Doctoral theses at NTNU, 2023:205

Jakob Peder Pettersen

Integrating large-scale data for biological network construction and analysis

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Jakob Peder Pettersen

Integrating large-scale data for biological network construction and analysis

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2023

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Natural Sciences
Department of Biotechnology and Food Science

© Jakob Peder Pettersen

ISBN 978-82-326-7112-0 (printed ver.)
ISBN 978-82-326-7111-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:205

Printed by NTNU Grafisk senter

Acknowledgements

The work presented in this thesis was carried out at the Norwegian University of Science and Technology (NTNU) for the degree of Philosophiae Doctor (Ph.D.). The research was funded by the Research Council of Norway as part of the ERA-NET project CoolWine (project number 283862).

First of all, I would like to thank my main (90% according to the contract) supervisor Professor Eivind Almaas for his committed and critical guidance on the project. He is a true scientist with all that entails and I appreciate his efforts to teach me how to become one. The remaining 10% of the supervision duty has been undertaken by Professor Olav Vadstein who has given me good advice and made me love microorganisms.

My collaborators in the CoolWine project have played a crucial role in my research and I therefore express my gratitude for supporting me despite having different scientific approaches and backgrounds. Paula and Sandra require a special mention because of their patient and skilful support in improving and troubleshooting the CoolWine models for Paper IV.

I am grateful for my colleges who have been important for my working environment and academic development. This does in particular include Madeleine for showing diligence, an excellent ability for cooperation and always being in a good mood. In addition, I would like to thank Vetle for reviewing this thesis and providing constructive feedback.

Furthermore, I would like to thank Henrik Bengtsson at University of California, San Francisco for our collaboration and discussions on the `matrixStats` R package. I have found it fulfilling to have such a collaborator who has broadened and developed my view on scientific software engineering.

My friends in Oslo; Diamond, Fredrik and Jenusiyia have also been important for me during the time as a PhD candidate. Their open-mindedness, friendliness and sense of humour have cheered me up and motivated me to remain vigilant in my efforts.

Last, but not least, this thesis would not have been possible without the practical and mental support from my mother who has always thought of me and been there when I have needed it. I also admire her fascination for my work even though she most likely only will read the acknowledgements section of this thesis.

Trondheim *Jakob Peder Pettersen*
March, 2023

”*Never in the history of mankind has it been possible to produce so many wrong answers so quickly!*”
Carl-Erik Fröberg- numerical analyst

Abstract

Collection of biological data proceeds at a rate greater than ever thanks to advances in high throughput technologies. This has created large amounts of data to be systematized and interpreted. Networks represent the entities in question and the connections between them, and have proven to be one of the most useful means of storing and representing such knowledge. There are two major tasks when working with biological networks: (1) inference of the network from experimental data and (2) application of network models which utilize the networks in order to make predictions about the system's behavior. In this doctoral thesis, we will cover three applications of biological networks: microbial co-occurence, differential gene co-expression, and genome-scale metabolic models.

Microbial communities represent a vibrant field of study, primarily due to its relevance to human medicine. In addition, ensuring a good microbial community is important in rearing of fish larva in aquaculture. It has been shown that selecting for K-strategist bacteria over opportunistic r-strategists has the potential to dramatically improve fish health and survival. Therefore, in Paper I, we analyze a dataset of bacteria in seawater reactors being exposed to r- and K-selection with the selection regimes being switched halfway in the experiment. From the generated microbial co-occurrence networks, we observe that most associations are found between bacteria which are taxonomically related and most likely share the same environmental preferences. Furthermore, the microbial communities do not show signs of resilience to changes in the selection regime, so over time, signatures of past selection regimes are wiped out.

Faulty regulation of genes is often seen in diseases such as cancer, allergy and chronic fatigue. By measuring and comparing gene expression in sick patients and healthy controls, we can summarize the gene co-expression in a differential gene co-expression network. This helps to understand the gene regulatory mechanisms of the diseases and may result in novel targeted treatments. In paper II, we introduce the R package `csdR` which is an efficient and user-friendly implementation of the existing CSD (Conserved, Specific, and Differentiated) algorithm for differential co-expression. This package is available in the Bioconductor repository and is shown to be orders of magnitude faster than the original CSD implementation.

Genome-scale metabolic models (GEMs) cover the metabolic conversions of an organism represented by a network of reactions and metabolites. These models are generated from the enzymes present in the organism's genome and can be used to predict metabolic fluxes through methods such as Flux Balance Analysis (FBA).

Over time, many varieties and extensions of FBA have been developed. This includes incorporating enzyme constraints, accounting for time dynamics and taking the ambient temperature into consideration. In paper III, we assess a Bayesian modeling approach for inferring parameters for a temperature and enzyme constrained genome-scale metabolic model (etcGEM). We show that the existing procedure is unstable and therefore implement an alternative evolutionary algorithm.

GEMs are often used for genetic engineering and selection of organisms with desired properties. In Paper IV we use automatically reconstructed GEMs of five non-*Saccharomyces* yeast strains in order to evaluate their ability to produce wine with reduced alcohol content. This includes *Metschnikowia pulcherrima* which has already proven to be a good candidate for this task. Our results suggest that the desired properties of *Metschnikowia pulcherrima* is partially due to having Complex I of the electron transport chain which is missing in the other yeast strains.

A recurring problem encountered throughout the thesis is a lack of sufficient data to make reliable models and predictions. As a result, the supply of data is a more limiting factor than the algorithms used to analyze the data. However, we hope that these challenges can be alleviated by continued efforts into improvements to collection and management of biological data. This includes measures such as laboratory automation and collaborations on data management. In particular, we find the current formats for exchanging genome-scale models insufficient for dealing with FBA extensions such as dynamic FBA and enzyme constrained FBA in a reproducible manner. Hence, we encourage further development into the SBML standards for genome-scale models.

Contents

List of Tables

List of Figures

List of Abbreviations

(detc)FBA Extensions of FBA, consult table 4.1

ALE Adaptive Laboratory Evolution

ATP Adenosine TriPhosphate, the energy currency molecule of cells

BiGG BIochemical Genetic and Genomic, database of large scale metabolic models

BRENDA BRaunschweig ENzyme DAtabase, a public database on protein function

cDNA Complementary DNA, DNA synthesized from mRNA, commonly used in gene expression analyzes

CSD Name of the differential co-expression analysis method used in Paper II. Short for Conserved/Specific/Differentiated.

DGGE Denaturing Gradient Gel Electrophoresis, a method for studying 16S amplicons in a microbial community

FAIR Findable, Accessible, Interoperable and Reusable, guidelines for data management

FBA Flux Balance Analysis, a method to predict metabolic fluxes using linear programming

FBAwMC FBA With Molecular Crowding

FVA Flux Variability Analysis, a method to predict the varibility of fluxes in a GEM

GCN Gene Co-expression Network

GECKO Genome-scale model to account for Enzyme Constraints, using Kinetics and Omics, a framework for GEMs with enzymatic constraints

GEM GEnome-scale metabolic Model

gLV Generalized Lotka-Volterra, a differential equation model for describing population dynamics in a community of interacting organisms

GWAS Genome-Wide Association Study, a statistical procedure for establishing the connection between phenotypic features or diseases with genotype

LP Linear program, an optimiation problem with linear constrains and objective function

MMS Microbially Matured water System

MOMENT MetabOlic Modeling with ENzyme kineTics, another framework for GEMs with enzymatic constraints

mRNA Messenger RNA, molecules which are transcribed from DNA and contain the instruction to build a protein

NGAM Non-Growth Associated Maintenance, a pseudoreaction in a GEM acting as an ATP (energy) sink

ODE Ordinary Differential Equation, an equation that involves a function of one independent variable and its derivatives

OTU Operational Taxonomical Unit, used for grouping organisms on a custom level based on sequence similarity

PCR Polymerase Chain Reaction, laboratory method for amplifying specific DNA sequences

RAS Recirculating Aquaculture System

SABIO-RK System for the Analysis of Biochemical Pathways - Reaction Kinetics, a public database of biochemical reactions and information about their kinetics

SBML Systems Biology Markup Language, a software format for describing biological models sMOMENT short MOMENT, a simplification of the MOMENT approach yielding the same results

T-RFLP Terminal Restriction Fragment Length Polymorphism, a method for studying 16S amplicons in a microbial community

WGCNA Weighted Gene Co-expression Network Analysis, an R package for studying GCNs

List of Papers

Paper I

Robust bacterial co-occurence community structures are independent of r- and K-selection history
Jakob Peder Pettersen, Madeleine S. Gundersen, Eivind Almaas. Scientific Reports 11(1) 2021.

Project conceived by JPP and EA. Data provided by MSG. Software and analysis by JPP. Writing of article by JPP, EA and MSG.

Paper II

csdR, an R package for differential co-expression analysis
Jakob Peder Pettersen, Eivind Almaas. BMC Bioinformatics 23, 79 (2022).

Development, testing and benchmarking by JPP. Writing of article by JPP and EA.

Paper III

Parameter inference for enzyme and temperature constrained genome-scale models
Jakob Peder Pettersen, Eivind Almaas. Under review in Scientific Reports

Project conveived by JPP and EA. Software, analysis and visualisations by JPP. Writing of article by JPP and EA.

Paper IV

Genome-scale metabolic models reveal determinants of phenotypic differences in non-*Saccharomyces* yeasts

Jakob Peder Pettersen, Sandra Castillo, Paula Jouhten, Eivind Almaas. To be submitted

Creation of models from genome sequences and bioinformatics by SC and PJ. Incorporation of enzyme constraints and analysis by JPP. Writing of article by JPP, SC and EA.

Chapter 1

Introduction

1.1 Why do we care about networks?

In order to feed a growing world population, new means of producing food are required. This includes aquaculture of new fish species such as Atlantic cod (*Gadus morhua*)[1]. However, initial attempts to produce cod by aquaculture has suffered from considerable difficulties, limiting its adoption[2, 3]. This is to a great extent due to high mortality of the cod larva during the first mounts after hatching[3]. It has been suggested that a major cause of this mortality is predominance of opportunistic bacteria in the rearing environment, causing dysbiosis and considerable stress to the fish[4, 5, 6].

Climate change has resulted in warmer and sunnier summers in wine producing regions. This has resulted in increased amounts of sugar in the mature grapes. In turn, when the grape must is fermented by brewing yeast, the result is wine with increased alcohol content[7]. Not only does this stronger wine cause health issues and higher taxation, the increased alcohol content negatively affects the sensory profile[8, 9, 10].

Moving over to the field of medicine, cancer is a major cause of death and suffering. More refined and personalized approaches have been developed for cancer treatment the last years with decent success. Still, we lack a good understanding of the causes and progression of cancer and having the ability to cure any patient is a distant vision[11, 12, 13].

All of these problems have in common that they can partially be described, and hopefully solved, by network science. A network is a collection of nodes and links interconnecting the nodes. Besides being a pure mathematical concept, any system which has entities with some connections between them can be represented as a network.

In the case of the rearing a cod larva, a network can be constructed of the interactions between the bacteria surrounding the fish. Using the network, we can gain understanding of the interplay between the bacteria, the fish and their environment. As a result, we can create favorable rearing conditions for the fish and thus help increase survival[14, 5]. Similarly, a metabolic network can be created of the yeast metabolism and guide us to metabolic engineering approaches for reducing the alcohol output of the wine fermentation process[15, 16]. Finally, for understanding the causes and development of cancer, networks can be created of the gene co-expression in sick and healthy patients. These networks can in turn be compared and analyzed in order to obtain understanding into the causes of the disease and suggest therapeutic interventions[17, 18].

As the preceding examples illustrate, network science can be defined as the study of the network representation of systems in order to gain understanding and make predictions of the phenomena occurring in the systems[19, 20]. In this context, a network-based model is a conceptual or computational model where the constituents of the model are represented as a network or a collection of networks. In statistics, we speak of two different, yet related processes; inference and prediction[21]. Inference is the process by applying data to create or parametrize a model, whereas prediction is the process of forecasting results based on a model. Often for network models, scientists do both. Research and data collection is required for creating the network model, but the true benefits from the model comes when it is used for studying the properties and making predictions about the system in question.

1.2 Structure of this thesis

In this thesis, we will go through various applications of networks in biology. Even if the areas of applications differ, the overall procedure is similar: data is generated from experiments on the system in question. Modern high-throughput omics technologies such as DNA sequencing and mass spectroscopy have made it feasible to collect large amounts of data in a single experiment[22]. These data are then handed directly to the modeler or stored in databases for further use. From the experimental data, a network is created with dedicated computational tools (inference). After construction, the structure of the network is analyzed and integrated with other data. This produces information which can explain the behavior of the system in question. In turn, the new understanding can help the researchers come up with questions and hypothesis which can be tested in follow-up experiments[23, 24].

The first topic of this thesis is co-occurrence networks of microbial communities. In our case, we will focus on aquatic bacterial communities as running experiments and sampling from such environments is relatively easy. Until high-throughput sequencing was introduced, there was no easy method for characterizing all bacteria living in a natural environment, and research was focused on qualitative production capabilities of the community as a whole and isolation of individual strains. Today however, it is possible to get almost complete coverage of a bacterial community[25]. The primary motivation behind such studies is to get an understanding of the community in question and to provide an overview of the ecological role of each organism and the interplay between the organisms (interactions). Even if the real ecological interactions between bacteria is difficult to infer from quantification of the present organisms, statistical methods[26, 27, 28] are used to provide a meaningful representation of the community. Networks generated from such statistical methods can then be used to pinpoint the underlying ecological structure of the community. In paper I, we will use network inference and analysis for studying the microbial community dynamics in seawater chemostats.

Just like modern advances in sequencing technology has enabled creation of networks of microbial communities, novel high-throughput techniques, such as microarray[29, 30], RNAseq[31, 32], and Liquid Chromatography–Mass Spectrometry(LC-MS)[33] have enabled comprehensive studies of gene expression. Genes are expressed in different conditions and shed light on the underlying physiological and regulatory processes in the organism. For instance, if two genes are always expressed together, it might mean that the same regulation applies to both genes and that they are involved in the same process. Gene co-expression analysis is thus widely used in genetic studies for finding regulatory mechanisms, obtaining insight into relationships between genotypes and phenotypes, and associate diseases with the genes which trigger them[34]. One of the methods used in this context is the CSD (Conserved, Specific, and Differentiated) algorithm for differential co-expression analysis[35]. In paper II, we present the `R` package `csdR` which provides an efficient implementation of the CSD algorithm.

The third and final topic of this thesis is genome-scale metabolic modeling. A genome-scale metabolic model (GEM) is in its most basic sense a network of the reactions and metabolites presents in an organism. Unlike the two aforementioned applications of networks, the topology of GEMs is usually *not* created as a result of statistical inference. Most reactions taking place in living organisms are catalyzed by enzymes. Consequently, the basic approach for constructing metabolic model is based on finding the enzymes of the organism and determining which reaction(s) they catalyze[36]. The term *genome-scale* is used to emphasize that the model is supposed to cover all reactions catalyzed by the enzymes encoded

Table 1.1: A comparison of the types of networks covered in this thesis.

Field of study	**Nodes**	**Edges**	**Method of network inference**	**Common methods for description and prediction**
Microbial co-occurrence	Microbial OTUs (species)	Co-occurrence associations	Similarity measures to find all-to-all associations followed by significance testing	Module detection, phylogentic clustering
Gene differential co-expression	Genes	Differential co-expression (three different types)	Similarity measures to find all-to-all associations followed by selection of the strongest links	Module detection, gene enrichment analysis
Genome-scale metabolism	Metabolites	(Enzymatic) reactions	Reconstruction from genome annotations and literature	Flux Balance Analysis (FBA), Flux Variability Analysis (FVA)

in the genome[37]. However, creating genome-scale metabolic models is not a straight-forward task. Genome annotations may be missing, some reactions occur without enzyme catalysis, and some enzymes utilize alternative substrates (underground metabolism)[38, 39]. Over time, the scope and details captured by GEMs have widened[40, 41]. This includes including enzymatic constraints, time dynamics, and accounting for temperature dependence. In paper III, we will assess the properties of an enzyme and temperature constrained metabolic model (etcGEM). Also, in paper IV we will present a study of the properties of automatically generated models of non-*Saccharomyces* yeasts which includes temporal dynamics and enzyme constraints.

