2005). Prokaryotic sulfur oxidation. *Current Opinion in Microbiology*, *8*(3), 253–259. https://doi.org/10.1016/j.mib.2005.04.005

Friedrich, C. G., Rother, D., Bardischewsky, F., Ouentmeier, A., & Fischer, J. (2001). Oxidation of Reduced Inorganic Sulfur Compounds by Bacteria: Emergence of a Common Mechanism? *Applied and Environmental Microbiology*, *67*(7), 2873. https://doi.org/10.1128/AEM.67.7.2873-2882.2001

Frioux, C., Singh, D., Korcsmaros, T., & Hildebrand, F. (2020). From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, *18*, 1722–1734. https://doi.org/10.1016/J.CSBJ.2020.06.028

García-Jiménez, B., García, J. L., & Nogales, J. (2018). FLYCOP: metabolic modeling-based analysis and engineering microbial communities. *Bioinformatics*, *34*(17), i954. https://doi.org/10.1093/BIOINFORMATICS/BTY561

García-Jiménez, B., Torres-Bacete, J., & Nogales, J. (2021). Metabolic modelling approaches for describing and engineering microbial communities. *Computational and Structural Biotechnology Journal*, *19*, 226–246. https://doi.org/10.1016/J.CSBJ.2020.12.003

Ghosh, W., & Dam, B. (2009). Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiology Reviews*, *33*(6), 999–1043. https://doi.org/10.1111/j.1574-6976.2009.00187.x

Gianchandani, E. P., Chavali, A. K., & Papin, J. A. (2010). The application of flux balance analysis in systems biology. *Biology and Medicine*, *2*, 372–382. https://doi.org/10.1002/wsbm.60

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine 2018 24:4*, *24*(4), 392–400. https://doi.org/10.1038/nm.4517

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC Biology 2014 12:1*, *12*(1), 1–4. https://doi.org/10.1186/S12915-014-0069-1

Glowacki, R. W. P., & Martens, E. C. (2020). In sickness and health: Effects of gut microbial metabolites on human physiology. *PLoS Pathogens*, *16*(4). https://doi.org/10.1371/JOURNAL.PPAT.1008370

Gomez, J. A., Höffner, K., & Barton, P. I. (2014). DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinformatics 2014 15:1*, *15*(1), 1–10. https://doi.org/10.1186/S12859-014-0409-8

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Goyal, N., Widiastuti, H., Karimi, I. A., & Zhou, Z. (2014). A genome-scale metabolic model of *Methanococcus maripaludis* S2 for CO2 capture and conversion to methane. *Mol. BioSyst.*, *10*(5), 1043–1054. https://doi.org/10.1039/C3MB70421A

Graham, E. D., Heidelberg, J. F., & Tully, B. J. (2017). Binsanity: Unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ*, *2017*(3), e3035. https://doi.org/10.7717/PEERJ.3035

Grunditz, C., & Dalhammar, G. (2001). Development of nitrification inhibition assays using pure cultures of *Nitrosomonas* and *Nitrobacter*. *Water Research*, *35*(2), 433–440. https://doi.org/10.1016/S0043-1354(00)00312-2

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., & Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biology 2019 20:1*, *20*(1), 1–18. https://doi.org/10.1186/S13059-019-1730-3

Gudmundsson, S., & Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC*

Bioinformatics, *11*(1), 1–3. https://doi.org/10.1186/1471-2105-11-489/TABLES/2

Gupta, R. S., & Lali, R. (2013). Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order Aquificales, containing the families *Aquificaceae* and *Hydrogenothermaceae*, and a new order <i>Desulf. *Antonie van Leeuwenhoek*, *104*(3), 349–368. https://doi.org/10.1007/s10482-013-9957-6

Hallberg, K. B., Dopson, M., & Lindström, E. B. (1996). Reduced sulfur compound oxidation by *Thiobacillus caldus*. *Journal of Bacteriology*, *178*(1), 6–11. https://doi.org/10.1128/jb.178.1.6-11.1996

Hallberg, K. B., & Lindstromt, E. B. (1994). a Moderately Thermophilic Acidophile. *Microbiology*, *140*(1994), 3451–3456.

Hamamura, N., Meneghin, J., & Reysenbach, A. L. (2013). Comparative community gene expression analysis of Aquificales-dominated geothermal springs. *Environmental Microbiology*, *15*(4), 1226–1237. https://doi.org/10.1111/1462-2920.12061

Hamarneh, S. (1960). Measuring the Invisible World. The life and works of Antoni van Leeuwenhoek. A. Schierbeek. Abelard-Schuman, New York, 1959. 223 pp. $4. *Science*, *132*(3422), 289–290. https://doi.org/10.1126/SCIENCE.132.3422.289

Han, H., Ling, Z., Khan, A., Virk, A. K., Kulshrestha, S., & Li, X. (2019). Improvements of thermophilic enzymes: From genetic modifications to applications. *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2019.01.087

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, *5*(10). https://doi.org/10.1016/S1074-5521(98)90108-9

Hanly, T. J., & Henson, M. A. (2011). Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnol. Bioeng.*, *108*(2), 376–385. https://doi.org/10.1002/bit.22954

Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., … Segrè, D. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, *7*(4), 1104–1115. https://doi.org/10.1016/j.celrep.2014.03.070

Hedlund, B. P., Reysenbach, A. L., Huang, L., Ong, J. C., Liu, Z., Dodsworth, J. A., ... Dong, H. (2015). Isolation of diverse members of the Aquificales from geothermal springs in Tengchong, China. *Frontiers in Microbiology*, *6*(FEB), 1–8. https://doi.org/10.3389/fmicb.2015.00157

Heinken, A, Sahoo, S., Fleming, R. M., & Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*, *4*(1), 28–40. https://doi.org/10.4161/gmic.22370

Heinken, Almut, Acharya, G., Ravcheev, D. A., Hertel, J., Nyga, M., Okpala, O. E., ... Thiele, I. (2020). AGORA2: Large scale reconstruction of the microbiome highlights wide-spread drug-metabolising capacities. *BioRxiv*, 2020.11.09.375451. https://doi.org/10.1101/2020.11.09.375451

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., ... Fleming, R. M. T. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nature Protocols 2019 14:3*, *14*(3), 639–702. https://doi.org/10.1038/s41596-018-0098-2

Henry, C. S., Bernstein, H. C., Weisenhorn, P., Taylor, R. C., Lee, J. Y., Zucker, J., & Song, H. S. (2016). Microbial community metabolic modeling: A community data-driven network reconstruction. *J Cell Physiol*, *231*(11), 2339–2345. https://doi.org/10.1002/jcp.25428

Henry F. A., Devereux R., Maki J. S., Gilmour C. C., Woese C. R., Mandeleo L., Schauder R., Remsen C. C., M. R. (1994). Characterization of a new thermophilic sulfate-reducing bacterium *Thermodesulfovibrio yellowstonii*, ge. nov. and sp. nov.: its phylogenetic relationship to Thermodesulfobacterium commune and their origins deep within the bacterial domain. *Archives of Microbiology*, *161*, 62–69.