In Table 1.1 we compare the three topics in terms of the networks we generate and analyze.

Chapter 2

Network inference for microbial communities

2.1 Microbial communities

Microorganisms are found in almost all natural and man-made habitats and play a crucial role for all living organisms[42]. We call the collection of all microorganisms present in a habitat a *microbial community*. Even though microorganisms encompasses both prokaryotes (including bacteria and archaea), viruses and unicellular eukaryotes, we will limit our discussion to prokaryotes for practical purposes and refer to them as bacteria. On a global scale, microorganisms are essential actors in recycling of nutrient elements such as nitrogen, phosphorus and carbon[43]. Hence, despite the microorganisms being the most primitive living entities, they provide functions which larger organisms cannot conduct on their own. Besides being part of nutrient utilization on a global scale, the microbes also affect and benefit macroorganisms more directly. A prime example of this is how the microorganisms in ruminants live in symbiotic relationship with their hosts and help degrading cellulose, enabling cows and sheep to digest grass[44]. Most commonly, services such as nutrient recycling and symbiotic nutrient degradation are not carried out by a single type of organism, instead being a feature of the microbial community as a whole. This community consists of the populations of different types of microorganisms.

2.1.1 Why are microbial communities different?

The composition of a microbial community is defined by which bacteria are present and in what abundances. According to the quote ”Everything is everywhere, but, the environment selects" by Lourens Gerhard Marinus Baas Becking, the inventory of bacteria ready to colonize a habitat is static, but the environment determines which bacteria will be able to establish themselves in the community as a result of selection pressure[45, 46]. Although modern studies in biogeography have challenged the notion of global dispersal of all microorganisms[47], the common understanding is still that the environment is the most important factor for the composition of the microbial community. For instance, an underwater geothermal vent features entirely different nutrients and adaptational challenges compared to the rumen of a cow, so the microbial communities of these two habitats are totally different from each other.

Furthermore, microbial communities are not static over time and change both due to external pressure and intrinsic activities carried out by the community itself. This process of change in a microbial community is referred to as *community dynamics*[48]. For instance, the microbial community in a lake will adapt over the seasons as availability of inorganic nutrients, sunlight and carbon sources will vary.

2.1.2 r- and K-strategists

Alfred J. Lotka provided a classical ordinary differential equation (ODE) model for describing the populations dynamics of living organisms[49]. His logistic equation for population growth is as follows:

$$\frac{\mathrm{d}\,N}{\mathrm{d}\,t} = Nr\left(1 - \frac{N}{K}\right), \tag{2.1}$$

where N is the population size and r, K are parameters. According to this equation, the population will stabilize to the *carrying capacity* K in the long run. The parameter r is referred to as the *specific growth rate* and determines how fast the population grows when $N << K$. This equation has given rise to the concept of r and K life strategies[50]. An r-strategist is an organism which is optimized for fast growth and proliferation, but struggles when competing for limiting resources (high r, low K). On the contrary, a K-strategist reproduces slowly, but is better at competing for limited resources and thus dominate crowded environments (low r, high K). Even though this concept was originally applied for macroorganisms, Andrew and Harris[51] started to apply it for microorganisms. r-strategist bacteria, such as *Vibrio natriegens*, prefer unstable environments where nutrients are in excess and there is little competition for resources. In contrast, K-strategist bacteria, such as *Pelagibacter ubique*, prefer stable environments where nutrient supply is a limiting factor.

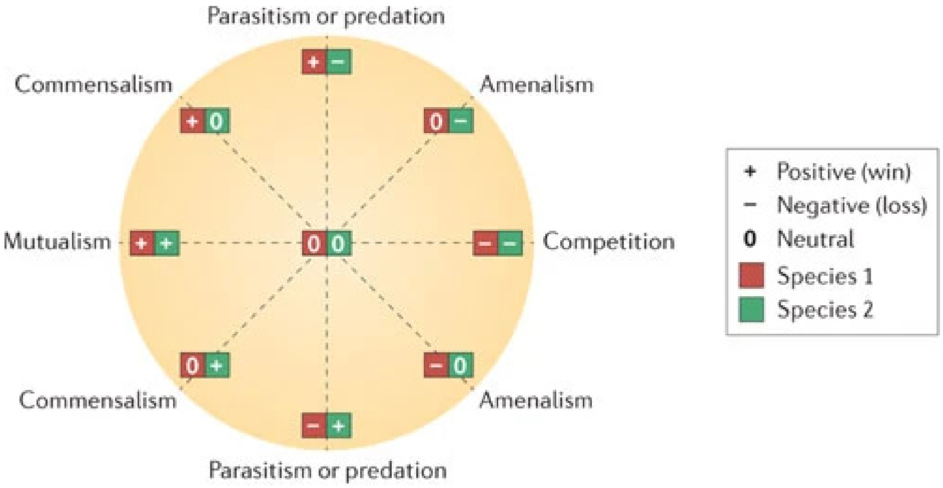
2.1.3 Implications for fish health

The health of fish is affected by the microbiota surrounding them and colonizing their surfaces and intestines, especially at early life stages[52, 53]. For this reason, survival of fish larva in aquaculture has shown enormous variability depending on the microbes in their rearing environment[14, 54, 55]. Most fish pathogens causing disease are r-strategists[51, 5]. Still, in many cases of mass fish death, no apparent pathogen could be pinpointed as the cause. Conversely, known pathogens can be detected in small quantities without causing any apparent harm. This has lead to the hypothesis that predominance of opportunistic cause disease when in excess through dysbiosis in fish without being true pathogens[5, 54]. This means that the opportunists can be present in small quantities without causing disease, but cause disease when they predominate the microbial community. Furthermore, research has suggested that it is not the bacterial load per se which contributes to the dysbiosis, but rather the composition of the microbial community and predominance of opportunistic r-strategists[56, 57].

Understanding of r- and K selection regimes has provided insight in how aquaculture systems should be designed[5]. A flow-through aquaculture system imports water from an external source at intake and discharges the water after passing through the rearing tanks. Since the incoming water is oligotrophic, the microbial carrying capacity is increased dramatically when feed is added to rearing tanks. This causes selection for r-strategists which are more likely to cause dysbiosis and death to the fish[58, 5]. To prevent such situations, strategies such as microbially matured water system (MMS)[59] and recirculating aquaculture system (RAS)[60] have been proposed in order to improve the quality of the microbiome. In a MMS system, the incoming water passes through a biofilter which seeds the water with non-opportunistic bacteria and thus reduces the potential for r-strategists to proliferate[59]. For a RAS system, the water is recirculated in the system and the water is only exchanged to a limited degree. Hence, a properly operated RAS creates a large advantage for K-strategists as the microbial population is close to the carrying capacity throughout the cycle and thus there is little room for r-strategists to proliferate[61, 5].

2.1.4 Microbial interactions and their importance for microbial communities

In a community, the bacteria are constantly affecting the life of each other. Hence, an ecological interaction is an effect that one microorganism has on another[28, 62]. Interactions are typically classified according to the net effect on the two actors involved (Figure 2.1). For two bacteria A and B, A positively interacts with B if B is getting a benefit of A's presence. Conversely, A negatively interacts with B if B suffers from a detrimental effect due to A's presence. A neutral interaction is



Nature Reviews | Microbiology

Figure 2.1: Wheel of microbial interactions showing the different forms of pairwise ecological interactions between two types of organisms where the effect for each of the two species involved is shown. For instance, amenalism adversarially affect one of the species (-), whereas the other species is not affected (0). Figure is taken from [28] with permission.

characterized by being an interaction where there is no net effect of being present in the same location. Note that interactions are asymmetrical in nature as the effect of A on B may not be the same as the effect of B on A. An example of this is parasitism where A is a parasite on B. B has then a positive interaction on A, whereas B is negatively affected by the interaction. We call the relationship between A and B mutualistic or cooperative if both A and B benefit from the interaction. This is the case in cross-feeding where A produces a compound which is essential for B, while B produces another compound which is essential for A. The opposite of mutualism is competition where A and B are negatively affected by each other presence. This often occurs when competing for the same limiting nutrients.

In our discussion of r- and K-selection, the K-strategists negatively affect the r strategists by being better for competing for resources[5, 51]. We know that certain biogeochemical pathways such as nitrification and methanogenesis depend on mutalistic interactions and that biogeochemical processes are due to the community as a whole rather than each of the species in isolation[63, 64]. Still, we do not know how ecological interactions between bacteria contribute to community assembly, stability and robustness to external perturbations. These are the main questions assessed in Paper I. Are the bacterial communities more based on collaboration (positive interactions) than competition (negative interactions)? Are there any specific interactions which can be exploited in order to favor the prevalence of K-strategists in aquaculture environments?

2.2 Bacterial systematics

Given the large variety of bacteria in existence, we need to systematize and structure the diversity in a systematic and hierarchical manner. Hence, a taxonomy is required for naming, describing and classifying bacteria. Taxonomical ranks are used for placing organisms into the taxonomy and follows a strict hierarchical order, meaning all organisms within a taxon at a certain rank also share the same taxonomical assignment for the ranks above. For bacteria, the standard set of ranks, in ascending order are Domain, Phylum, Class, Order, Family, Genus, and Species[65]. For instance, the gut bacterium species *Escherichia coli* belongs to the domain Bacteria, the phylum Proteobacteria, the class Gammaproteobacteria, the order Enterobacterales, the family Enterobacteriaceae and the genus *Escherichia*.

The International Code of Nomencalture of Prokaryotes[66] contains the guidelines for assigning scientific names to prokaryotic organisms. Moderns taxonomy is based on phylogeny, the representation of evolutionary history and relationship between organisms. The ideal of a phylogentically defined taxon is that it is monophylic, meaning that all individuals of the taxon decent from a common ancestor and that no other taxon on the same level contains decedents of this same ancestor[67]. This history can be recorded in a phylogentic tree which is a graphical representation of the evolutionary history of a collection of organisms[68]. Thus, every branch of the tree is intended to contain all and the only organisms decending from a single ancestor and the length of the branches are supposed to represent the phylogentic distance.

2.2.1 The 16S ribosomal RNA gene as a marker

Since the millions of years of bacterial evolution have not been observed directly, researchers have to look for markers of the evolutionary history. Classical bacteriology used phenotypic and morphological characteristics to systematize and determine bacteria. These classical approaches did however result in major inconsistencies with the phylogenic history[69]. When genome sequencing became available, this revolutionized the field of bacterial systematics as evolution would be analyzed with respect to changes in the genome.

The 16S ribosome RNA gene has proven to a good candidate for reconstructing phylogenies[70, 25]. It is universally distributed among all prokaryotes. Furthermore, it contains both regions which are conserved enough to be used as target sites for primers in addition to regions which are variable enough for differentiation to be done at genus level. These variable regions are usually very similar in closely related organisms and more different for distantly related organisms. Finally, the 16S rRNA gene is generally not believed to be exchanged by horizontal gene transfer[71].

The most direct approach to analyze the 16S sequence is to use amplify a specific section of the 16S rRNA gene using polymerase chain reaction (PCR) and then sequence it [72]. Originally, this was done through Sanger sequencing and worked well for pure cultures. However, for samples from natural environments this technique could not be used directly. Techniques such as T-RFLP[73], DGGE[74] and clone libraries[75] were developed for dealing with samples containing multiple sorts of bacteria, but either only gave a granular view of the microbial community or suffered from low throughput. After the advent of high-throughput sequencing, 16S amplicons stemming from a community of bacteria could be sequencing directly. This has made it possible to capture the entire diversity of the microbial community in a sample[76, 77, 78].

2.2.2 The species concept for bacteria

The species is the basic unit of taxonomic classification. Still, the species concept for prokaryotes is fuzzy and difficult because many of the species definitions applied for animals and plants are impossible or difficult to apply for microorganisms[69, 65]. According to the Prokaryote Code[66], in order to get a microbial species approved, there must exist a type strain which viable cultures are deposited in at least two separate locations. For this reason, the species concept is closely linked to studies done on pure cultures of isolates. Having bacteriology being based on studies of pure cultures works well for handbooks[79] for isolating, identifying, handling and storing of bacterial strains. The Prokaryote Code does not specify which criteria to apply for delimitating a species from strains within a species. DNA-DNA hybridization[80] was long the preferred method for defining new species, but there has been a development into using 16S rRNA similarity[81] and more recently average nucleotide identity[82] for this task.

2.2.3 Operational Taxonomical Unit (OTU)

For analyzing bacteria diversity in natural environments, adhering strictly to the Prokaryote Code's species concept by isolating and cultivating bacteria, is not a sound approach. First, it is observed that for many natural environments, only a small fraction of the bacteria can be grown and identified through such culture-dependent methods, referred to as ”The Great Plate Count Anomaly”[83]. Moreover, since these studies involve a large number of different organisms, characterizing and quantifying all the bacteria would be too labor intensive and time consuming even if they were all cultivable. As these diversity analyses usually are based on the 16S rRNA gene, an operational definition of the taxonomical relatedness is applied instead. Thus, all reads with an 16S similarity above a certain threshold (usually 97 or 98%) are grouped into a single Operational Taxonomical Unit (OTU)[84].

2.2.4 Typical workflow for 16S microbial community analysis

We will now briefly explain how the 16S microbial community analyses typically are conducted[85, 86, 87]: DNA from the samples is isolated and the V3-V4 region of the 16S-rRNA gene is amplified by PCR using primers with wide coverage. Typically, the primers also contain sequencing specific primers in addition to barcodes labeling the sample. The amplicons are run on an electrophoresis gel, purified and pooled together. This amplicon library is then sequenced on a benchtop sequencing machine such as Illumina MiSeq. When the sequencing data is collected, it is fed through bioinformatics pipelines such as USEARCH[88]. This bioinformatics pipeline trims away adapter sequences, demultiplexes reads originating from different samples, filters away low quality read, clusters the sequencing reads into OTUs and assigns taxonomical classification to the OTUs. The result of the pipeline is usually an OTU table showing the number of reads for each OTU in each sample.

2.3 Microbial datasets

In Paper I we analyze the selection-switch dataset[85] which stems from an experiment with laboratory-scale mesocosms inoculated with seawater. The reactors were subject to different selection regimes which were switched halfway during the experiment.

We also analyzed the Tara dataset[89] which stems from natural seawater samples taken from most of the world's oceanic regions without publishing our results. Finally, we considered the Carbon-Cycle dataset[90] which stems from samples taken in the Trondheimsfjord at various depths and seasons, combined with measurements of chemical composition of the seawater. An overview of the datasets can be found in Table 2.1.

2.4 Methods for analyzing microbial interactions

The most direct way of inferring interactions between microbes is through co-culture studies. Here, the growth of the two microbes in co-culture is compared to the growth rate of each microbe in a pure culture[91, 28]. Even though co-

Table 2.1: Overview of the microbial abundance datasets discussed in this thesis.

Name	Number of samples	Number of OTUs	Source
Selection-Switch	202	1 537	[85]
Tara	128	35 650	[89]
Carbon-Cycle	12	1 939	[90]

culture studies is considered the gold standard for determining interactions, it has numerous caveats making it difficult to apply in practice. First, only a fraction of natural bacteria can be grown in isolation. Second, replicating the natural context of the interaction is often difficult. Finally, natural environments consist of far more species than can be isolated and co-cultured. For these reasons, co-culture experiments is often an infeasible approach for untangling the numerous possible interactions in natural microbial communities. Instead, the most common methods for inferring microbial interactions are based on co-occurrence of all bacteria present in the environment. The most common and natural way to represent and model interactions is a network where the nodes are the different species (OTUs) of bacteria, while the interaction are the links between the nodes[27, 28, 26].

2.4.1 Co-occurrence based approaches

Various approaches exist for such network inference[27]. The simplest of these methods are based on similarity and dissimilarity measures such as Bray-Curtis similarity or Kullback-Leibler distance, where the pairwise similarity between vectors of OTU abundances are calculated and the most similar OTUs are grouped together using guilt by association. Some varieties of this method instead apply correlation metrics such as Person's product-moment correlation coefficient or the non-parametric Spearman rank coefficient[92]. Applying correlation- or similarity-based metrics on 16S microbial datasets has some bespoke challenges. First of all, the datasets are sparse, as most OTUs are present in a few samples only. This means that presence of shared zeros may cause two OTUs to appear more correlated than ecological theory would suggest[93]. Second, 16S sequencing typically only yields relative estimates of microbial abundances. Hence, the dataset becomes compositional with a sum-to-one constraint which distorts the underlying data[94, 95]. Therefore, correlation-based or dissimilarity-based methods are typically combined with different types of heuristics designed to deal the mentioned problems with microbial datasets. These heuristics include log-transforming abundances[93], applying shrinking methods[96, 97] and automated determination of the inclusion threshold for interactions[98]. Additionally, there exists various other approaches such as sparse graphical models which use more advanced statistical methods in order to generate a network of interactions[99].

2.4.2 Time series based approaches

A different class of microbial interaction inference methods are based on the generalized Lotka-Volterra (gLV) equation[100] which describes the community dynamics as a system of differential equations. These approaches require that the microbial data come from a time series and have the added benefit that the behavior of the system can be predicted in addition to being described. Depending on the approach, the OTU abundances are typically discretized and the gLV coefficients are found by solving a linear system[101, 102, 103], or the abundances of microbes are approximated by spline and forward-difference methods and maximum-likelihood or Bayesian algorithms are then applied to estimate the coefficients[104].

2.4.3 ReBoot

For Paper I, we chose to use a correlation- or similarity-based network inference method named ReBoot (Permutation-Renormalization and Bootstrap)[94]. While not being one of the most advanced methods for inferring microbial interactions, it has the benefit of being relatively simple to understand and customize.

The ReBoot approach is an enhancement to the basic similarity-based network approaches as it can estimate statistical significance of each of the resulting links, and applies heuristics such as bootstrapping and renormalization. As illustrated in Figure 2.2, the ReBoot method repetitively measures the similarity of abundances of each possible pair of OTUs under two conditions. Under the first condition, the samples of the original data are selected by bootstrapping[105] and the similarity of the relative abundance vectors of the two OTUs are calculated. For the other condition, the sample labels are randomly permuted (swapped) for one of the two OTUs before measuring the similarity of the relative abundances of the two OTUs. Permuting the samples breaks the sum-to-zero constraint of compositional data which is why ReBoot uses renormalization after permutation in order to retain the sum-to-zero constraint after permutation.

By taking the mean similarity obtained from bootstrapping the data and comparing it with the similarity when the samples are randomly permuted, we obtain a p-value reporting the probability of the association to be generated by chance. As all possible pairs of OTUs are checked as possible candidates for interactions, we have an extreme case of multiple testing. Hence, a p-value reported is only valid when considering a pair of OTUs in isolation, but not in context of constructing an interaction network. Hence, the Benjamini-Hochberg-Yekutieli procedure[106] is used to convert the p-values into q-values which represent the *false discovery rate* and are valid when considering all interactions at once.

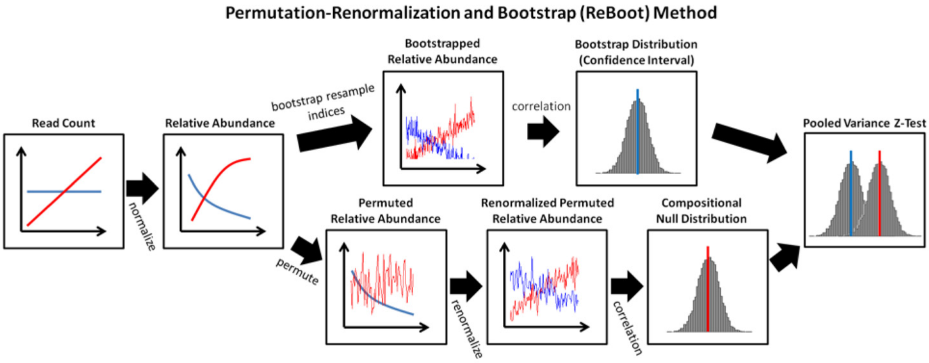

Permutation-Renormalization and Bootstrap (ReBoot) Method
Read Count
normalize
Relative Abundance
bootstrap resample indices
Bootstrapped Relative Abundance
correlation
Bootstrap Distribution (Confidence Interval)
Pooled Variance Z-Test
permute
Permuted Relative Abundance
renormalize
Renormalized Permuted Relative Abundance
correlation
Compositional Null Distribution

Figure 2.2: Graphical overview of the ReBoot method. The raw read counts are normalized, yielding relative abundances. From the relative abundances and each pair of OTUs, a bootstrap distribution and a compositional null distribution are estimated. These two distributions are compared and tested for significance. The figure is taken from [94] under license CC-BY-4.0.

By applying a threshold on the q-values to consider or specifying the number of links to include, a network can be constructed. Because all the OTUs are compared symmetrically, this network is undirected, meaning that the method has the downside of not being able to represent asymmetric interactions such as amenalism and parasitism. We refer to a link in such a network as an *inferred interaction*. The original idea was that these inferred interactions should represent the ecological interaction between the microbes. However, the presence of an inferred interaction does not automatically imply that a real ecological interaction is present[107] and the results must therefore we interpreted with caution.

When creating a ReBoot data analysis pipeline there are many different options available which may influence the results:

- There exist many similarity measures such as Pearson correlation, Spearman correlation, Kendall correlation, cosine similarity, and Bray-Curtis similarity. Unlike the original paper introducing the ReBoot method[94], we consider the results for each similarity measure independently instead of merging the networks for different similarity measures.

- Rare OTUs should be filtered out prior to analysis [108], the reason being that rare OTUs may cause noise, spurious correlations and reduced statistical power. Hence, how to set the threshold for filtering the rare OTUs is an important choice to make.

- OTU tables obtained from 16S sequencing do by default only provide estimates of the *relative* abundance of an OTU within a sample. This means that it is harder to accurately compare OTU abundances across samples[95]. For datasets where estimations of total cell count is available through qPCR or flow cytometry, absolute abundances can nevertheless be estimated from the relative abundances. This is the case for both the selection-switch dataset and the Tara dataset, giving us the opportunity to compare analysis of relative and absolute abundances. When handling datasets with estimated absolute abundances, renormalization of reads is counter-productive, so it should be turned off in these cases.
- Random noise can be added to the abundances during bootstrapping in order to distort patterns of shared zeros and limited resolution in the data. Especially the non-parametric similarity measures such as the Spearman correlation suffer from the fact that common zeros lead to a high degree of similarity for OTUs present in a few samples only. In our case, the noise is normally distributed with a standard deviation proportional to the resolution of the sequencing data.

The ReBoot implementation used in this thesis is based on the `ccrepe` `R`[109] package[110]. However, in order to enhance performance and pipeline the aforementioned variations of the ReBoot algorithm, we adapted the original code into a package named `micInt` (`https://github.com/AlmaasLab/micInt`). A graphical summary of the pipeline employed in Paper I (as well as for the Tara dataset) is shown in Figure 2.3.

2.4.4 Similarity measures

For the ReBoot algorithm the choice of the similarity measure is important. Mathematically speaking, a similarity measure is a function which reports how similar two vectors of observations are[111]. In our research, we tried out a variety of similarity measures, but focused on the Pearson product-moment correlation coefficient[112] and the Spearman rank coefficient[92] as they are commonly used, and yet have different properties. The Pearson correlation determines linear relationships and reports perfect similarity ($r = 1$) if the two abundance vectors follow a straight line with a positive slope and perfect dissimilarity ($r = -1$) if the straight line has a negative slope. On the other hand, the Spearman correlation is a non-parametric measure and is defined as the Pearson correlation of the ranks of the observations in the vector. This means that all data which form a monotonic increasing (but not necessarily linear) relationship will be reported as completely similar, whereas data which form a monotonically decreasing relationship will be reported as completely dissimilar[113].

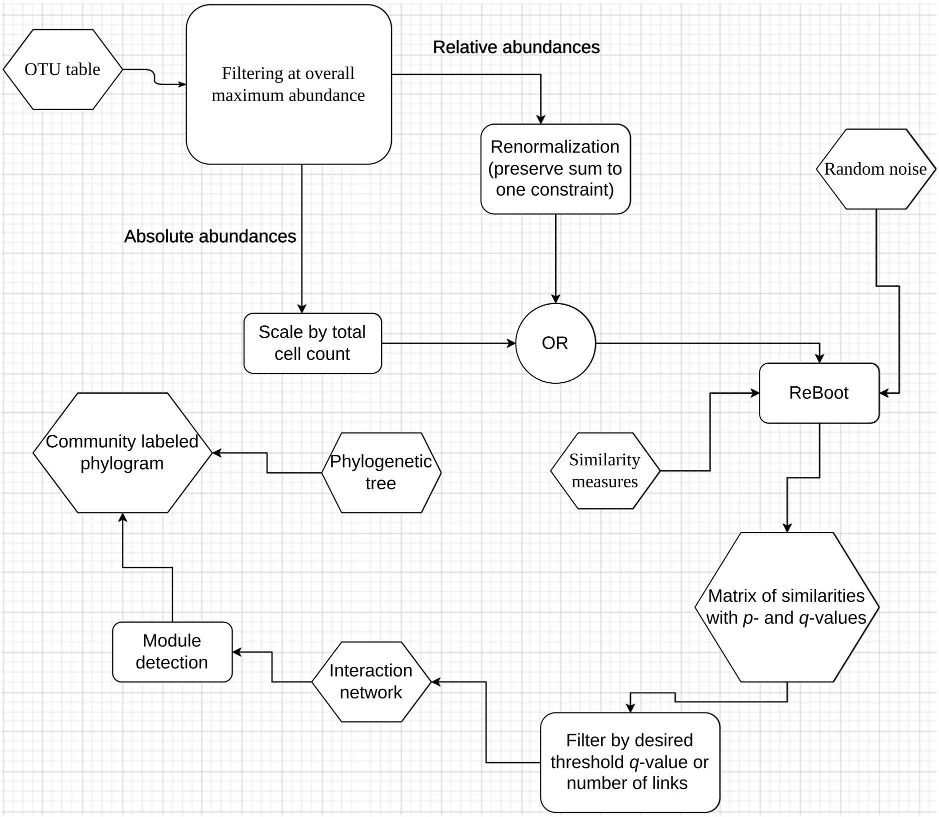

OTU table
Filtering at overall maximum abundance
Relative abundances
Renormalization (preserve sum to one constraint)
Random noise
Absolute abundances
Scale by total cell count
OR
ReBoot
Community labeled phylogram
Phylogenetic tree
Similarity measures
Matrix of similarities with *p*- and *q*-values
Module detection
Interaction network
Filter by desired threshold *q*-value or number of links

Figure 2.3: Flowchart of the data processing pipeline for Paper I. The OTU table is first filtered in order to remove rare OTUs. Subsequently, the table is preprocessed either by renormalization or scaling by total cell counts. This preprocessed OTU table is fed to ReBoot together with similarity measures and generators of random noise. The resulting matrix of q-values is filtered, the involved OTUs are presented in a network and module detection is run. The module assignments are then used to label the OTUs in the phylogenetic tree. Figure adapted from [111] with permission.

2.5 How to interpret microbial interaction networks

2.5.1 Network modules

In a network, a collection of nodes which is more interconnected with itself than the rest of the network, is called a *community* or *module*. We will refer to this concept as network module as not to confuse it with the term *microbial community*. The nodes in a module and the links between them constitute a subgraph, meaning that it is a network by itself if considered in isolation. There exists various module concepts and definitions of what a module is[114, 20]. The main idea is that a tightly interconnected part of a network might have something in common and therefore is worth studying. For a social network this can correspond to a workplace or a school class and in a scientific co-publishing network, this may correspond to a field of study such as mathematics or chemistry.

We will focus on the hierarchical network module concept where the network is partitioned into multiple layers of modules[20]. On one extreme, the entire network itself is a module, and on the other extreme, the individual nodes themselves are modules. However, we are generally interested in dividing the network into modules which reside between these extremes. This requires us to decide on a criterion determining which partitioning is the best one.

2.5.2 The walktrap algorithm

Just as there are various definitions of what a network module is, there are also a wide range of tools for finding modules[20]. The main classes of methods for hierarchical module detection are divisive and agglomerative algorithms where the former infer modules top-down and the latter infer the modules bottom-up. Divisive algorithms start with the entire network and recursively split the network into smaller and smaller pieces, where some kind of optimality criterion is satisfied at each step. By contrast, agglomerative algorithms merges nodes and smaller modules together until a certain optimality threshold is reached.

The *walktrap algorithm*[115] is an agglomerative method where short random walks among the edges of the network are used to estimate the probability that the walk stays inside a module. Probabilities P_{ij} (probability of ending up at node j given start at node i) are first estimated. These probabilities serve as a starting point for clustering the nodes in a hierarchical tree by Ward clustering[116]. Where to cut the tree and thus obtain a module decomposition is decided by the *modularity score* of each of the possible cuts. The cut having the largest modularity score is chosen as the final module decomposition. As the estimation of the transition probabilities P_{ij} are estimated by a Monte-Carlo method, the results may vary depending on random seed, but with a sufficiently high number of random walks, the stochastic effects should be minimal.

2.5.3 Network modules for ReBoot networks

While the most intuitive approach is to assume that each link in a ReBoot network corresponds to some direct ecological between two bacteria, the results tell a different story. As Paper I shows, we end up with networks consisting of mostly positive inferred interactions where there are tightly clustered modules of related bacteria. Instead of telling us about the ecological interactions, the resulting networks are more indicative of habitat preferences, an observation reported in other studies[117, 118]. This effect is robust to the level of applied noise, similarity measure and the rate of filtering rare OTUs. In addition, using other and more sophisticated tools than ReBoot may not help the problem either as inferring ecological interaction from co-occurence data has been shown to be difficult despite efforts to make the inference algorithms more robust[119, 120].

Results from the Tara dataset tell a similar story even though the dataset is very different in nature. Another finding evident in the Tara dataset is that due to a large number of samples and OTUs, there is a cognitive overload of interactions having low q-values. Depending on the settings, it is easy to get more than 1 000 000 interactions having a q-value of less than 0.05. Therefore, it is more sensible to pick a certain number of edges for network visualization than setting a specific q-value threshold.

When focusing on the network modules, we also see tendencies of phylogentic clustering (Figure 2.4 and 2.5). The main distinction is between cluster 3 and 6 which mostly contain *Proteobacteria* and *Cyanobacteria*, respectively. In module 3, the SAR11 clade including *Candidatus Pelagibacter* dominates, whereas *Prochlorococcus* and *Synechococcus* are the dominant genera in module 6. Ecologically, the SAR11 clade is the most dominant order of heterotrophic microorganisms in the oceans and have widespread geographical distribution[121, 122]. On the other hand, *Prochlorococcus* and *Synechococcus* are the oceans' most abundant primary producers[123]. Hence, the most significant interactions seem to cluster the most abundant groups of organisms together. The observation could make sense in the respect that phototropes are found in the same samples when sunlight is abundant, whereas heterotrophes are found together when organic carbon and other nutrients are available.

Still, our findings defy the competitive exclusion principle stating that if two species compete for the same resources, one of the species will out-compete the other[124]. The fact that far more planktonic species co-exist in natural environment than there are resources available, is referred to as the ”paradox of the plankton”[125] and numerous theories and models have been suggested in order to explain this discrepancy[126]. One of the popular theories in this field is Hubbell's neutrality theory which states that different competitors are almost ecologically equivalent and therefore selection is not predominant enough to ensure co-exclusion[127, 128].

During our research, we noticed that a relatively high number of samples are required in order to obtain sufficient statistical power for network construction. This was not a problem for the selection-switch and Tara datasets which contained more than a hundred samples each. However, the carbon-cycle dataset having only twelve samples proved to be insufficient for any meaningful interaction network construction.