Hernández-Arriaga, A. M., del Cerro, C., Urbina, L., Eceiza, A., Corcuera, M. A., Retegi, A., & Auxiliadora Prieto, M. (2019). Genome sequence and characterization of the bcs clusters for the production of nanocellulose from the low pH resistant strain *Komagataeibacter medellinensis* ID13488. *Microbial Biotechnology*, *12*(4), 620–632. https://doi.org/10.1111/1751-7915.13376

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5). https://doi.org/10.1038/NMICROBIOL.2016.48

Hügler, M., Huber, H., Molyneaux, S. J., Vetriani, C., & Sievert, S. M. (2007). Autotrophic CO2 fixation via the reductive tricarboxylic acid cycle in different lineages within the phylum Aquificae: Evidence

for two ways of citrate cleavage. *Environmental Microbiology*, *9*(1), 81–92. https://doi.org/10.1111/j.1462-2920.2006.01118.x

Hügler, M., & Sievert, S. M. (2010). Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Http://Dx.Doi.Org/10.1146/Annurev-Marine-120709-142712*, *3*, 261–289. https://doi.org/10.1146/ANNUREV-MARINE-120709-142712

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential Model-Based Optimization for General Algorithm Configuration. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6683 LNCS*, 507–523. https://doi.org/10.1007/978-3-642-25566-3_40

Ilgrande, C., Leroy, B., Wattiez, R., Vlaeminck, S. E., Boon, N., & Clauwaert, P. (2018). Metabolic and proteomic responses to salinity in synthetic nitrifying communities of *Nitrosomonas* spp. And *Nitrobacter* spp. *Frontiers in Microbiology*, *9*(NOV), 2914. https://doi.org/10.3389/FMICB.2018.02914/BIBTEX

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., & Rusch, D. B. (2013). The YNP metagenome project: Environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. *Frontiers in Microbiology*, *4*(MAY), 67. https://doi.org/10.3389/FMICB.2013.00067

Isalan, M. (2012). A cell in a computer. *Nature 2012 488:7409*, *488*(7409), 40–41. https://doi.org/10.1038/488040a

Ishino, S., & Ishino, Y. (2014). DNA polymerases as useful reagents for biotechnology - The history of developmental research in the field. *Frontiers in Microbiology*, *5*(AUG), 465. https://doi.org/10.3389/FMICB.2014.00465/BIBTEX

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., … Rice, P. (2013). EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, *29*(10), 1325–1332. https://doi.org/10.1093/BIOINFORMATICS/BTT113

Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., & Banfield, J. F. (2020). The rise of diversity in metabolic platforms across the Candidate Phyla Radiation. *BMC Biology*, *18*(1). https://doi.org/10.1186/S12915-020-00804-5

Jansma, J., & El Aidy, S. (2021). Understanding the host-microbe interactions using metabolic modeling. *Microbiome*, *9*(1), 1–14. https://doi.org/10.1186/S40168-020-00955-1

Jay, Z. J., Beam, J. P., Kozubal, M. A., Jennings, R. de M., Rusch, D. B., & Inskeep, W. P. (2016). The distribution, diversity and function of predominant Thermoproteales in high-temperature environments of Yellowstone National Park. *Environmental Microbiology*, *18*(12), 4755–4769. https://doi.org/10.1111/1462-2920.13366

Jiang, Y., Wang, D., Wang, W., & Xu, D. (2021). Computational methods for protein localization prediction. *Computational and Structural Biotechnology Journal*, *19*, 5834–5844. https://doi.org/10.1016/J.CSBJ.2021.10.023

Johns, N. I., Blazejewski, T., Gomes, A. L., & Wang, H. H. (2016). Principles for designing synthetic microbial communities. *Curr. Opin. Microbiol.*, *31*, 146–153. https://doi.org/10.1016/j.mib.2016.03.010

Jones, B. E., Grant, W. D., Duckworth, A. W., Owenson, G. G., Horikoshi, K., Jones, B. E., … Owenson, · G G. (1998). Microbial diversity of soda lakes. *Extremophiles 1998 2:3*, *2*(3), 191–200. https://doi.org/10.1007/S007920050060

Jullesson, D., David, F., Pfleger, B., & Nielsen, J. (2015). Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances*, *33*(7), 1395–1402. https://doi.org/10.1016/J.BIOTECHADV.2015.02.011

Kambourova, M., Radchenkova, N., Tomova, I., & Bojadjieva, I. (2016). Thermophiles as a Promising Source of Exopolysaccharides with Interesting Properties. *Grand Challenges in Biology and Biotechnology*, *1*, 117–139. https://doi.org/10.1007/978-3-319-13521-2_4

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/NAR/GKV1070

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., … Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, *20*(4), 1085–1093. https://doi.org/10.1093/BIB/BBX085

Kashefi, K., & Lovley, D. R. (2003). Extending the Upper Temperature Limit for Life. *Science*, *301*(5635),

934–934. https://doi.org/10.1126/science.1086823

Kawano, Y., Saotome, T., Ochiai, Y., Katayama, M., Narikawa, R., & Ikeuchi, M. (2011). Cellulose accumulation and a cellulose synthase gene are responsible for cell aggregation in the cyanobacterium *Thermosynechococcus vulcanus* RKN. *Plant and Cell Physiology*, *52*(6), 957–966. https://doi.org/10.1093/pcp/pcr047

Kennedy, N. A., Walker, A. W., Berry, S. H., Duncan, S. H., Farquarson, F. M., Louis, P., ... Hold, G. L. (2014). The Impact of Different DNA Extraction Kits and Laboratories upon the Assessment of Human Gut Microbiota Composition by 16S rRNA Gene Sequencing. *PLOS ONE*, *9*(2), e88982. https://doi.org/10.1371/JOURNAL.PONE.0088982

Khandelwal, R A, Olivier, B. G., Roling, W. F., Teusink, B., & Bruggeman, F. J. (2013). Community flux balance analysis for microbial consortia at balanced growth. *PLoS One*, *8*(5), e64567. https://doi.org/10.1371/journal.pone.0064567

Khandelwal, Ruchir A., Olivier, B. G., Röling, W. F. M., Teusink, B., & Bruggeman, F. J. (2013). Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064567

Kim, W. J., Kim, H. U., & Lee, S. Y. (2017). Current state and applications of microbial genome-scale metabolic models. *Current Opinion in Systems Biology*, *2*, 10–18. https://doi.org/10.1016/J.COISB.2017.03.001

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., ... Lewis, N. E. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, *44*(D1), D515–D522. https://doi.org/10.1093/nar/gkv1049

Kitadai, N., Kameya, M., & Fujishima, K. (2017). Origin of the Reductive Tricarboxylic Acid (RTCA) Cycle-Type CO2 Fixation: A Perspective. *Life*, *7*(4). https://doi.org/10.3390/life7040039

Kitano, H. (2002). Systems Biology: A Brief Overview. *Science*, *295*(5560), 1662–1664. https://doi.org/10.1126/SCIENCE.1069492

Klamt, S., Saez-Rodriguez, J., & Gilles, E. D. (2007). Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Systems Biology*, *1*(1), 2. https://doi.org/10.1186/1752-0509-1-2

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glöckner, F. O. (2013).

Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, *41*(1), e1–e1. https://doi.org/10.1093/NAR/GKS808

Klitgord, N., & Segrè, D. (2010). Environments that Induce Synthetic Microbial Ecosystems. *PLOS Computational Biology*, *6*(11), e1001002. https://doi.org/10.1371/JOURNAL.PCBI.1001002

Koch, S, Kohrs, F., Lahmann, P., Bissinger, T., Wendschuh, S., Benndorf, D., … Klamt, S. (2019). RedCom: A strategy for reduced metabolic modeling of complex microbial communities and its application for analyzing experimental datasets from anaerobic digestion. *PLoS Comput. Biol.*, *15*(2), e1006759. https://doi.org/10.1371/journal.pcbi.1006759

Koch, Sabine, Benndorf, D., Fronk, K., Reichl, U., & Klamt, S. (2016). Predicting compositions of microbial communities from stoichiometric models with applications for the biogas process. *Biotechnology for Biofuels*, *9*(1), 17. https://doi.org/10.1186/s13068-016-0429-x

Krasteva, P. V., Bernal-Bayard, J., Travier, L., Martin, F. A., Kaminski, P. A., Karimova, G., … Ghigo, J. M. (2017). Insights into the structure and assembly of a bacterial cellulose secretion system. *Nature Communications*, *8*(1), 25–28. https://doi.org/10.1038/s41467-017-01523-2

Kristjánsson, J. K., Hjörleifsdóttir, S., Marteinsson, V. T., & Alfredsson, G. A. (1994). *Thermus scotoductus*, sp. nov., a Pigment-Producing Thermophilic Bacterium from Hot Tap Water in Iceland and Including *Thermus* sp. X-1. *Systematic and Applied Microbiology*, *17*(1), 44–50. https://doi.org/10.1016/S0723-2020(11)80030-5

Kumar, S., Dangi, A. K., Shukla, P., Baishya, D., & Khare, S. K. (2019). Thermozymes: Adaptive strategies and tools for their biotechnological applications. *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2019.01.088

Lagoa, D., Faria, J. P., Liu, F., Cunha, E., Henry, C. S., & Dias, O. (2021). TranSyT, the Transport Systems Tracker. *BioRxiv*, 2021.04.29.441738. https://doi.org/10.1101/2021.04.29.441738

Lalonde, S. V., Konhauser, K. O., Reysenbach, A.-L., & Ferris, F. G. (2005). The experimental silicification of Aquificales and their role in hot spring sinter formation. *Geobiology*, *3*(1), 41–52. https://doi.org/10.1111/J.1472-4669.2005.00042.X

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*,

9(4), 357. https://doi.org/10.1038/NMETH.1923

Lapierre, P., Shial, R., & Gogarten, J. P. (2006). Distribution of F- and A/V-type ATPases in *Thermus scotoductus* and other closely related species. *Systematic and Applied Microbiology*, *29*(1), 15–23. https://doi.org/10.1016/j.syapm.2005.06.004

Lawson, C. E., Harcombe, W. R., Hatzenpichler, R., Lindemann, S. R., Löffler, F. E., O'Malley, M. A., … McMahon, K. D. (2019). Common principles and best practices for engineering microbiomes. *Nature Reviews Microbiology 2019 17:12*, *17*(12), 725–741. https://doi.org/10.1038/s41579-019-0255-9

Lawson, C. E., Mundinger, A. B., Koch, H., Jacobson, T. B., Weathersby, C. A., Jetten, M. S. M., … Lücker, S. (2021). Investigating the chemolithoautotrophic and formate metabolism of *Nitrospira moscoviensis* by constraint-based metabolic modeling and 13C-tracer analysis. *BioRxiv*, 2021.02.18.431926. https://doi.org/10.1101/2021.02.18.431926

Lee, Y. S., Lee, D. H., Kahng, H. Y., Sohn, S. H., & Jung, J. S. (2011). *Polaribacter gangjinensis* sp. nov., isolated from seawater. *International Journal of Systematic and Evolutionary Microbiology*, *61*(6), 1425–1429. https://doi.org/10.1099/IJS.0.024869-0/CITE/REFWORKS

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., … Palsson, B. Ø. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, *6*, 390. https://doi.org/10.1038/MSB.2010.47

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. https://doi.org/10.1093/BIOINFORMATICS/BTV033

Li, F., Xie, W., Yuan, Q., Luo, H., Li, P., Chen, T., … Ma, H. (2018). Genome-scale metabolic model analysis indicates low energy production efficiency in marine ammonia-oxidizing archaea. *AMB Express*, *8*(1), 106. https://doi.org/10.1186/S13568-018-0635-Y

Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., … Zhang, C. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nature Biotechnology 2020 38:3*, *38*(3), 272–276. https://doi.org/10.1038/s41587-020-0446-y

Lindemann, S. R., Bernstein, H. C., Song, H. S., Fredrickson, J. K., Fields, M. W., Shou, W., … Beliaev,

A. S. (2016). Engineering microbial consortia for controllable outputs. *ISME J.*, *10*(9), 2077–2084. https://doi.org/10.1038/ismej.2016.26

Liu, Z., Wang, K., Chen, Y., Tan, T., & Nielsen, J. (2020). Third-generation biorefineries as the means to produce fuels and chemicals from CO2. *Nature Catalysis 2020 3:3*, *3*(3), 274–288. https://doi.org/10.1038/s41929-019-0421-5

Löffler, F. E., & Edwards, E. A. (2006). Harnessing microbial activities for environmental cleanup. *Current Opinion in Biotechnology*, *17*(3), 274–284. https://doi.org/10.1016/J.COPBIO.2006.05.001

Löhne, A., & Weißing, B. (2017). The vector linear program solver Bensolve – notes on theoretical background. *European Journal of Operational Research*, *260*(3), 807–813. https://doi.org/10.1016/J.EJOR.2016.02.039