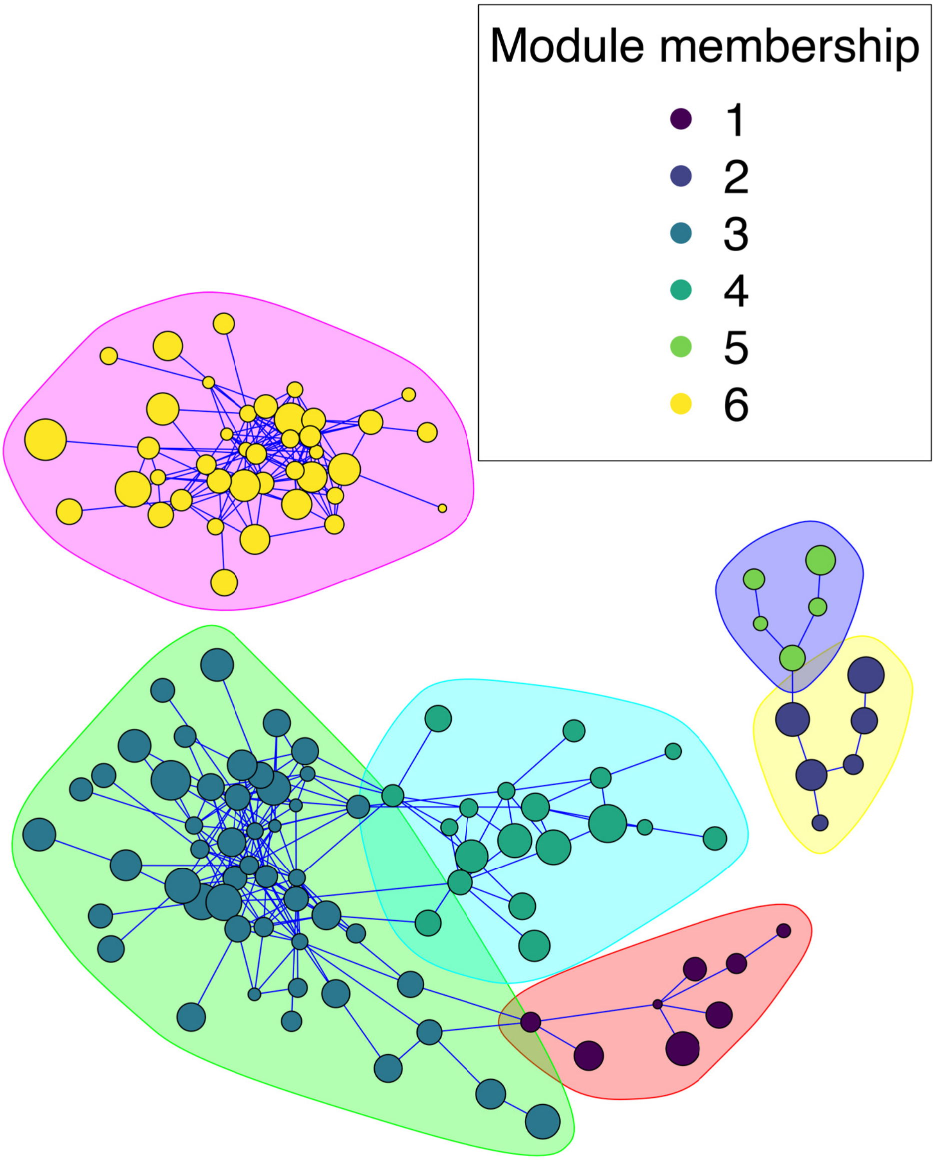


Figure 2.4: Module labeled network (computed by the Walktrap algorithm with 20 steps) of the 400 most significant interactions in the Tara dataset with Spearman similarity, absolute abundances, low noise level and low filtering threshold. Blue edges indicate positive interactions, whereas red edges indicate negative interactions. The node-sizes are scaled logarithmically according to overall mean abundance.

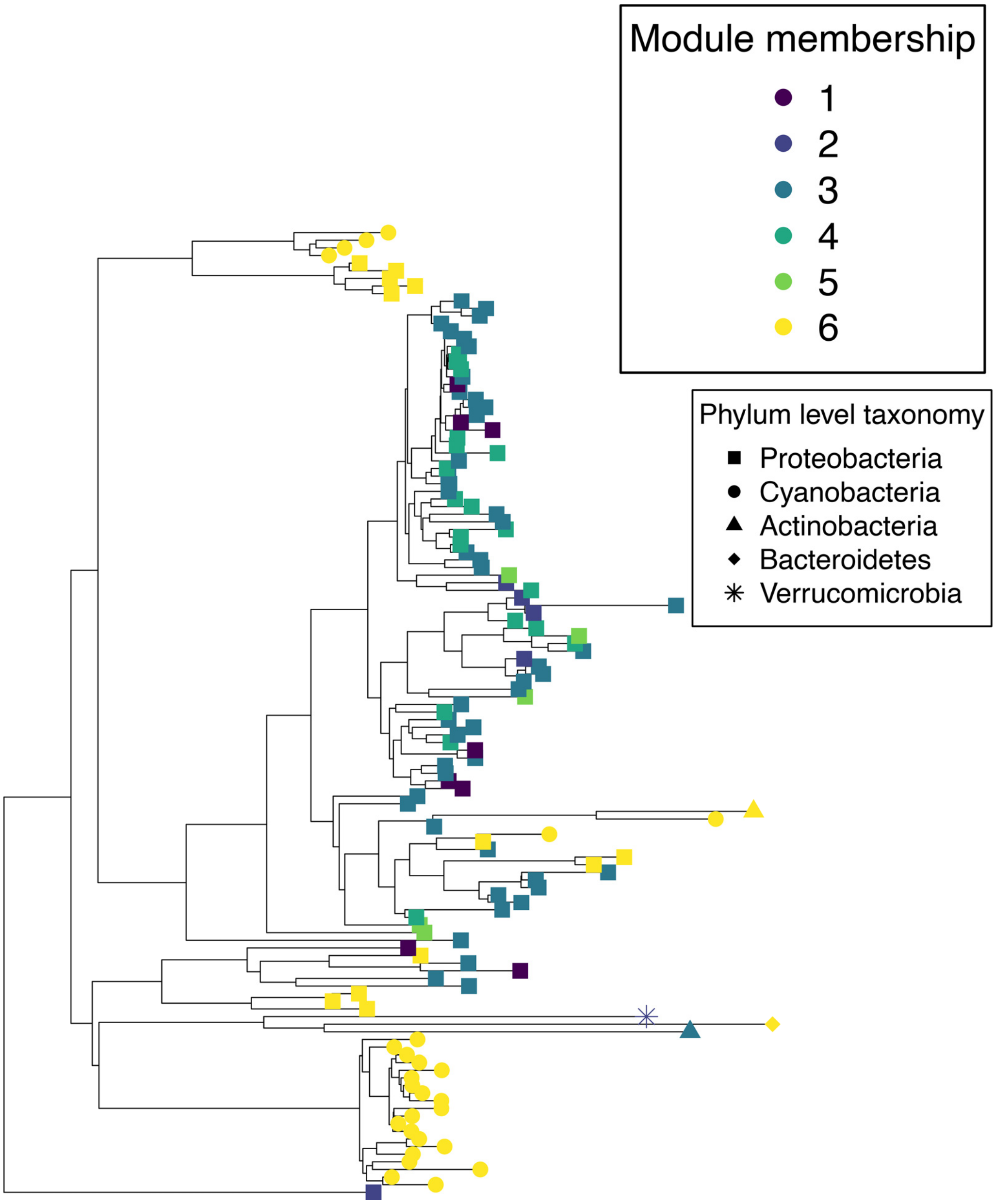


Figure 2.5: The OTUs shown as nodes in figure 2.4 in a phylogentic tree together with the class level taxonomical assignment. Notice that there is some inconsistencies between the phylogentic tree and the assigned taxonomy.

Chapter 3

Network inference for gene co-expression

3.1 What is gene expression?

A gene in a living organism is understood as an inherited construct with a specific function[129]. For our discussion, we will limit ourselves to protein coding genes, this is coding regions in the DNA which are translated into a functional protein[130]. The genome encoding the genes is generally static over the entire lifespan of the organism and cells of multicellular organisms usually contain the same genome. However, organisms respond to their environment and exhibit cell differentiation by altering the *gene expression* which is a dynamic entity. A gene is expressed if it is actively transcribed and translated from the genome. Hence, the gene expression is the key determinant differentiating nerve cells and muscle cells[131]. Also, gene expression reflects the state of the human body because different situations demand different responses. For this reason, gene expression of a person recently having eaten a meal is different from that of a person how is starving[132].

Genetic background, intrinsic factors such as age and environmental disturbances all affect the gene expression. In certain cases, these factor lead to malfunctioning in regulation of gene expression, and diseases can arise as a consequence. In many cases, such as in cancer, allergy, type 2 diabetes, and chronic fatigue there is not a single gene responsible for the development of the disease, but still dysfunctional regulation of a gene compared to other genes can lead to disease progression[133]. Thus, obtaining insight into the orchestration of the gene expression, referred to as gene co-expression can shed light on the causes and progression of diseases[134].

Hopefully, this will help implement measures and treatments in order to prevent or treat the diseases in question[135].

3.2 How is gene expression measured?

According to the central dogma of molecular biology[136, 137], the information flow in a living organism starts from the DNA comprising the genome, to messenger RNA (mRNA) relaying the genetic information, and finally to protein which carries out the function encoded in the genome. Hence, gene expression studies target either the proteins directly or the mRNA containing the transcripts for building the proteins. As with microbial co-occurence networks, high-throughput modern omics technologies have enabled large-scale analyses of gene expression.

For detecting mRNA transcripts, the mRNA from the samples has typically been isolated, converted to cDNA through reverse transcriptase and analyzed on a DNA microarray[29, 30]. A DNA microarray is a chip on which fluorescent gene probes are attached. When a sample is presented on the chip surface, cDNA matching the probes hybridizes and changes the level of fluorescence. This effect is then recorded for each probe on the chip on an automated scanner, thus reporting the level of expression for each gene. In the recent years, the cDNA has been sequenced directly through high-throughput sequencing in a technique called RNAseq[31, 32]. This has largely replaced DNA microarray as it is more sensitive and provides more information of gene expression.

For proteomics analyses, approaches based on mass spectrometry (MS) have become the most popular for large scale analyzes[138, 139]. Here, proteins are usually fractionated and partially degraded to peptides before sent into the mass spectrometry apparatus. Here, the peptides are exposed to an ion source which adds electric charge to the peptide, turning them into ions. The fragment ions are then deflected by a magnetic field before the fragment hits a detector. This allows the mass spectrometer to determine the mass to charge (m/z) ratio of the peptide. The resulting data are assembled and mapped back to the proteins encoded in the genome, providing an estimate of the protein composition of the sample.

3.3 What is a differential gene co-expression network?

Once the gene expression data are acquired, the data must be presented such that researchers can make sense of it. Organizing the data into gene co-expression networks (GCNs) is by far the most common approach in this regard. In a GCN, the genes are nodes and the links represent co-expression between the genes. Typically, the expression values are evaluated by similarity measures and each pairs of genes having a similarity or dissimilarity above a certain threshold is connected by a link, see Figure 3.1.

Further analyses of the network are usually necessary for drawing conclusions. This includes identification of network modules, finding hub genes and enrichment analysis. The `R` package `WGCNA` (Weighted Correlation Network Analysis)[140] is currently the most popular computational tool for such analysis and support creating a co-expression network in addition to module reporting.

Differential gene co-expression is an extension of gene co-expression by comparing gene co-expression across two or more different conditions[34, 141]. This extension is supposed to capture regulatory patterns which are difficult to capture by simple co-expression network analysis alone. This is because knowing how gene regulation in a healthy individual works helps us to differentiate unhealthy gene regulation from healthy gene regulation. In a typical setting, two datasets are analyzed, one with sick patients and one with healthy controls. With the underlying idea being that dysfunctional regulation of the genes contribute to disease development, the co-expression networks for each of the conditions are compared and used to generate a differential co-expression network. This differential GCN is then further analyzed for features which may explain the different phenotypes between the sick patients and the healthy controls.

Currently, there exists several different computational tools for differential gene co-expression analysis. One key characteristic of these algorithms is what metrics they use to infer differentially expressed genes. These include[135]:

- Changes in correlation coefficient across conditions
- Changes in entropy across conditions
- Expected conditional F-statistic
- Interaction test

As a result, the tools vary considerably in what kind of output they provide. Some tools report rankings of genes which are differentially co-expressed, whereas other give the full network of differentially expressed genes. Also, the tools differ in respect to how they treat signed changes of co-expression, meaning that the two genes are co-expressed in both conditions, but the sign of co-expression differs. Some of the tools treat such sign changes the same as loss of co-expression, whereas other tools identify this as conserved co-expression.

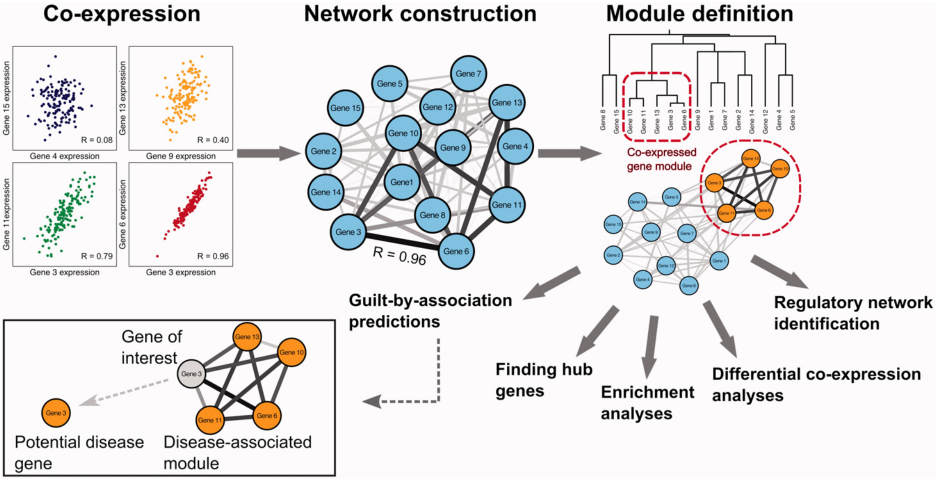


Figure 3.1: Example of a gene co-expression workflow. The co-expression of each gene pair is analyzed and the relationships being stronger than a certain threshold are then included as links in the co-expression network. From this network, genes of interest can be isolated through guild-by-association (i.e. a gene is considered important if it is connected with certain other genes). Furthermore, modules can be detected, allowing for finding sets of tightly co-expressed genes. Other types of analyses include finding hub (highly connected) genes, enrichment analysis (finding phenotypes or biological processes which occur more frequently than expected by chance), prediction of the regulatory network behind the gene co-expression, and differential co-expression analysis. Figure taken from [135] under license CC-BY-4.0.

3.4 The CSD algorithm

The CSD algorithm[35] (Conserved, Specific, and Differentiated) is a method for differential gene co-expression analysis and quantifies differences in gene co-expression among two different conditions. It is designed to discriminate between conserved co-expression (Conserved), loss of gene expression (Specific) and change of sign of co-expression (Differentiated). The procedure is as follows:

1. Treat the two datasets independently and calculate the Spearman correlation ρ[92] between the expression levels of every pair of genes. This process is repeated by resampling in order to estimate the standard error σ for the correlation.

2. The two datasets are combined in order to produce Conserved (C), Specific (S) and Differentiated (D) scores for each gene pair. Given the gene co-expression ρ_1 and ρ_2 in the two conditions and the associated estimation of standard deviation σ_1 and σ_2, these values are given by:

$$C = \frac{|\rho_1 + \rho_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \tag{3.1}$$

$$D = \frac{||\rho_1| + |\rho_2| - |\rho_1 + \rho_2||}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \tag{3.2}$$

$$S = \frac{||\rho_1| - |\rho_2||}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{3.3}$$

   The $C-$ value indicates to what extent two genes have the same co-expression across the two conditions, the $D-$ value indicates whether the two genes are co-expressed with opposite signs under the two conditions, whereas the $S-$ value indicates whether the genes are co-expressed in one condition, but not in the other. This logic is illustrated in Figure 3.2. The result of this stage is a table where each gene pair is an entry and contains the values of ρ_1, ρ_2, σ_1, σ_2, C, S and D.

3. The gene pairs with the highest values (according to some thereshold) for C, S and D are selected and tables containing the gene pairs with the highest values for C, S and D are produced. These tables are interpreted as networks where the nodes are genes and the links are the gene pairs in these tables. External network analysis and visualization tools are thereafter used to interpret the networks and provide biological knowledge.