Lu, H., Li, F., Sánchez, B. J., Zhu, Z., Li, G., Domenzain, I., … Nielsen, J. (2019). A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nature Communications 2019 10:1*, *10*(1), 1–13. https://doi.org/10.1038/s41467-019-11581-3

Machado, D., Andrejev, S., Tramontano, M., & Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Research*, *46*(15), 7542–7553. https://doi.org/10.1093/NAR/GKY537

Machado, D., Maistrenko, O. M., Andrejev, S., Kim, Y., Bork, P., Patil, K. R., & Patil, K. R. (2021). Polarization of microbial communities between competitive and cooperative metabolism. *Nature Ecology & Evolution 2021 5:2*, *5*(2), 195–203. https://doi.org/10.1038/s41559-020-01353-4

Mahadevan, R., & Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, *5*(4), 264–276. https://doi.org/10.1016/J.YMBEN.2003.09.002

Mahadevan, Radhakrishnan, Edwards, J. S., & Doyle, F. J. (2002). Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical Journal*, *83*(3), 1331–1340. https://doi.org/10.1016/S0006-3495(02)73903-9

Maia, P., Rocha, M., & Rocha, I. (2016). *In Silico* Constraint-Based Strain Optimization Methods: the Quest for Optimal Cell Factories. *Microbiology and Molecular Biology Reviews*, *80*(1), 45–67.

https://doi.org/10.1128/MMBR.00014-15

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., ... Hermjakob, H. (2020). BioModels—15 years of sharing computational models in life science. *Nucleic Acids Research*, *48*(D1), D407–D415. https://doi.org/10.1093/NAR/GKZ1055

McCarty, P. L., Bae, J., & Kim, J. (2011). Domestic Wastewater Treatment as a Net Energy Producer–Can This be Achieved? *Environmental Science and Technology*, *45*(17), 7100–7106. https://doi.org/10.1021/ES2014264

Mellbye, B L, Giguere, A. T., Murthy, G. S., Bottomley, P. J., Sayavedra-Soto, L. A., & Chaplen, F. W. R. (2018). Genome-scale, constraint-based modeling of nitrogen oxide fluxes during coculture of *Nitrosomonas europaea* and *Nitrobacter winogradskyi*. *MSystems*, *3*(3). https://doi.org/10.1128/mSystems.00170-17

Mellbye, Brett L., Giguere, A. T., Murthy, G. S., Bottomley, P. J., Sayavedra-Soto, L. A., & Chaplen, F. W. R. (2018). Genome-Scale, Constraint-Based Modeling of Nitrogen Oxide Fluxes during Coculture of *Nitrosomonas europaea* and *Nitrobacter winogradskyi* . *MSystems*, *3*(3), 1–13. https://doi.org/10.1128/msystems.00170-17

Mendoza, S. N., Olivier, B. G., Molenaar, D., & Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biology*, *20*(1), 1–20. https://doi.org/10.1186/S13059-019-1769-1

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. https://doi.org/10.1093/BIOINFORMATICS/BTV697

Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H. J., Cuenca, M., ... Sunagawa, S. (2019). Microbial abundance, activity and population genomic profiling with mOTUs2. *Nature Communications 2019 10:1*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-08844-4

Miller, S. R., Strong, A. L., Jones, K. L., & Ungerer, M. C. (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. *Applied and Environmental Microbiology*, *75*(13), 4565–4572. https://doi.org/https://doi.org/10.1128/AEM.02792-08

Mo, M. L., Palsson, B. Ø., & Herrgård, M. J. (2009). Connecting extracellular metabolomic measurements

to intracellular flux states in yeast. *BMC Systems Biology*, *3*(1), 37. https://doi.org/10.1186/1752-0509-3-37

Montalvo-Rodríguez, R., López-Garriga, J., Vreeland, R. H., Oren, A., Ventosa, A., & Kamekura, M. (2000). *Haloterrigena thermotolerans* sp. nov., a halophilic archaeon from Puerto Rico. *International Journal of Systematic and Evolutionary Microbiology*, *50*(3), 1065–1071. https://doi.org/10.1099/00207713-50-3-1065/CITE/REFWORKS

Monteiro, C., Saxena, I., Wang, X., Kader, A., Bokranz, W., Simm, R., … Römling, U. (2009). Characterization of cellulose production in *Escherichia coli* Nissle 1917 and its biological consequences. *Environmental Microbiology*, *11*(5), 1105–1116. https://doi.org/10.1111/j.1462-2920.2008.01840.x

Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Sundaramurthi, J. C., Lee, J., … Reddy, T. B. K. (2021). Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Research*, *49*(D1), D723–D733. https://doi.org/10.1093/NAR/GKAA983

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*(1), 263–273. https://doi.org/10.1101/SQB.1986.051.01.032

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., & Yarmush, M. L. (2007). Integrated Energy and Flux Balance Based Multiobjective Framework for Large-Scale Metabolic Networks. *Annals of Biomedical Engineering*, *35*(6), 863–885. https://doi.org/10.1007/s10439-007-9283-0

Nakagawa, S., Shataih, Z., Banta, A., Beveridge, T. J., Sako, Y., & Reysenbach, A. L. (2005). *Sulfurihydrogenibium yellowstonense* sp. nov., an extremely thermophilic, facultatively heterotrophic, sulfur-oxidizing bacterium from Yellowstone National Park, and emended descriptions of the genus *Sulfurihydrogenibium*. *International Journal of Systematic and Evolutionary Microbiology*, *55*(6), 2263–2268. https://doi.org/10.1099/ijs.0.63708-0

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., … Eloe-Fadrosh, E. A. (2020). A genomic catalog of Earth's microbiomes. *Nature Biotechnology 2020 39:4*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nazaries, L., Pan, Y., Bodrossy, L., Baggs, E. M., Millard, P., Murrell, J. C., & Singh, B. K. (2013). Evidence

of microbial regulation of biogeochemical cycles from a study on methane flux and land use change. *Applied and Environmental Microbiology*, *79*(13), 4031–4040. https://doi.org/10.1128/AEM.00095-13

NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(Database issue), D8. https://doi.org/10.1093/NAR/GKX1095

Neidhardt, F. C., Ingraham, J. L., & Schaechter, M. (1990). *Physiology of the Bacterial Cell: a Molecular Approach*. Sunderland, MA: Sinauer Associates.

Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., … Zhu, D. (2010). A catalog of reference genomes from the human microbiome. *Science*, *328*(5981), 994–999. https://doi.org/10.1126/SCIENCE.1183605

Nicolaus, B., Kambourova, M., & Oner, E. T. (2010). Exopolysaccharides from extremophiles: From fundamentals to biotechnology. *Environmental Technology*, *31*(10), 1145–1158. https://doi.org/10.1080/09593330903552094

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., … Erwin, Z. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology 2014 32:8*, *32*(8), 822–828. https://doi.org/10.1038/nbt.2939

Nogales, J., Gudmundsson, S., Duque, E., Ramos, J. L., & Palsson, B. O. (2017). Expanding the computable reactome in *Pseudomonas putida* reveals metabolic cycles providing robustness. *BioRxiv*, 139121. https://doi.org/10.1101/139121

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/GR.213959.116/-/DC1

O'Connell, K. P., Goodman, R. M., & Handelsman, J. (1996). Engineering the rhizosphere: Expressing a bias. *Trends in Biotechnology*, *14*(3), 83–88. https://doi.org/10.1016/0167-7799(96)80928-0

Oliveira, A. P., Nielsen, J., & Förster, J. (2005). Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiology*, *5*, 39. https://doi.org/10.1186/1471-2180-5-39

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–248. https://doi.org/10.1038/nbt.1614

Oyama, K., Shimada, K., Ishibashi, J. ichiro, Miki, H., & Okibe, N. (2018). Silver-catalyzed bioleaching of enargite concentrate using moderately thermophilic microorganisms. *Hydrometallurgy*, *177*, 197–204. https://doi.org/10.1016/J.HYDROMET.2018.03.014

Pacheco, A. R., & Segrè, D. (2021). An evolutionary algorithm for designing microbial communities via environmental modification. *Journal of the Royal Society Interface*, *18*(179). https://doi.org/10.1098/RSIF.2021.0348

Pandey, H. M., Chaudhary, A., & Mehrotra, D. (2014). A comparative review of approaches to prevent premature convergence in GA. *Applied Soft Computing*, *24*, 1047–1077. https://doi.org/10.1016/J.ASOC.2014.08.025

Park, S. Y., Yang, D., Ha, S. H., & Lee, S. Y. (2018). Metabolic Engineering of Microorganisms for the Production of Natural Compounds. *Advanced Biosystems*, *2*(1), 1700190. https://doi.org/10.1002/adbi.201700190

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996. https://doi.org/10.1038/NBT.4229

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., & Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLOS Computational Biology*, *12*(7), e1004977. https://doi.org/10.1371/JOURNAL.PCBI.1004977

Patil, K. R., Rocha, I., Förster, J., & Nielsen, J. (2005). Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics*, *6*(1), 1–12. https://doi.org/10.1186/1471-2105-6-308

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*, *27*(13), i94. https://doi.org/10.1093/BIOINFORMATICS/BTR216

Pereira, V., Cruz, F., & Rocha, M. (2021). MEWpy: a computational strain optimization workbench in Python. *Bioinformatics*, *37*(16), 2494–2496.

https://doi.org/10.1093/BIOINFORMATICS/BTAB013

Perevalova, A. A., Kublanov, I. V., Bidzhieva, S. K., Mukhopadhyay, B., Bonch-Osmolovskaya, E. A., & Lebedinsky, A. V. (2016). Reclassification of *Desulfurococcus mobilis* as a synonym of *Desulfurococcus mucosus*, *Desulfurococcus fermentans* and *Desulfurococcus kamchatkensis* as synonyms of *Desulfurococcus amylolyticus*, and emendation of the <i>D. m. *International Journal of Systematic and Evolutionary Microbiology*, *66*(1), 514–517. https://doi.org/10.1099/IJSEM.0.000747

Perez-Garcia, O., Chandran, K., Villas-Boas, S. G., & Singhal, N. (2016). Assessment of nitric oxide (NO) redox reactions contribution to nitrous oxide (N2O) formation during nitrification using a multispecies metabolic network model. *Biotechnology and Bioengineering*, *113*(5), 1124–1136. https://doi.org/10.1002/BIT.25880

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, *98*(17), 9748–9753. https://doi.org/10.1073/PNAS.171285098

Ponomarova, O., Gabrielli, N., Sevin, D. C., Mulleder, M., Zirngibl, K., Bulyha, K., ... Patil, K. R. (2017). Yeast creates a niche for symbiotic lactic acid bacteria through nitrogen overflow. *Cell Syst*, *5*(4), 345-357 e6. https://doi.org/10.1016/j.cels.2017.09.002

Popp, D., & Centler, F. (2020). μBialSim: Constraint-Based Dynamic Simulation of Complex Microbiomes. *Frontiers in Bioengineering and Biotechnology*, *8*, 574. https://doi.org/10.3389/FBIOE.2020.00574/BIBTEX

Probst, A. J., Weinmaier, T., DeSantis, T. Z., Santo Domingo, J. W., & Ashbolt, N. (2015). New Perspectives on Microbial Community Distortion after Whole-Genome Amplification. *PLOS ONE*, *10*(5), e0124158. https://doi.org/10.1371/JOURNAL.PONE.0124158

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology 2017 35:9*, *35*(9), 833–844. https://doi.org/10.1038/nbt.3935

Raman, K., & Chandra, N. (2009). Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, *10*(4), 435–449. https://doi.org/10.1093/BIB/BBP011

Ramon, C., Gollub, M. G., & Stelling, J. (2018). Integrating –omics data into genome-scale metabolic network models: principles and challenges. *Essays in Biochemistry*, *62*(4), 563–574. https://doi.org/10.1042/EBC20180011

Rana, S., & Upadhyay, L. S. B. (2020). Microbial exopolysaccharides: Synthesis pathways, types and their commercial applications. *International Journal of Biological Macromolecules*, *157*, 577–583. https://doi.org/10.1016/J.IJBIOMAC.2020.04.084

Raposo, P., Padrão, J., & Dias, O. (2017). *Reconstruction of the genome-scale metabolic model of* Nitrosomonas europaea. University of Minho.

Reischl, B., Ergal, İ., & Rittmann, S. K. M. R. (2018). Metabolic reconstruction and experimental verification of glucose utilization in *Desulfurococcus amylolyticus* DSM 16532. *Folia Microbiologica*, *63*(6), 713–723. https://doi.org/10.1007/s12223-018-0612-5

Riessen, S., & Antranikian, G. (2001). Isolation of *Thermoanaerobacter keratinophilus* sp. nov., a novel thermophilic, anaerobic bacterium with keratinolytic activity. *Extremophiles : Life under Extreme Conditions*, *5*(6), 399–408. https://doi.org/10.1007/S007920100209

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature 2013 499:7459*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Rocha, I., Förster, J., & Nielsen, J. (2008). Design and Application of Genome-Scale Reconstructed Metabolic Models. *Microbial Gene Essentiality - Protocols and Bioinformatics*, *416*, 409–431. https://doi.org/10.1007/978-1-59745-321-9

Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J. P., … Rocha, M. (2010). OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC Systems Biology 2010 4:1*, *4*(1), 1–12. https://doi.org/10.1186/1752-0509-4-45

Rocha, M., Maia, P., Mendes, R., Pinto, J. P., Ferreira, E. C., Nielsen, J., … Rocha, I. (2008). Natural computation meta-heuristics for the *in silico* optimization of microbial strains. *BMC Bioinformatics*, *9*(1), 1–16. https://doi.org/https://doi.org/10.1186/1471-2105-9-499

Römling, U. (2002). Molecular biology of cellulose production in bacteria. *Research in Microbiology*, *153*(4), 205–212. https://doi.org/10.1016/S0923-2508(02)01316-5

Römling, U., & Galperin, M. Y. (2015). Bacterial cellulose biosynthesis: Diversity of operons, subunits, products, and functions. *Trends in Microbiology*, *23*(9), 545–557. https://doi.org/10.1016/j.tim.2015.05.005

Romling, U., Galperin, M. Y., & Gomelsky, M. (2013). Cyclic di-GMP: the First 25 Years of a Universal Bacterial Second Messenger. *Microbiology and Molecular Biology Reviews*, *77*(1), 1–52. https://doi.org/10.1128/mmbr.00043-12

Rothschild, L. J., & Mancinelli, R. L. (2001). Life in extreme environments. *Nature 2001 409:6823*, *409*(6823), 1092–1101. https://doi.org/10.1038/35059215

Rudolph, B., Gebendorfer, K. M., Buchner, J., & Winter, J. (2010). Evolution of *Escherichia coli* for Growth at High Temperatures. *The Journal of Biological Chemistry*, *285*(25), 19029. https://doi.org/10.1074/JBC.M110.103374

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., … Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLOS Biology*, *5*(3), e77. https://doi.org/10.1371/JOURNAL.PBIO.0050077

Sahm, K., John, P., Nacke, H., Wemheuer, B., Grote, R., Daniel, R., & Antranikian, G. (2013). High abundance of heterotrophic prokaryotes in hydrothermal springs of the Azores as revealed by a network of 16S rRNA gene-based methods. *Extremophiles*, *17*(4), 649–662. https://doi.org/10.1007/s00792-013-0548-2

Saier, M. H., Reddy, V. S., Moreno-Hagelsieb, G., Hendargo, K. J., Zhang, Y., Iddamsetty, V., … Medrano-Soto, A. (2021). The transporter classification database (TCDB): 2021 update. *Nucleic Acids Research*, *49*(D1), D461–D467. https://doi.org/10.1093/nar/gkaa1004

Salvy, P., & Hatzimanikatis, V. (2020). The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models. *Nature Communications 2020 11:1*, *11*(1), 1–17. https://doi.org/10.1038/s41467-019-13818-7

Sánchez-Andrea, I., Guedes, I. A., Hornung, B., Boeren, S., Lawson, C. E., Sousa, D. Z., … Stams, A. J. M. (2020). The reductive glycine pathway allows autotrophic growth of *Desulfovibrio desulfuricans*. *Nature Communications 2020 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-18906-7

Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E. J., & Nielsen, J. (2017). Improving

the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology*, *13*(8), 935. https://doi.org/10.15252/MSB.20167411

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/PNAS.74.12.5463

Santos, M. A., Williams, R. A. D., & Da Costa, M. S. (1989). Numerical Taxonomy of *Thermus* Isolates from Hot Springs in Portugal. *Systematic and Applied Microbiology*, *12*(3), 310–315. https://doi.org/10.1016/S0723-2020(89)80079-7

Santos, S., & Rocha, I. (2016). *A computation tool for the estimation of biomass composition from genomic and transcriptomic information*. *Advances in Intelligent Systems and Computing* (Vol. 477). https://doi.org/10.1007/978-3-319-40126-3_17

Sarmiento, F., Peralta, R., & Blamey, J. M. (2015). Cold and Hot Extremozymes: Industrial Relevance and Current Trends. *Frontiers in Bioengineering and Biotechnology*, *3*(OCT). https://doi.org/10.3389/FBIOE.2015.00148

Schellenberger, J., & Palsson, B. (2009). Use of Randomized Sampling for Analysis of Metabolic Networks. *Journal of Biological Chemistry*, *284*(9), 5457–5461. https://doi.org/10.1074/JBC.R800048200

Schellenberger, J., Park, J. O., Conrad, T. M., & Palsson, B. Ø. (2010). BiGG : a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions Database. *BMC Bioinformatics*, *11*, 213–313.

Schilling, C. H., & Palsson, B. (2000). Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *Journal of Theoretical Biology*, *203*(3), 249–283. https://doi.org/10.1006/jtbi.2000.1088

Schleper, C., Puehler, G., Holz, I., Gambacorta, A., Janekovic, D., Santarius, U., … Zillig, W. (1995). *Picrophilus* gen. nov., fam. nov.: A novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. *Journal of Bacteriology*, *177*(24), 7050–7059. https://doi.org/10.1128/jb.177.24.7050-7059.1995

Schuetz, R., Kuepfer, L., & Sauer, U. (2007). Systematic evaluation of objective functions for predicting

intracellular fluxes in *Escherichia coli*. *Molecular Systems Biology*, *3*(1), 119. https://doi.org/10.1038/MSB4100162

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature Methods 2008 5:1*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods 2017 14:11*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Seaver, S. M. D., Liu, F., Zhang, Q., Jeffryes, J., Faria, J. P., Edirisinghe, J. N., … Henry, C. S. (2021). The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Research*, *49*(D1), D575–D588. https://doi.org/10.1093/NAR/GKAA746

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, *12*(6), 1–18. https://doi.org/10.1186/GB-2011-12-6-R60

Segerer, A., Langworthy, T. A., & Stetter, K. O. (1988). *Thermoplasma acidophilum* and *Thermoplasma volcanium* sp. nov. from Solfatara Fields. *Systematic and Applied Microbiology*, *10*(2), 161–171. https://doi.org/10.1016/S0723-2020(88)80031-6

Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, *99*(23), 15112–15117. https://doi.org/10.1073/pnas.232349399

Sgobba, E., Stumpf, A. K., Vortmann, M., Jagmann, N., Krehenbrink, M., Dirks-Hofmeister, M. E., … Wendisch, V. F. (2018). Synthetic *Escherichia coli-Corynebacterium glutamicum* consortia for l-lysine production from starch and sucrose. *Bioresource Technology*, *260*, 302–310. https://doi.org/10.1016/J.BIORTECH.2018.03.113