Earlier, there existed two implementations of CSD. This includes the original implementation introducing the method[35] (`https://github.com/andre-voigt/`

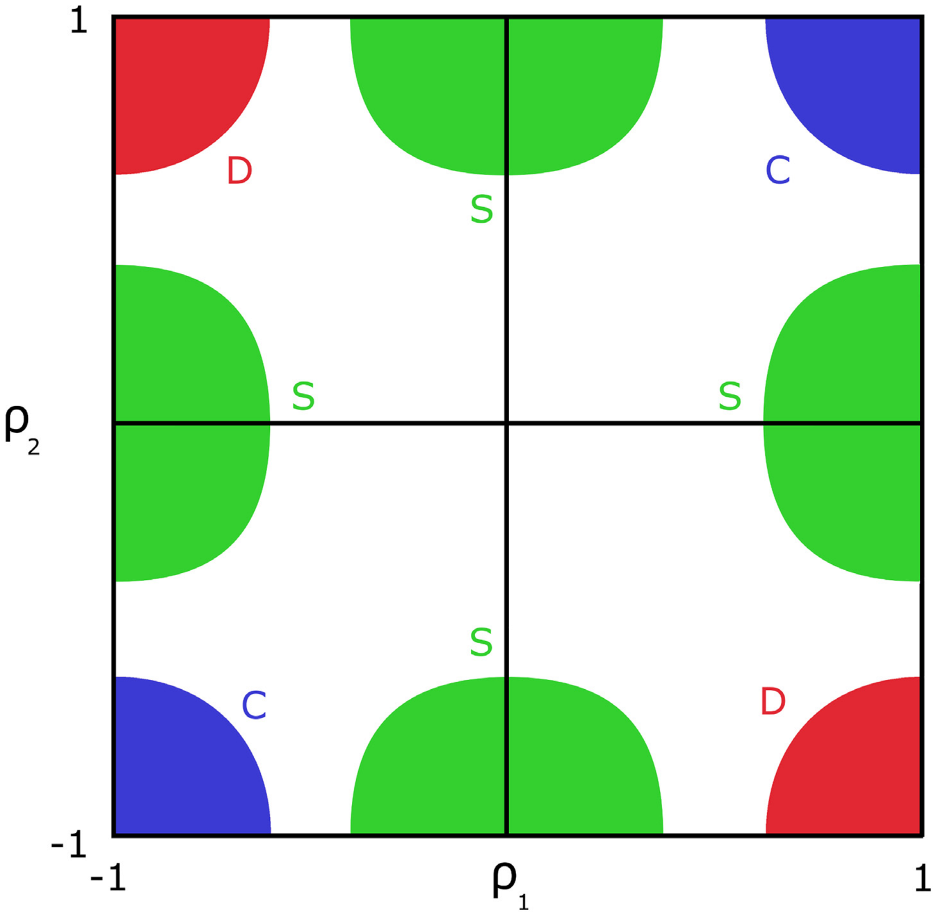


Figure 3.2: The scores Conserved (C), Specific (S) and Differentiated (D) for a gene pair depend on the gene co-expression values ρ_1 and ρ_2 under the two conditions. A high or low value for both ρ_1 and ρ_2 result in a high C-score (upper right and lower left in the figure). A high value for either ρ_1 or ρ_2 and a corresponding low value for the other co-expression value result in a high D-score. Finally, a high or low value for one of ρ_1 or ρ_2 and a close to zero value for the other co-expression value result in a high S-score (middle of the top, left, right and bottom faces). Figure taken from [35] under license CC-BY-4.0.

CSD). It is provided as a collection of three programs each carrying out one of the steps of the CSD procedure and are expected to be used in tandem. The program for part 1 is written in C++, whereas the two others programs are written in Python. The other implementation, named CSD C++ (https://githuxb.com/magnusolavhelland/CSD-Software) is the result of a Master's thesis[142] and is written in C++, aiming for maximal performance.

3.5 The csdR package

Paper II introduces csdR which is an R package written with the performance, interoperability and user-friendliness in mind. In addition to R, the C++ interface Rcpp[143, 144, 145] and the parallelism interface openMP[146] is used for writing performance critical code. The package is available on Bioconductor (https://doi.org/doi:10.18129/B9.bioc.csdR) which is a comprehensive repository of R packages. Residing on Bioconductor makes the package easy to find and install[147].

Both of the older CSD implementations have hard-coded the procedure for file operations. Besides adding complexity to the codebase, coding file operations directly into the source code opens a larger potential for errors due to unexpected input formats. For CSD C++, the vagueties about the accepted file format caused serious file reading problems to the extent that we gave up benchmarking the implementation for Paper II. Fortunately, this did not make cause any severe problems for the the orignal CSD implementation.

The original CSD implementation passes the results from one step in the pipeline to the next step through files. The size of the intermediate files in a typical CSD analysis is often on the magnitude of tens of gigabytes and much larger than the input and the end result. This consumes large amounts of disk space and file operations slows down the process. csdR avoids these problems by not including any file operation code. All computations are done on R datastructures which gives larger flexibility in import of data and downstream handling of the results. Typically, the user will offload reading and writing of data to preexisting R procedures for data import and export.

In terms of computational performance, the most demanding part is first step of the algorithm where the co-expression of the gene pairs are computed with resampling. If there was no need for calculating the standard deviation of the co-expression, this step could be skipped, saving the bulk of the computational burden. Some studies have suggested that not normalizing the C-, S- and D-values by the standard deviations does not alter the results to a large degree[148]. Also, calculating standard deviations is omitted by CoDiNA, a multi-condition variety of CSD[149]. However, for csdR we chose to retain the conservative approach of estimating standard errors.

In csdR, the resampling is done through bootstrapping of the entire dataset[105] compared to a custom subsampling algorithm on the individual gene pairs in the two other implementations. The way this resampling is implemented for csdR saves large amounts of computational overhead compared to the original CSD implementation. Even though csdR runs the bootstrap iterations sequentially due to constraints on memory, it can still leverage parallelism as the Spearman correlation is calculated by a heavily optimized and multithreaded implementation.

Unfortunately, Paper II shows that csdR fails to achieve a parallel speedup higher than a factor 2.72, no matter how many CPU cores are available. This demonstrates that there must be another bottleneck in the system. Given the fact that the CPU has to load and store gigabytes of data for each bootstrap iteration, we believe that this bottleneck is the bandwidth to system memory. Still, the performance of csdR is suffcient to conduct realistic analyses within a day compared to the original implementation where weeks could pass before the results were ready.

Chapter 4

Genome-scale metabolic models

4.1 What is a genome-scale metabolic model?

The collection of all chemical conversions taking place inside a living organism is called the metabolism. As the goal of bioengineering often is to make the organisms produce or consume a compound of interest or carry out a specific process, the metabolism of is usually the most important focus in bioengineering. The biochemical reactions themselves are difficult to measure directly as they happen on a microscopic scale and cannot be isolated from the environment in which they occur. Still, the metabolic compounds inside an organism and its surroundings can be isolated and detected[150], and thus shed light on the reactions taking place inside the cell. The process of creating an *in silico* model of the metabolism occurring inside a cell is referred to as *metabolic reconstruction*[151].

Most biological reactions are catalyzed by enzymes. These enzymes can too be isolated and studied for their catalytic properties[152]. Furthermore, the enzymes are encoded in the organism's genome, so the enzymes and hence the organism's metabolic capabilities can be predicted by the genome sequence. As a consequence, the common starting point for metabolic reconstruction is to sequence the genome of the organism. The genome is then annotated for the enzymes it encodes. Gene-reaction rules are then used to associate the enzymes with the metabolic reactions they encode. The resulting listing of the organism's metabolic capabilities is referred to as a genome-scale metabolic model (GEM)[36, 151]. In addition, a good metabolic model should also include non-enzymatic reactions such as diffusion and non-catalyzed reactions.

For modeling purposes, GEMs often include pseudo-reactions which are not real

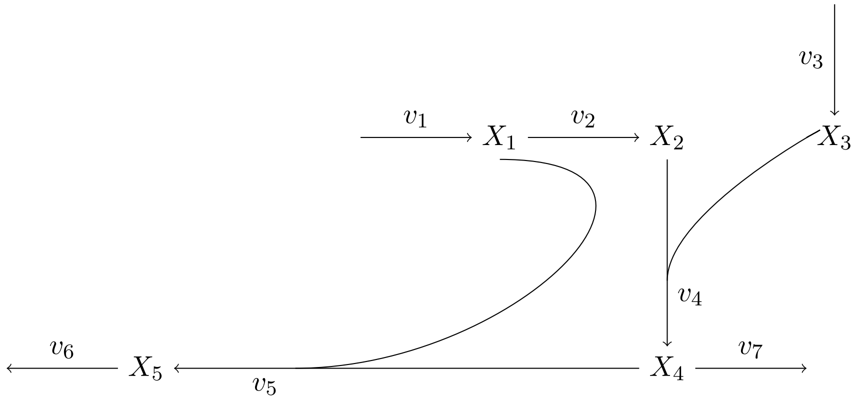


Figure 4.1: A small example of a metabolic network. X_1 through X_5 denotes metabolites, whereas the arrows denote reactions with their corresponding fluxes v_1 through v_7.

metabolic reactions per se, but are still included in order to account for biological effects. Some of the most common applications of such pseudo-reactions relate to biomass production (growth), enzyme pools (for enzyme-constrained models), and non-growth associated maintenance (NGAM)[153, 154]. NGAM simulates the ATP demand for the cell in order to maintain homeostasis and is thus implemented as an ATP hydrolysis function through which a minimum flux is required. We will discuss production of biomass and enzyme pools later in this chapter.

A toy metabolic network is shown in Figure 4.1. This network contains five metabolites and seven reactions and has the reaction representation:

$$R_1: \quad \xrightarrow{v_1} X_1 \tag{4.1}$$
$$R_2: \quad X_1 \xrightarrow{v_2} X_2 \tag{4.2}$$
$$R_3: \quad \xrightarrow{v_3} X_3 \tag{4.3}$$
$$R_4: \quad X_2 + X_3 \xrightarrow{v_4} X_4 \tag{4.4}$$
$$R_5: \quad X_1 + X_4 \xrightarrow{v_5} X_5 \tag{4.5}$$
$$R_6: \quad X_5 \xrightarrow{v_6} \tag{4.6}$$
$$R_7: \quad X_4 \xrightarrow{v_7} \tag{4.7}$$

We note that reactions R_1 and R_3 produce metabolites without requiring participation of other metabolites. These are called *import reactions*. Conversely, reactions R_6 and R_7 consume metabolites without producing any metabolites and are hence called *export reactions*. Collectively, import and export reactions are called *exchange reactions*. $v_1, \ldots, v_7$ denominates how fast the reactions run (usually reported in milimoles per gram dry weight per second) and are called the *fluxes* of the reactions. The stochiometric matrix of the model is a mathematical representation of the network where the metabolites constitute the rows and the columns are the reactions. Entry (i, j) in the stochiometric matrix S gives the coefficient of metabolite i in reaction j. By convention $S_{i,j}$ is positive if the metabolite is produced in the reaction and negative if the metabolite is consumed in the reaction. In our example, the stochiometric matrix is given by:

$$S = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{array} \begin{array}{c} \begin{array}{ccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 \end{array} \\ \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix} \end{array} \tag{4.8}$$

4.2 Flux Balance Analysis (FBA)

Once the network of reactions is known, the next problem is to use it to predict and explain metabolic behavior. The most direct approach for modeling the metabolism is having a mechanistic differential equation based model of the reactions and the metabolites[155]. However, this approach requires large amounts of knowledge into the kinetics of the system and how the processes are regulated. For this reason, mechanistic models are more common for small, well-studied metabolic and signaling pathways, whereas creating genome-scale mechanistic models of entire organisms is usually infeasible[156]. Constraint-based methods provide viable alternatives to mechanistic models for larger models as they usually require less data[157]. Of the constraint-based methods, Flux Balance Analysis (FBA) [158, 159, 160, 157] is one of the most popular ones. It relies on two basic assumptions:

1. *(Quasi) Steady-state*: We assume that the internal concentration of metabolites does not change. This means that the consumption of a metabolite equals its consumption. Even though cells in real life *can* change internal metabolite concentrations, the steady-state assumption still is a good approximation in many situations such as under exponential growth. Note that this does *not* imply that the cell is in chemical equilibrium which means that there is no net chemical conversion and all fluxes are zero. In fact, a cell which has reached chemical equilibrium is dead[161].

2. *Optimal regulation*: Even though the details of internal regulation of metabolism is unknown, we nevertheless assume that evolution has shaped the regulation to maximize a certain objective linked to the organism's survival and proliferation. Hence, we talk about the objective function, the function of fluxes to maximize given the constraints of the model are respected. The most popular approach for microorganisms is to assume that the objective for the organism is to grow as quickly as possible and thus maximize production of biomass. The biomass reaction is a pseudo-reactions which purpose is to simulate the consumption of metabolites, co-factors and energy currency molecules in order to make more cell material.

4.2.1 Mathematical formulation

Using the notation from the toy metabolic network, we can describe the FBA framework mathematically:

For all i, we have

$$\frac{\mathbf{d}\,X_i}{\mathbf{d}\,t} = \sum_j S_{i,j} v_j, \tag{4.9}$$

or on a vectorized form:

$$\frac{\mathbf{d}\,\mathbf{X}}{\mathbf{d}\,t} = S\mathbf{v}, \tag{4.10}$$

The steady state assumption of FBA requires that

$$\frac{\mathbf{d}\,X_i}{\mathbf{d}\,t} = 0 \tag{4.11}$$

for all i or equivalently

$$\frac{\mathbf{d}\,\mathbf{X}}{\mathbf{d}\,t} = \mathbf{0} \Leftrightarrow S\mathbf{v} = \mathbf{0}. \tag{4.12}$$

The objective function is given by a linear combination $\mathbf{c}$ of the fluxes[1]:

[1] Non-linear objective functions are in theory possible to define, but are rare in practice due to computational issues

$$Z = \mathbf{c}^{\mathsf{T}}\mathbf{v} \tag{4.13}$$

Also, we set bounds on the reactions to limit the fluxes, such that for each reaction j, there exist flux bounds v_j^{lb} and v_j^{ub} such that:

$$v_j^{lb} \leq v_j \leq v_j^{ub}. \tag{4.14}$$

In the default FBA setting, only the import reactions are constrained to a certain number, the reason being is that uptake rates are relatively easy to measure experimentally, whereas the fluxes for internal reactions are not. Otherwise, the flux bounds are most commonly set to unlimited or high numbers which for all practical purposes gives an unlimited flux. The exception of this rule is that many reactions are considered irreversible and only run in a certain direction, hence their fluxes are only allowed to take on positive or negative values.

The entire FBA problem can then be formulated as a linear program (LP):

$$\begin{array}{ll} \text{Maximize} & Z = \mathbf{c}^{\mathsf{T}}\mathbf{v} \\ \text{Subject to:} & S\mathbf{v} = \mathbf{0} \\ & \mathbf{v}^{lb} \leq \mathbf{v} \leq \mathbf{v}^{ub} \end{array} \tag{4.15}$$

4.2.2 Solution space of FBA

The set of flux vectors **v** which satisfy the equality and inequality constraints in equation 4.15 is called the *solution space* of the problem. Geometrically, the solution space is a convex polytype in a high-dimensional room. If the solution space is empty, i.e. no **v** exists such that the constraints can be satisfied, we refer to the problem as infeasible and no solution can be obtained.

The simplex algorithm is the classical method for solving LP problems such as the one in equation 4.15. Variations and enhancements of the simplex algorithm are used in efficient LP solvers such as CPLEX or Gurobi in order to obtain an ideal solution to the problem. Usually, the FBA problem possesses degeneracy, meaning that there is more than one point in the solution space yielding the maximal objective value. Still, LP solvers usually only report one of these optimal solutions[162].