Sgobba, E., & Wendisch, V. F. (2020). Synthetic microbial consortia for small molecule production. *Current Opinion in Biotechnology*, *62*, 72–79. https://doi.org/10.1016/J.COPBIO.2019.09.011

Shi, Z. (2019). Gut Microbiota: An Important Link between Western Diet and Chronic Diseases. *Nutrients*, *11*(10). https://doi.org/10.3390/NU11102287

Shinfuku, Y., Sorpitiporn, N., Sono, M., Furusawa, C., Hirasawa, T., & Shimizu, H. (2009). Development and experimental verification of a genome-scale metabolic model for *Corynebacterium glutamicum*. *Microbial Cell Factories*, *8*, 43. https://doi.org/10.1186/1475-2859-8-43

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/MSB.2011.75

Singh, D., & Lercher, M. J. (2020). Network reduction methods for genome-scale metabolic models. *Cellular and Molecular Life Sciences*, *77*(3), 481–488. https://doi.org/https://doi.org/10.1007/s00018-019-03383-z

Skirnisdottir, S., Hreggvidsson, G. O., Hjörleifsdottir, S., Marteinsson, V. T., Petursdottir, S. K., Holst, O., & Kristjansson, J. K. (2000). Influence of sulfide and temperature on species composition and community structure of hot spring microbial mats. *Applied and Environmental Microbiology*, *66*(7), 2835–2841. https://doi.org/10.1128/AEM.66.7.2835-2841.2000

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/CPMB.59

Smith, P. F., Langworthy, T. A., Mayberry, W. R., & Hougland, A. E. (1973). Characterization of the membranes of *Thermoplasma acidophilum*. *Journal of Bacteriology*, *116*(2), 1019–1028. https://doi.org/10.1128/jb.116.2.1019-1028.1973

Stephanopoulos, G. (1999). Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering*, *1*(1), 1–11. https://doi.org/10.1006/MBEN.1998.0101

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., & Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.*, *3*, 92. https://doi.org/10.1038/msb4100131

Strazzulli, A., Fusco, S., Cobucci-Ponzano, B., Moracci, M., & Contursi, P. (2017). Metagenomics of microbial and viral life in terrestrial geothermal environments. *Reviews in Environmental Science and Bio/Technology 2017 16:3*, *16*(3), 425–454. https://doi.org/10.1007/S11157-017-9435-0

Sun, F., Zhang, X. Z., Myung, S., & Zhang, Y. H. P. (2012). Thermophilic *Thermotoga maritima* ribose-5-

phosphate isomerase RpiB: Optimized heat treatment purification and basic characterization. *Protein Expression and Purification*, *82*(2), 302–307. https://doi.org/10.1016/J.PEP.2012.01.017

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237). https://doi.org/DOI: 10.1126/science.1261359

Takacs-vesbach, C., Inskeep, W. P., Jay, Z. J., Herrgard, M. J., Rusch, D. B., Tringe, S. G., … Xie, G. (2013). Metagenome Sequence Analysis of Filamentous Microbial Communities Obtained from Geochemically Distinct Geothermal Channels Reveals Specialization of Three Aquificales Lineages. *Frontiers in Microbiology*, *0*(MAY), 84. https://doi.org/10.3389/FMICB.2013.00084

Takai, K., Kobayashi, H., Nealson, K. H., & Horikoshi, K. (2003). *Sulfurihydrogenibium subterraneum* gen. nov., sp. nov., from a subsurface hot aquifer. *International Journal of Systematic and Evolutionary Microbiology*, *53*(3), 823–827. https://doi.org/10.1099/ijs.0.02506-0

Tanner, M. A., Goebel, B. M., Dojka, M. A., & Pace, N. R. (1998). Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Applied and Environmental Microbiology*, *64*(8), 3110–3113. https://doi.org/10.1128/AEM.64.8.3110-3113.1998

Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W. M., Siezen, R. J., & Smid, E. J. (2006). Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *The Journal of Biological Chemistry*, *281*(52), 40041–40048. https://doi.org/10.1074/jbc.M606263200

Thiele, I., & Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. https://doi.org/10.1038/nprot.2009.203.A

Thommes, M., Wang, T., Zhao, Q., Paschalidis, I. C., & Segrè, D. (2019). Designing Metabolic Division of Labor in Microbial Communities. *MSystems*, *4*(2), 263–281. https://doi.org/10.1128/MSYSTEMS.00263-18

Tonda, A. (2020). Inspyred: Bio-inspired algorithms in Python. *Genetic Programming and Evolvable Machines*, *21*(1–2), 269–272. https://doi.org/10.1007/S10710-019-09367-Z/FIGURES/1

Tsoi, R., Dai, Z., & You, L. (2019). Emerging strategies for engineering microbial communities. *Biotechnology Advances*, *37*(6), 107372. https://doi.org/10.1016/J.BIOTECHADV.2019.03.011

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. *Nature 2007 449:7164*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Uk Kim, H., Yong Kim, T., & Yup Lee, S. (2008). Metabolic flux analysis and metabolic engineering of microorganisms. *Molecular BioSystems*, *4*(2), 113–120. https://doi.org/10.1039/B712395G

Ulas, T., Riemer, S. A., Zaparty, M., Siebers, B., & Schomburg, D. (2012). Genome-Scale Reconstruction and Analysis of the Metabolic Network in the Hyperthermophilic Archaeon Sulfolobus Solfataricus. *PLoS ONE*, *7*(8), 43401. https://doi.org/10.1371/journal.pone.0043401

Van den Burg, B. (2003). Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology*, *6*(3), 213–218. https://doi.org/10.1016/S1369-5274(03)00060-2

Varma, A., & Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology 1994 12:10*, *12*(10), 994–998. https://doi.org/10.1038/nbt1094-994

Volkl, P., Huber, R., Drobner, E., Rachel, R., Burggraf, S., Trincone, A., & Stetter, K. O. (1993). *Pyrobaculum aerophilum* sp. nov., a novel nitrate-reducing hyperthermophilic archaeum. *Applied and Environmental Microbiology*, *59*(9), 2918–2926. https://doi.org/10.1128/aem.59.9.2918-2926.1993

Wang, R., Zhao, S., Wang, Z., & Koffas, M. A. (2020). Recent advances in modular co-culture engineering for synthesis of natural products. *Current Opinion in Biotechnology*, *62*, 65–71. https://doi.org/10.1016/J.COPBIO.2019.09.004

Ward, D. M., Weller, R., & Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature 1990 345:6270*, *345*(6270), 63–65. https://doi.org/10.1038/345063a0

Watkin, E. L. J., Keeling, S. E., Perrot, F. A., Shiers, D. W., Palmer, M. L., & Watling, H. R. (2009). Metals tolerance in moderately thermophilic isolates from a spent copper sulfide heap, closely related to *Acidithiobacillus caldus*, *Acidimicrobium ferrooxidans* and *Sulfobacillus thermosulfidooxidans*. *Journal of Industrial Microbiology and Biotechnology*, *36*(3), 461–465.

https://doi.org/10.1007/s10295-008-0508-5

Watling, H. R., Collinson, D. M., Fjastad, S., Kaksonen, A. H., Li, J., Morris, C., ... Shiers, D. W. (2014). Column bioleaching of a polymetallic ore: Effects of pH and temperature on metal extraction and microbial community structure. *Minerals Engineering*, *58*, 90–99. https://doi.org/10.1016/j.mineng.2014.01.022

White, J. R., Nagarajan, N., & Pop, M. (2009). Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLOS Computational Biology*, *5*(4), e1000352. https://doi.org/10.1371/JOURNAL.PCBI.1000352

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., ... Wilmes, P. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME Journal 2016 10:11*, *10*(11), 2557–2568. https://doi.org/10.1038/ismej.2016.45

Wilson, M. C., Mori, T., Rückert, C., Uria, A. R., Helf, M. J., Takada, K., ... Piel, J. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature*, *506*(7486), 58–62. https://doi.org/10.1038/NATURE12959

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/PNAS.74.11.5088

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), 1–12. https://doi.org/10.1186/GB-2014-15-3-R46

Woolston, B. M., Edgar, S., & Stephanopoulos, G. (2013). Metabolic Engineering: Past and Future. *Annual Review of Chemical and Biomolecular Engineering*, *4*(1), 259–288. https://doi.org/10.1146/annurev-chembioeng-061312-103312

Xavier, J. C., Patil, K. R., & Rocha, I. (2017). Integration of Biomass Formulations of Genome-Scale Metabolic Models with Experimental Data Reveals Universally Essential Cofactors in Prokaryotes. *Metabolic Engineering*, *39*, 200–208. https://doi.org/10.1016/j.ymben.2016.12.002

Ye, C., Zou, W., Xu, N., & Liu, L. (2014). Metabolic model reconstruction and analysis of an artificial microbial ecosystem for vitamin C production. *Journal of Biotechnology*, *182–183*(1), 61–67. https://doi.org/10.1016/j.jbiotec.2014.04.027

Yim, L. C., Hongmei, J., Aitchison, J. C., & Pointing, S. B. (2006). Highly diverse community structure in a remote central Tibetan geothermal spring does not display monotonic variation to thermal stress. *FEMS Microbiology Ecology*, *57*(1), 80–91. https://doi.org/10.1111/J.1574-6941.2006.00104.X

Yishai, O., Bouzon, M., Döring, V., & Bar-Even, A. (2018). *In Vivo* Assimilation of One-Carbon via a Synthetic Reductive Glycine Pathway in Escherichia coli. *ACS Synthetic Biology*, *7*(9), 2023–2028. https://doi.org/10.1021/acssynbio.8b00131

Yoshino, J. I., Sugiyama, Y., Sakuda, S., Kodama, T., Nagasawa, H., Ishii, M., & Igarashi, Y. (2001). Chemical structure of a novel aminophospholipid from *Hydrogenobacter thermophilus* strain TK-6. *Journal of Bacteriology*, *183*(21), 6302–6304. https://doi.org/10.1128/JB.183.21.6302-6304.2001

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., … Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608–1615. https://doi.org/10.1093/BIOINFORMATICS/BTQ249

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., & Forney, L. J. (2012). Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. *PLOS ONE*, *7*(3), e33865. https://doi.org/10.1371/JOURNAL.PONE.0033865

Zaramela, L. S., Moyne, O., Kumar, M., Zuniga, C., Tibocha-Bonilla, J. D., & Zengler, K. (2021). The sum is greater than the parts: exploiting microbial communities to achieve complex functions. *Current Opinion in Biotechnology*, *67*, 149–157. https://doi.org/10.1016/J.COPBIO.2021.01.013

Zayulina, K. S., Kochetkova, T. V., Piunova, U. E., Ziganshin, R. H., Podosokorskaya, O. A., & Kublanov, I. V. (2020). Novel Hyperthermophilic Crenarchaeon *Thermofilum adornatum* sp. nov. Uses GH1, GH3, and Two Novel Glycosidases for Cellulose Hydrolysis. *Frontiers in Microbiology*, *10*, 2972. https://doi.org/10.3389/FMICB.2019.02972/BIBTEX

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., & Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc. Natl. Acad. Sci. U. S. A.*, *112*(20), 6449–6454. https://doi.org/10.1073/pnas.1421834112

Zerfaß, C., Chen, J., & Soyer, O. S. (2018). Engineering microbial communities using thermodynamic principles and electrical interfaces. *Current Opinion in Biotechnology*, *50*, 121–127.

https://doi.org/10.1016/J.COPBIO.2017.12.004

Zhang, W., Liu, H., Li, X., Liu, D., Dong, X. T., Li, F. F., ... Yuan, Y. J. (2017). Production of naringenin from D-xylose with co-culture of *E. coli* and *S. cerevisiae*. *Engineering in Life Sciences*, *17*(9), 1021–1029. https://doi.org/10.1002/ELSC.201700039

Zhou, J., Xue, K., Xie, J., Deng, Y., Wu, L., Cheng, X., ... Luo, Y. (2011). Microbial mediation of carbon-cycle feedbacks to climate warming. *Nature Climate Change 2012 2:2*, *2*(2), 106–110. https://doi.org/10.1038/nclimate1331

Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., & Lovley, D. R. (2011). Genome-scale dynamic modeling of the competition between *Rhodoferax* and *Geobacter* in anoxic subsurface environments. *The ISME Journal*, *5*(2), 305–316. https://doi.org/10.1038/ismej.2010.117

Zomorrodi, A R, & Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.*, *8*(2), e1002363. https://doi.org/10.1371/journal.pcbi.1002363

Zomorrodi, Ali R., Islam, M. M., & Maranas, C. D. (2014). D-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, *3*(4), 247–257. https://doi.org/10.1021/sb4001307

Zomorrodi, Ali R., & Segrè, D. (2016). Synthetic Ecology of Microbes: Mathematical Models and Applications. *Journal of Molecular Biology*, *428*(5), 837–861. https://doi.org/10.1016/J.JMB.2015.10.019

Zorrilla, F., Buric, F., Patil, K. R., & Zelezniak, A. (2021). metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Research*, *49*(21), e126–e126. https://doi.org/10.1093/NAR/GKAB815

Zorrilla, F., Patil, K. R., & Zelezniak, A. (2021). metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *BioRxiv*, 2020.12.31.424982. https://doi.org/10.1101/2020.12.31.424982