4.2.3 Flux Variability Analysis (FVA)

Sometimes it is not desirable to only obtain one of many possible optimal solutions from a degenerate FBA problem. This is often the case with dynamic FBA simulations (Paper IV) which are sensitive to which solutions are chosen from degenerate problems. There are two main strategies for dealing with the challenge of degeneracy:

- Add secondary objectives or constraints in order to narrow down the solution space. Parsimonious flux balance analysis[163] is a prime example of this methodology which picks the optimal solution having the minimum absolute sum of reaction fluxes. In Paper IV we solve the challenge of degenerate solutions by applying multiple objectives for each of the exchange fluxes in lexicographical order[164, 165].
- Explore the diversity of optimal solutions. This includes approaches such as flux sampling[166] and flux variability analysis (FVA)[167, 168].

We will explain FVA in more depth:

In FVA, the optimal value of the objective function Z_{obj} is first found through FBA. Then, the algorithm goes through each reaction in the model individually and queries the minimum and maximum fluxes which are obtainable under the optimal FBA objective value:

$$\begin{array}{ll} \text{Maximize/Minimize} & \mathbf{v}_i \\ \text{Subject to:} & S\mathbf{v} = \mathbf{0} \\ & \mathbf{c}^\mathsf{T}\mathbf{v} = Z_{obj} \\ & \mathbf{v}^{lb} \leq \mathbf{v} \leq \mathbf{v}^{ub} \quad \text{for } i \in \{1, \dots, n\} \end{array} \tag{4.16}$$

Often in practice, the constraint $\mathbf{c}^\mathsf{T}\mathbf{v} = Z_{obj}$ is relaxed to $\mathbf{c}^\mathsf{T}\mathbf{v} \leq (1 - \epsilon)Z_{obj}$ where ϵ is a small positive number (typically 0.01). This is done in order to avoid numerical problems. Even though FVA can find the allowed range of values for each flux in the space of optimal solutions, we still obtain no information on how the fluxes are coupled. For instance, imagine reactions R_1 and R_2 have $[1, 5]$ and $[8, 10]$, respectively as their FVA ranges. Then we know there is an optimal solution where R_1 has the value 3 and there exists an optimal solution where R_2 has the value 9. However, we don't know whether there exists any optimal solution which satisfies both $R_1 = 3$ *and* $R_2 = 9$ simultaneously.

4.2.4 How FBA is used

By its nature, FBA gives a quantitative prediction of the resulting fluxes. Hence, FBA is often used in bioprocess engineering to predict the growth rate, consumption of substrate and excretion of waste products given the nutrient environment[169].

In addition to the quantitative flux predictions, FBA is often used qualitatively where the feasibility of the FBA problem is checked. Often, infeasible problems are due to errors or missing reactions in the model. As such FBA is used to validate and test genome-scale reconstructions rather than just being a result of the reconstruction process[36]. This is utilized in a reconstruction process referred to as *gap-filling* where changes (often addition of missing reactions) in the metabolic model are made to improve its quality[170].

Another other major source of infeasible models is missing nutrients in the simulated growth medium. Experimental results may determine that the organism is able to grow in some environments and not in others[171]. This can be applied in gap-filling as a model which is infeasible in an environment where it should grow, most certainly has missing reactions. Conversely, a model which is feasible in an environment where it does not grow experimentally often has reactions which should be removed from the model.

Furthermore, FBA is used to predict effects of gene knock-outs and changes to the growth medium[172, 173]. This includes attempts to couple production of a desired compound with the organism's growth[174]. However, exposing the organism to a genetic makeup or an environment to which it has not evolved to adapt, often invalidates the optimality principle of FBA. This is because the regulatory mechanisms inside the cell have not been optimized for the new environment or genetic background and may therefore not achieve optimality.

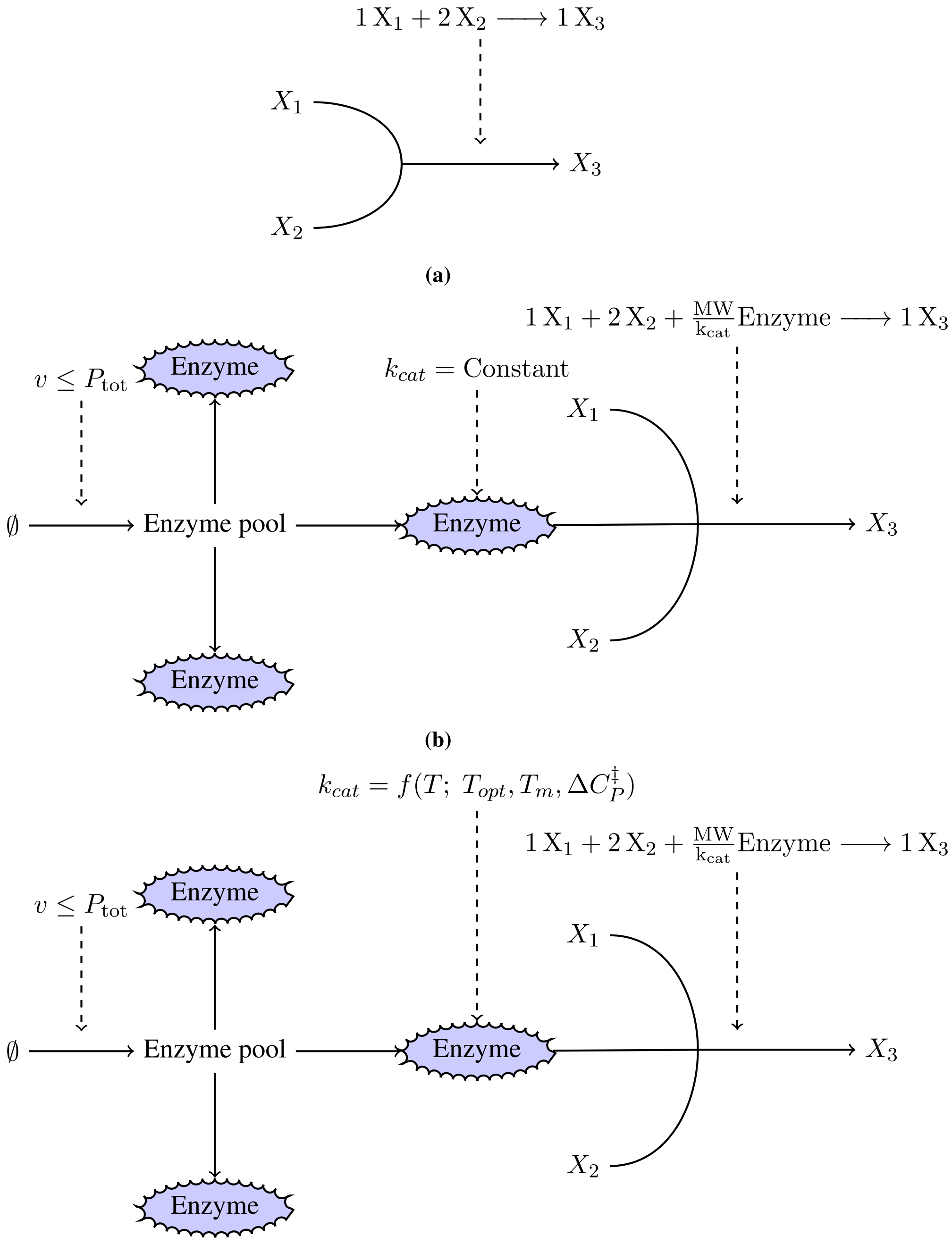
The problem of non-optimality can be mitigated by adaptive laboratory evolution (ALE) where microorganisms being exposed to genetic or environmental changes will be allowed to evolve for generations. Evolution will then usually rewire the regulation of metabolism and thus help the organism approach optimality[174, 175]. In addition, there exists heuristics designed to predict metabolic behavior immediately after environmental and genetic disturbances[176, 177, 178].

4.3 Extensions of FBA

Since its inception, a plethora of FBA varieties and extensions have been developed[40]. Although a complete listing of all varieties is out of scope, we will discuss FBA extensions following the organization in Table 4.1. Additionally, we will illustrate the difference between FBA, ecFBA and etcFBA in Figure 4.2.

4.3.1 Enzymatic constraints (ecFBA)

Baseline FBA considers only constraints related to uptake of metabolites and nutrients. This way, the cell is effectively thought of as a chemical factory which converts some input metabolites into cell material (growth) and metabolites which


$1\,\mathrm{X_1} + 2\,\mathrm{X_2} \longrightarrow 1\,\mathrm{X_3}$
X_1
X_2
X_3
$1\,\mathrm{X_1} + 2\,\mathrm{X_2} + \frac{\mathrm{MW}}{\mathrm{k_{cat}}}\mathrm{Enzyme} \longrightarrow 1\,\mathrm{X_3}$
$v \leq P_{\mathrm{tot}}$
Enzyme
$k_{cat} = \mathrm{Constant}$
$\emptyset$
Enzyme pool
Enzyme
Enzyme
X_1
X_2
X_3
$k_{cat} = f(T;\ T_{opt}, T_m, \Delta C_P^{\ddagger})$
$1\,\mathrm{X_1} + 2\,\mathrm{X_2} + \frac{\mathrm{MW}}{\mathrm{k_{cat}}}\mathrm{Enzyme} \longrightarrow 1\,\mathrm{X_3}$
$v \leq P_{\mathrm{tot}}$
Enzyme
$\emptyset$
Enzyme pool
Enzyme
Enzyme
X_1
X_2
X_3

(a)

(b)

(c)

Figure 4.2: Baseline FBA, ecFBA and etcFBA are incrementally more complex model frameworks. Baseline FBA (panel **(a)**) only involves the stoichiometry of the reactions and enforces a steady-state assumption. ecFBA (panel **(b)**) adds the enzyme as a pseudo-reactant. The enzyme has a fixed catalytic rate and is drawn from a finite enzyme pool which is shared with the other enzymes in the model. etcFBA (panel **(c)**), adds temperature dependence of the enzymes through a thermodynamic equation.

Table 4.1: Classification of FBA varieties presented in this thesis. The prefix *ec* is short for Enzyme Constrained, the prefix *d* is short for Dynamic and the prefix *etc* is short for Enzyme and Temperature Constrained.

	No internal constraints	**Internal enzyme constraints**	
		Fixed catalytic rate	**Temperature-dependent catalytic rate**
No time dependence	FBA	ecFBA	etcFBA
Static time dependence	dFBA	decFBA	detcFBA

are excreted. However, metabolic enzymes also have limitations on how much flux they can catalyze. This becomes evident when considering the Crabtree effect in yeast or overflow metabolism in *Escherichia coli*. Both *Saccharomyces cerevisiae* and *E. coli* have a facultative aerobic metabolism. This means that they can use oxygen for respiration, but are able to utilize energy sources through fermentation when oxygen supply falls short. Growing on glucose, full respiration yields water and carbon dioxide as the final products and has a superior energy yield per unit of glucose than fermentation to ethanol. Still, it can be observed that even though oxygen is available, fermentation steps in and supplements respiration[179, 180]. This is referred to as the Crabtree effect or overflow metabolism and surely provides a lower energy yield for the same amount of glucose consumed, so this would be judged as a wasteful process by import-constrained FBA, appeartly violating the principle of growth rate maximization.

Among modelers, a common explanation for the observed fermentative metabolism in presence of oxygen that it is due to internal enzyme constraints. Respiration of glucose requires more enzymes than fermentation and these enzymes are costly to produce and there is a physical limit to how much enzyme the cell can accommodate. Combined with the fact that the enzymes are only able to catalyze a certain number of conversions per time unit, the resporatory enzymes constitute a bottleneck in utilizing glucose when this nutrient is in excess. The protein cost for the efficient fermentative enzymes however, is much lower. This means that even though the yield of fermentative metabolism is much lower per molecule of glucose, the energy generating flux can still be higher than for respiration, allowing for faster growth[181, 15].

For incorporating internal enzyme constraints into a GEM, two key ingredients are needed: the molecular weight of the enzymes and their catalytic rates (k_{cat}). The molecular weight can be inferred from the amino acid sequence and is therefore easy to extract from genomic data. Errors can still occur resulting from post-translational modification and the presence of catalytic enzymes comprising multiple subunits. On the other hand, determining the enzyme catalytic rates for an entire GEM can be tricky. Databases such as Sabio-RK[182] and BRENDA[183, 184] contain catalytic data derived from experiments, but do not cover all enzymes and usually only provides broad coverage for model organisms such as *E. coli* and *S. cerevisiae*. In addition, the database measurements are conducted *in vitro* with the enzyme in isolation. Given that enzymes are affected by a multitude of factors such as temperature, pH, ionic strength, other metabolites and regulatory enzymes, it is often difficult, if not impossible to reproduce the natural environment of the enzymes. Hence, database measurements of catalytic rates can be off the real biological rates by several orders of magnitude[185].

Efforts have been made in order to apply machine learning and experimental omics data for obtaining better predictions of *in vivo* catalytic rates[186, 187]. Some of these refined approaches include extensive experimental data such as proteomics and fluxomics. In a similar fashion, there exists a method which combines deep learning and a Bayesian approach to calibrate k_{cat}s with experimental data[188]. However, the Bayesian approach used in this paper is very similar to the one presented by Gang Li *et al.*[189] and studied in Paper III, so we therefore suspect that it suffers from the same instability issues as demonstrated in Paper III.

FBA with Molecular Crowding (FBAwMC)[190] was the first attempt to create an enzyme constrainted FBA model (ecFBA). Here, it is assumed that the cell only can hold a limited density of proteins in their cytoplasm and the capacity for catalytic enzymes therefore is limited.

Later approaches have instead operated on the assumption that it is the amount of *protein mass* allocated to the enzymes, not the spatial crowing which causes the limitation for catalytic conversion. The GECKO[191] approach illustrated in Figure 4.3 assumes that each enzyme is in limited supply and provides this enzyme as a pseudo-metabolite to the reactions. Note that this extension of FBA still has the same overall mathematical representation as shown in equation 4.15. This is important because it allows for using existing FBA tools and data formats without modification.

One problem with the GECKO formalism is that each enzymes has its own resource pool. This may be useful when proteomics data are available to measure the concentration of each enzyme, but often experiments are done without determining the protein composition. The MOMENT[192] method relaxes this assumption by having a single pool from which all enzymes are drawn in quantities corresponding to their molecular weights. A further development of MOMENT, sMOMENT[193] gives equivalent predictions, but uses a simpler formulation while allowing adding constraints for usage of individual enzymes. sMOMENT is based on the notion that the size of the protein pool is limited by a constant P. Given the concentration g_i of enzyme i, and its molecular weight MW_i, we obtain:

$$\sum_i g_i \cdot MW_i \leq P. \tag{4.17}$$

Taken into account that the maximal flux which can be catalyzed by enzyme i is given by $v_i \leq g_i \cdot k_{cat,i}$, we obtain:

$$\sum_i v_i \cdot \frac{MW_i}{k_{cat,i}} \leq P. \tag{4.18}$$

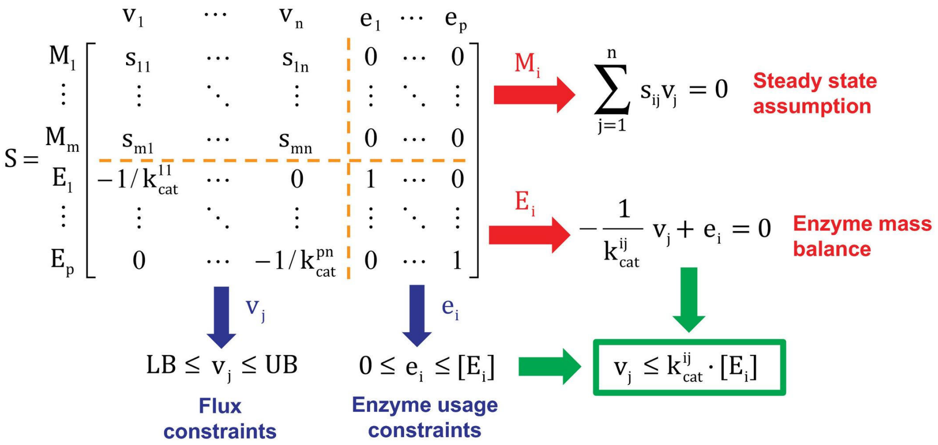
Equivalently, we can define an additonal pseudo-reaction R_{Pool} which yields:

$$-\sum_i v_i \cdot \frac{MW_i}{k_{cat,i}} + v_{Pool} = 0;\ v_{Pool} \leq P. \tag{4.19}$$

4.3.2 Time dependence (dFBA)

The standard FBA formulation only treats the cell metabolism in an instant. However, the microbes themselves change their own environment through their metabolism. This is of importance in a wine fermentation setting where the yeast consumes sugar and produces ethanol, so the behavior of the yeast will certainly change over time. Tracking these kinds of time dynamics can be done through dynamic FBA (dFBA). In addition to the wine fermentation study in Paper IV, dFBA has been applied to study diauxic growth where an organism consumes two different metabolites in sequence[194] and for guiding design of industrial biotechnological processes[195]. Two main strategies exist for creating a dynamic FBA model[196]:

- The dynamic optimization approach assuming that the organism will maximize its biomass across a certain time span. The primary shortcoming with this method is that the problem must be discretized into timesteps in advance and is computationally challenging to solve.



$$S = \begin{array}{c|ccc|ccc} & v_1 & \cdots & v_n & e_1 & \cdots & e_p \\ \hline M_1 & s_{11} & \cdots & s_{1n} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ M_m & s_{m1} & \cdots & s_{mn} & 0 & \cdots & 0 \\ \hline E_1 & -1/k_{cat}^{11} & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ E_p & 0 & \cdots & -1/k_{cat}^{pn} & 0 & \cdots & 1 \end{array}$$
M_i
$\sum_{j=1}^{n} s_{ij} v_j = 0$
Steady state assumption
E_i
$-\frac{1}{k_{cat}^{ij}} v_j + e_i = 0$
Enzyme mass balance
v_j
e_i
$LB \leq v_j \leq UB$
Flux constraints
$0 \leq e_i \leq [E_i]$
Enzyme usage constraints
$v_j \leq k_{cat}^{ij} \cdot [E_i]$

Figure 4.3: Illustration of how GECKO mathematically incorporates enzyme constraints into a FBA model. This is done by augmenting the stochiometric matrix where the enzymes are included in the matrix as well as the metabolites. The upper left quadrant of the matrix is identical to the stochiometric matrix in baseline FBA and hence the upper part of the matrix enforces the steady state assumption of the metabolites without considering the enzymes. The lower part of the augmented stochiometric matrix represents the enzyme mass constraints and ensures that the catalytic rate through a reaction never exceeds the enzyme concentration multiplied by the reaction turnover number. The figure is adapted from [191] under license CC-BY-4.0.

- The static optimization approach considering that the organism is in a quasi steady-state at all time points and biomass production is maximized given the current environment. This assumption can be motivated by the notion that the internal regulation of metabolism acts on a shorter time-scale than changes in the external environment. Computationally, this approach is more tractable as it allows the standard FBA problem in equation 4.15 to be integrated into an ODE solver for computing a solution of the problem.

We will for the remainder of the thesis focus on the static optimization approach. This can be characterized as an ODE system where the state variables are:

- $\mathbf{w}$: A vector of the external metabolite concentrations.
- X, the total biomass of the organism in the system.

The uptake constraints $\mathbf{v}^{lb}\left(\mathbf{w}(t)\right)$ and $\mathbf{v}^{ub}\left(\mathbf{w}(t)\right)$ of the FBA problem are now dependent on the microenvironment the cells can feel and hence the external metabolite concentrations, such that at any given time t, the flux vector $\mathbf{v}(t)$ must obey

$$\mathbf{v}^{lb}\left(\mathbf{w}(t)\right) \leq \mathbf{v}(t) \leq \mathbf{v}^{ub}\left(\mathbf{w}(t)\right) \tag{4.20}$$

Given these constraints, the flux vector $\mathbf{v}(t)$ is the one maximizing the growth rate:

$$\mu(t) = \mathbf{c}^{\mathsf{T}}\mathbf{v}(t) \tag{4.21}$$

In turn, the rate of change for the biomass is given by:

$$\frac{\mathrm{d}\,X}{\mathrm{d}\,t} = \mu(t) \cdot X(t) \tag{4.22}$$

and the external metabolite concentration

$$\frac{\mathrm{d}\,\mathbf{w}}{\mathrm{d}\,t} = X(t) \cdot \mathbf{g}\left(\mathbf{v}(t)\right). \tag{4.23}$$

$\mathbf{g}$ is typically a linear function of the exchange fluxes of the solution, meaning that there exist a matrix A such that

$$\mathbf{g}\left(\mathbf{v}\right) = A\mathbf{v} \tag{4.24}$$

4.3.3 Temperature dependence (etcFBA, Paper III)

Temperature is one of the most important factors to consider in enzyme kinetics. The temperature dependence of enzymes is the primary reason why food is preserved at low temperatures and high temperatures are used to kill germs. Life only thrives in a certain temperature range. If the temperature becomes too low, the enzymatic reactions will occur too slowly to be noticable, whereas high temperatures denaturate the enzymes and irreversibly destroy them. Also, an effect of temperature on heat capacity capacity of enzymes[197, 198] is believed to reduce the catalytic rate of enzymes without the effect of denaturation at sufficiently high temperatures. Li *et al.*[189] were the first to study the effect of temperature dependence on GEMs and proposed which enzymes limited the growth of the organism at high temperatures. Their approach, called enzyme and temperature constrained FBA (etcFBA), is an extension to ecFBA and assumes that the k_{cat} of an enzyme changes with temperature:

$$k_{cat}(T) \propto \frac{k_B T}{h} e^{-\frac{\Delta G^{\ddagger}(T)}{RT}}, \tag{4.25}$$

where $G^{\ddagger}(T)$ is the change of free energy from the ground state to the transition state of the reaction given by:

$$\Delta G^{\ddagger}(T) = \Delta H^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P (T - T_0) - T\left(\Delta S^{\ddagger}_{T_0} + \Delta C^{\ddagger}_P \left(\frac{T}{T_0}\right)\right), \tag{4.26}$$

where (relative to the transition state):

- $\Delta H^{\ddagger}_{T_0}$ is the change of enthalpy at a constant temperature T_0
- $\Delta C^{\ddagger}_P$ is the change in heat capacity
- $\Delta S^{\ddagger}_{T_0}$ is the change of entropy at a constant temperature T_0

Keep in mind that these parameter are *not* the same as the ΔH and ΔS of the overall reaction, which are independent of the enzyme. In addition to the temperature dependence of k_{cat}, some of the enzyme pool is modeled to be unavailable due to denaturated enzymes. The amount ($[E]_N$) of enzyme being present in the native (non-denaturated) state is given by:

$$[E]_N = \frac{1}{1 + e^{-\frac{\Delta G_u(T)}{RT}}}[E]_t, \tag{4.27}$$

where $[E]_t$ is the total amount of the enzyme and $G_u(T)$ is the free energy of denaturation given by:

$$\Delta G_u(T) = \Delta H^* + \Delta C_{p,u}(T - T_H^*) - T\Delta S^* - T\Delta C_{p,u}\left(\frac{T}{T_S^*}\right), \tag{4.28}$$

where ΔH^*, ΔS^* and $\Delta C_{p,u}$ are the enthalpy, entropy and heat capacity change of the denaturation, respectively. In addition to the enzymes, the model introduced by Li *et al.* captures the effect that heat stress increases demand for ATP maintenance, this is: The NGAM is a function of temperature $\text{NGAM} = f(T)$ which the authors pre-defined based on experimental data.

In total, there are six thermodynamic parameters which affect the temperature dependence. Direct measurement of these parameters is difficult. Using additional data such as k_{cat} at optimum temperature and various heuristics, Li *et al.* reduced the problem to three parameters:

- T_{opt}: Temperature optimum of enzyme
- T_m: Melting temperature of enzyme
- $\Delta C_P^{\ddagger}$: As mentioned earlier, the change in heat capacity from the initial state to the transition state

T_{opt} and T_m can be measured relatively easily by *in vitro* enzyme assays. However, such measurements do not exist for all of the enzymes in the ecGEM7.6 model employed by Li *et al.*. Furthermore, experimental determination of $\Delta C_P^{\ddagger}$ were missing. Finally, the experimentally determined values also needed correction to account for *in vivo* effects. For these reasons, Li *et al.* developed a Bayesian approach to infer the parameters. This Bayesian approach is based on the following components:

- Priors consisting of experimentally determined values of the enzyme parameters where available and imputed values in the other cases.
- Experimental data from aerobic batch, anaerobic batch and chemostat experiments under varying temperatures.
- A statistical model of the parameters. In this case the marginal distribution for each of the parameters is assumed to be independent and normally distributed.

- A distance function which accepts a proposed parameter set, run etcFBA on a set of scenarios and compares the results with the experimental data. The output is an R^2 value indicating the fitness with the experimental data. The fitness increases with R^2 which has a maximum value of 1.
- A sequential Monte-Carlo-based approximate Bayesian calculation method. This calculation method is seeded with the priors and uses the distance function to approximate the posterior distribution of the parameters. This posterior distribution is defined as all parameter sets for which $R^2 > 0.9$.

This Bayesian approach is the focus of Paper III, where we analyze the inference process and discuss how the design of each of these components affect the inferred parameters.

4.3.4 Combinations of extensions (decFBA)

Even though adding time dynamics, enzyme constraints and temperature dependence to FBA can be portrayed as separate processes, they can be combined in order to create more complex models. A temperature dependent GEM as we discussed in the previous subsection, is by default an enzyme constrained model because the temperature acts through the enzymatic constraints. In Paper IV, we take advantage of the fact that using an ecFBA model within the dFBA formalism yields a dynamic enzyme constrained FBA (decFBA) model without requiring any new ideas. For instance, Moreno-Paz *et al.*[15] demonstrated combining the enzyme constrained model Yeast8[199] with dynamic FBA to model the growth and metabolism of *S. cerevisiae* in batch and fed-batch reactors. The results matched experimental data to a great extent.

4.4 Automated tools for GEM reconstruction (Paper IV)

4.4.1 Creation of model

Although the genome sequence is the primary source of information about the metabolic capabilities, using an automated tool[200] to reconstruct a metabolic model only results in a *draft* model which is not suited to yield reliable results at its own. For creating a high-quality model, manual curation is required[201]. This includes filling metabolic gaps due to insufficient annotations and non-enzymatic reactions, ensuring reaction reversibility is correct, and adding export and import reactions. Hence, the process of curating a GEM often becomes a tiresome and time-consuming task.

For the reasons mentioned, such careful manual curations are only done on a handful of organisms. This raises problems when applying GEMs to non-model organisms. Such situations typically arise when studying ecological interactions between several organisms[202] or while screening a range of non-model organisms. The latter situation is present in Paper IV where five non-*Saccharomyces* strains are considered. While in theory, a manually curated model can be constructed for every single organism in question, this is often infeasible given the available time and resources.

Manual curation cannot be avoided altogether if the goal is to produce high quality GEMs. However, there exist semi-automated tools which reduces the manual labor burden when creating GEMs[203]. One of these tools is CarveME[204] in which a well curated universal model is used as a template for creating GEMs for a group of organisms i.e. yeasts. From the universal model, a model for a specific organism is constructed by ”carving” out the reactions annotated in the genome. Subsequently, having a high quality universal model makes creation of models of multiple organisms within the group much easier. Due to the simplification CarveME provides when creating models for related organisms, we applied CarveME in Paper IV for creating models of the non-*Saccharomyces* yeast strains. An example of such a model is visualized in Figure 4.4.

4.4.2 Incorporation of enzyme constraints

Generation of an enzymed constrained model is usually done by augmenting an existing GEM with enzyme constraints. Although public databases such as Uniprot[206], BRENDA[183, 184] and Sabio-RK[182] contain protein masses and enzyme catalytic constants, retrieving this data for every enzyme in a GEM is tedious and time-consuming if done manually. For this reason, tools such as GECKOMAT[191] and AutoPACMEN[193] have been developed for automatically retrieving these parameters and incorporate them into GECKO and sMOMENT models, respectively.

For Paper IV, we chose to apply AutoPACMEN. This package accepts a GEM with EC numbers and UniProt IDs as identifiers and fetches protein masses and k_{cat} values automatically from databases. Even though the automation saves a lot of work, it is not perfect. Heuristics must be in place to pick the k_{cat} value from a related organism when there is no measurement of the organism in question and EC numbers lacking k_{cat} values are inferred from the closest EC numbers. Even though autoPACMEN allows for specifying enzyme complexes with appropriate stochiometry, this would require lots of manual work and was not carried out in Paper IV.

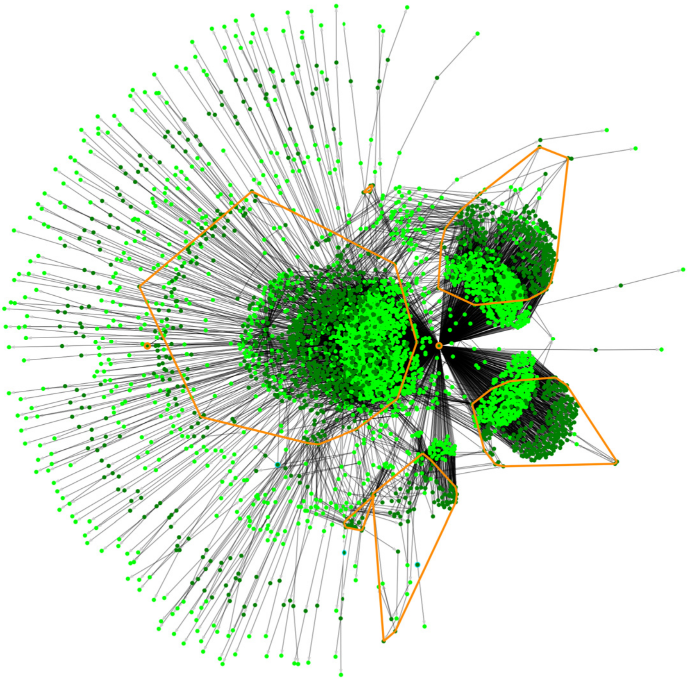


Figure 4.4: The sMOMENT GEM of *Metschnikowia pulcherrima* from paper IV visualized by ModelExplorer[205]. The GEM is shown as a bipartite network where the reactions are light green nodes whereas the metabolites are dark green nodes. The yellow polygons delimit the model cell compartments: the cytosol, the nucleus, the endoplasmatic reticulum, lipid particle, the Golgi apparatus, mitochondrion, and a pseudo compartment for the enzyme pool.

Chapter 5

Conclusion

5.1 Summary of the work

Dealing with different types of biological networks in this thesis, we found that there were recurring classes of challenges. All of these topics are well known within the scientific community, yet they materialized in different ways when working with the data, running the computations, and analyzing the results. We will discuss each of these topics in turn from the perspective of this thesis.

5.1.1 Computational performance

The concern of computational performance is twofold. First, an algorithm's demand for processing power, memory, and specialized hardware may require more powerful and expensive computers. This puts an additional burden on users to obtain access to sufficient hardware. Second, requesting computational resources, or waiting days for the computational analyses to complete, elevates the threshold for putting the algorithms into use. The latter is especially true for exploratory analyses when the analysis often is re-run with different parameters or slightly different algorithms.

As shown in this thesis, minor changes in implementation details can have large impact on performance, and thus alleviating the requirements for computer hardware and time. In paper II, we showed that csdR provided a 33 times speedup compared to the original CSD implementation. This difference is primary due to two optimizations: (1) avoiding re-calculating observation ranks for each gene pair and (2) offloading calculation of correlations to a highly optimized routine.

Likewise, we showed in Paper III that the processing power used by the Bayesian calculation methods is mostly spent on applying changes to the model before optimization. The main culprit of the original implementation using CobraPy, was that CobraPy calls the solver to update the model for every modification being done. However, updating the solver only needs to be done at few times when preparing the model. In the improved version, ReFramed's option of not updating the solver is thus used to avoid unnecessary overhead, reducing the time consumption by a factor of 8.5.

In some cases, there exist optimized algorithms for specific types of problems which perform much better than the more general ones. One example of this was the alignment the 16S sequences of the 1 537 OTUs in the Selection-Switch dataset (Paper I). Using the *de novo* multiple alignment algorithm MUSCLE[207] implemented in MEGA X[208], the alignment took approximately a day. However, the SINA[209] aligner is specialized for 16S sequences and uses a predetermined index of aligned sequences. Due to these customizations, we were able to align the sequences in a matter of seconds instead of hours.

5.1.2 Interpretation of results

Interpreting results for computational methods can sometimes be a challenging endeavour, and care must be taken in order to avoid leaping to erroneous conclusions. This is evident in Paper I where the inferred interactions *did not* correspond to ecological interactions as originally expected, but rather which bacteria have the same environmental preferences.

Furthermore, results generated from insufficient data or poor algorithms can be misleading. An example of this is the Bayesian calculation method studied in Paper III. Being a Bayesian calculation method, its target is to assess the uncertainty of the inferred parameters, and yet fails to do so. This in turns means that researchers trusting the method are likely to obtain misleading predictions.

A major challenge of Paper IV was to make sound conclusions from simulations which were confounded with large amounts of uncertainty. The models are automatically generated from a curated universal yeast model, and thus, were not individually manually curated, nor did we have enough high quality data to calibrate the models. Hence, it is unlikely that the curves of Paper IV can be reproduced in a laboratory setting. Still, we believe that the results sheds light on the ability of *Metschnikowia pulcherrima* to respire sugar more efficiently than the other yeasts, and therefore, helps explaining why this yeast is a good candidate for production of wine with lower alcohol content.

5.1.3 Requirements for sufficient data

All network inference algorithms and network models need data to function. The data must be of a certain quality for the algorithm to make use of it, and there must be a a sufficient quantity of data in order to arrive at reliable predictions. The Tara and the Selection-Switch datasets consist of 128 and 202 samples respectively, which is enough to generate interaction networks with ReBoot. However, when trying to generate ReBoot networks for the Carbon-Cycle dataset, with only 12 samples, we did not obtain any sensible results. This suggests that, even though the quality of the Carbon-Cycle dataset was sufficient, quantity was not.

Furthermore, the CSD algorithm discussed in Paper II, requires gene expression profiles of approximately 100 patients in order to give decent results. For experimentalists working with research animals such as mice, obtaining such large samples sizes is usually infeasible. In effect, CSD is restricted to large-scale clinical studies and requires large coordinated efforts to make usable data available.

The need for more data is also evident in Paper III, where the training datasets for inferring thermodynamic parameters are insufficient for trustworthy estimates. This is not very surprising, given that 2,292 parameters were estimated based on only 16 or 22 data points, where either growth rate or exchange fluxes of ethanol, carbon dioxide, and glucose are measured.

5.2 Opportunities for future work

5.2.1 Improved collection and dissemination of data

Given the large demand for data of high quality, future efforts must be put into facilitating high throughput experiments and store the data in a way which enables reuse and interoperability. In this thesis, none of the raw data were specifically produced for the papers written, but were taken from databases or earlier studies. High throughput experiments are often expensive to conduct and emphasis should therefore be put into designing experiments to be as efficient as possible given the data collected, and making the experimental data easily available for future use.

Analyses of data already collected takes place inside computers which are programmable and as a consequence, pipelines for analyzing large amounts of data can be automated in an arbitrary manner. Hence, the limiting factors are the analyst's ability to program the pipelines and the computer's processing power. Running and collecting data from experiments however, cannot be done just by specifying the protocol, but require manual work. This is the primary reason why it is usually easier to analyze generated data than to create the data in the first place.

Conducting experiments for high throughput data for a large number of samples would be more feasible if the process of running the experiment, collecting and preparing the samples was automated, and thus required less manual time and effort. For instance, the Selection-Switch dataset[85] used in Paper I could have been easier to produce and provide better time resolution if sampling, extraction of DNA and PCR amplification was automated by laboratory machinery. Automated platforms for collection and preparation of samples for high throughput analyses do exist. However, the commercial solutions available are usually expensive and hard to customize to lab-specific workflows, limiting their adaptation. Still, much progress have been done into making Do-it-yourself (DIY) approaches for laboratory automation available for the general researcher[210, 211, 212]. We believe that more emphasis and research into laboratory automation is likely to make large-scale experiments with high throughput more feasible.

Increasing throughput in the laboratories is necessary, but not sufficient for generating high-quality data. Initiatives are also needed at a higher level in order to coordinate and standardize the creation of research data. This ensures that there is an agreement among researchers about which types of data are generated, how it is collected and stored, and how it is made available for future use. In this respect, collaborations for human medical studies have seen the most success as large-scale efforts such as the The Trøndelag Health Study (HUNT)[213], The Cancer Genome Atlas Program (TCGA) and UK Biobank provide systematic collection, curation, storage and dissemination of experimental results at a large scale. This has enabled researchers to use tools such as genome-wide association studies (GWAS)[214] which has proven to be an invaluable tool for finding genetic risk factors for diseases. However, the mentioned resources are restricted to human medical data and parallels to these efforts do not, to your knowledge, exist outside this field of study.

Certainly, there exist initiatives and databases designed to convey knowledge for other organisms as well, such as *Saccharomyces cerevisiae* genome database[215] for information on *Saccharomyces cerevisiae*, SILVA[216] for 16S rRNA sequences, and BiGG models[217] for genome-scale models. While some of these databases are of high quality, the coverage and quality are variable, and some databases are at the risk of being discontinued or no longer receiving updates. Also, there are some kinds of data for which there not yet exits any good and comprehensive databases. This includes data from yeast fermentation experiments and analyses of microbial communities. As such, the raw data for Paper I and III in this thesis are not indexed in any public database and thus must be accessed in an ad-hoc manner through general data repositories without opportunities for good indexing.

In an ideal world, all research data of importance should be stored and indexed in a structured manner. The FAIR principles (Findability, Accessibility, Interoperability, and Reusability)[218] communicate the ideals for how scientific data should be stored and managed. However, the scientific community as a whole is currently far from conforming to these guidelines. Still, we believe that continued efforts into coordination of data generation, management, and storage has the potential to improve data availability and reuse. The ongoing push from scientific journals and funding bodies into data management will likely help accelerate the data literacy of researchers[219, 220, 221] and contribute to a culture of accessible and interoperable research data.

5.2.2 Standards and tools for handling FBA extensions

The Systems Biology Markup Language (SBML)[222] is a standard for exchanging quantitative models of biological systems. It is supported by various software tools and packages, and is widespread for working with and exchanging GEMs. The format is based on XML and has been revised and extended, the recent revision being Level 3 Version 2[223]. While we found the current SBML format satisfactory for baseline FBA models, we consider SBML insufficient when working with dFBA and ecFBA. We think there is large room for improvements of SBML in these areas which could make exchanging and working with FBA extensions easier.

SBML supports both kinetic models and constraint-based models, but a model defined in SBML cannot integrate both simultaneously. dFBA models integrate constraint-based models for determining the fluxes based on the external environment and a kinetic model of how the external environment is affected by the constraint-based model. Therefore, implementations of dFBA typically rely on scripts which keep track of the interaction between the nutrient environment and the constraint-based model. This approach is often tailor-made for each application as in Paper IV. In effect, porting dFBA models to other programming languages and modeling frameworks requires considerable amounts of manual work and is error-prone. Recently, however, progress has been made on this topic, as there now exists a suggested format for incorporating dFBA into SBML models[224], but acceptance of such a standard and widespread software support most likely remains years into the future.

In a similar manner, the current SBML standard has poor support for enzyme constrained models. Currently, sMOMENT stores proteins constraints in the SBML file by adding the enzyme usage as a pseudo-metabolite to the reactions. This generates models which are reproducible for all constraint-based modelling frameworks which support SMBL. However, the stoichiometry of pseudo-metabolites in the reactions is due to both protein mass, saturation factor, and turnover numbers.

Most often, protein mass is known with a far higher confidence than the turnover number, but the way the model is stored in SBML makes it difficult to disentangle these two numbers, for instance when running parameter inference or knocking out isoenzymes with different turnover numbers.

For etcGEMs, the shortcomings of SBML and the modelling tools which use them, become even more evident, as etcGEMs are built on top of enzyme constrained models and contain additional parameters. Hence, the current etcGEM implementation created by Li *et al.*[189] is difficult to understand, modify, and reproduce. When working on Paper III, we found it challenging that the computational model did not allow extracting and setting k_{cat} values directly. Thus, cumbersome workarounds were required for modifying effective k_{cat} values based on the thermodynamic parameters of the model. COBREXA[225], a novel constraint based framework, has implemented support for sMOMENT and GECKO models by incorporating the protein masses and k_{cat} values for each of the isoenzymes in each reaction. Still, COBREXA does not provide any specific format for reading and writing this information to file.

The good news is that SBML Level 3 is extensible and allows users to implement their own packages, supporting new types kinds of annotations. There also exists a software tool named Deviser (`https://github.com/sbmlteam/deviser`) to facilitate the process of creating new packages. Implementing an SBML package does not, by itself, solve the aforementioned problems or shortcomings of the SBML standard. Any such extension must gain enough acceptance in the scientific community and be implemented by many enough software tools that it becomes practical to create, read, and distribute SBML files with the extensions for dFBA and ecGEMs. Still, we believe that the demand for interoperability and reproducibility, in addition to the growing complexity of genome-scale metabolic models, will cause a high demand for domain-specific SBML packages. In turn, we believe that this demand will result in SBML packages for dFBA and ecGEMs to be implemented and widely used.

Bibliography

[1] Duarte CM, Holmer M, Olsen Y, Soto D, Marbà N, Guiu J, et al. Will the Oceans Help Feed Humanity? BioScience. 2009 December;59(11):967–976. Available from: https://doi.org/10.1525/bio.2009.59.11.8.

[2] Brown JA, Minkoff G, Puvanendran V. Larviculture of Atlantic cod (Gadus morhua): progress, protocols and problems. Aquaculture. 2003;227(1):357–372. 3rd Fish and Shellfish Larviculture Symposium. Available from: https://www.sciencedirect.com/science/article/pii/S0044848603005143.

[3] Puvanendran V, Mortensen A, Johansen LH, Kettunen A, Hansen OJ, Henriksen E, et al. Development of cod farming in Norway: Past and current biological and market status and future prospects and directions. Reviews in Aquaculture. 2022;14(1):308–342. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12599.

[4] Vadstein O, Bergh O, Gatesoupe FJ, Galindo-Villegas J, Mulero V, Picchietti S, et al. Microbiology and immunology of fish larvae. Reviews in Aquaculture. 2013;5(s1):S1–S25. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1753-5131.2012.01082.x.

[5] Vadstein O, Attramadal KJK, Bakke I, Olsen Y. K-Selection as Microbial Community Management Strategy: A Method for Improved Viability of Larvae in Aquaculture. Frontiers in Microbiology. 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02730.

[6] Lauzon HL, Gudmundsdottir S, Steinarsson A, Oddgeirsson M, Petursdottir SK, Reynisson E, et al. Effects of bacterial treatment at early stages of Atlantic cod (Gadus morhua L.) on larval survival and development. Journal of Applied Microbiology. 2010;108(2):624–632. Available from: https://ami-journals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2672.2009.04454.x.

[7] Mira de Orduña R. Climate change associated effects on grape and wine quality and production. Food Research International. 2010;43(7):1844–1855. Climate Change and Food Science. Available from: https://www.sciencedirect.com/science/article/pii/S0963996910001535.

[8] Alston J, Fuller KB, Lapsley JT, Soleas G. Too Much of a Good Thing? Causes and Consequences of Increases in Sugar Content of California Wine Grapes*. Journal of Wine Economics. 2011;6(2):135–159. Available from: https://EconPapers.repec.org/RePEc:cup:jwecon:v:6:y:2011:i:02:p:135-159_00.

[9] Varela C, Dry PR, Kutyna DR, Francis IL, Henschke PA, Curtin CD, et al. Strategies for reducing alcohol concentration in wine. Australian Journal of Grape and Wine Research. 2015;21(S1):670–679. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajgw.12187.

[10] Longo R, Blackman JW, Torley PJ, Rogiers SY, Schmidtke LM. Changes in volatile composition and sensory attributes of wines during alcohol content reduction. Journal of the Science of Food and Agriculture. 2017;97(1):8–16. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jsfa.7757.

[11] Upadhyay A. Cancer: An unknown territory; rethinking before going ahead. Genes & Diseases. 2021;8(5):655–661. Available from: https://www.sciencedirect.com/science/article/pii/S2352304220301203.

[12] Krieghoff-Henning E, Folkerts J, Penzkofer A, Weg-Remers S. Cancer - an overview. Medizinische Monatsschrift fur Pharmazeuten. 2017 February;40:48–54.

[13] National Institutes of Health (US). Understanding Cancer. Bethesda (MD); 2007. Available from: https://www.ncbi.nlm.nih.gov/books/NBK20362/.

[14] Vadstein O, Øie G, Olsen Y, Salvesen I, Skjermo J, Skjåk-Bræk G. A strategy to obtain microbial control during larval development of marine fish. Reinertsen H, Dahle LA, Jørgensen L, Tvinnereim K, editors. Balkema; 1993.

[15] Moreno-Paz S, Schmitz J, Martins Dos Santos VAP, Suarez-Diez M. Enzyme-constrained models predict the dynamics of Saccharomyces cerevisiae growth in continuous, batch and fed-batch bioreactors. Microbial biotechnology. 2022 January;15(5):1434–1445. Available from: https://ami-journals.onlinelibrary.wiley.com/doi/abs/10.1111/1751-7915.13995.

[16] Scott WT Jr, Smid EJ, Notebaart RA, Block DE. Curation and Analysis of a Saccharomyces cerevisiae Genome-Scale Metabolic Model for Predicting Production of Sensory Impact Molecules under Enological Conditions [Article]. Processes. 2020 September;8(9). Available from: https://www.mdpi.com/2227-9717/8/9/1195.

[17] Silverman EK, Schmidt HHHW, Anastasiadou E, Altucci L, Angelini M, Badimon L, et al. Molecular networks in Network Medicine: Development and applications [Review]. Wiley Interdisciplinary Reviews-Systems Biology and Medicine. 2020 November;12(6). Available from: https://doi.org/10.1002/wsbm.1489.

[18] Pawson T, Linding R; Inst Res Biomed. Network medicine. Febs Letters. 2008 April;582(8):1266–1270. Conference on Targeting and Tinkering with Interaction Networks, Barcelona, SPAIN, APR 14-16, 2008. Available from: https://doi.org/10.1016/j.febslet.2008.02.011.

[19] National Research, Council and Division on Engineering and Physical, Sciences and Board on Army Science and, Technology and Committee on Network Science for Future Army, Applications. Network Science. National Academies Press; 2005. Available from: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=147002&site=ehost-live&scope=site.

[20] Barabási AL. Network science. Cambridge: Cambridge University Press; 2016.

[21] James G, Witten D, Hastie T, Tibshirani R. Statistical Learning. In: An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer US; 2021. p. 15–57. Available from: https://doi.org/10.1007/978-1-0716-1418-1_2.

[22] Charitou T, Bryan K, Lynn DJ. Using biological networks to integrate, visualize and analyze genomics data. Genetics Selection Evolution. 2016;48(1):27. Available from: https://doi.org/10.1186/s12711-016-0205-1.

[23] Robinson SW, Fernandes M, Husi H. Current advances in systems and integrative biology. Computational and Structural Biotechnology Journal. 2014;11(18):35–46. Available from: https://www.sciencedirect.com/science/article/pii/S2001037014000221.

[24] Kholodenko BN, Bruggeman FJ, Sauro HM. Mechanistic and modular approaches to modeling and inference of cellular regulatory networks. In: Alberghina L, Westerhoff HV, editors. Systems Biology: Definitions and Perspectives. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 143–159. Available from: https://doi.org/10.1007/b136809.

[25] Di Bella JM, Bao Y, Gloor GB, Burton JP, Reid G. High throughput sequencing methods and analysis for microbiome research. Journal of Microbiological Methods. 2013;95(3):401–414. Available from: https://www.sciencedirect.com/science/article/pii/S0167701213002741.

[26] Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol. 2014 May;5(219). Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2014.00219/full.

[27] Layeghifard M, Hwang DM, Guttman DS. Disentangling Interactions in the Microbiome: A Network Perspective. Trends Microbiol. 2017 March;25(3):217–228. Available from: https://www.sciencedirect.com/science/article/pii/S0966842X16301858?via=ihub.

[28] Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012 July;10(8):538–50. Available from: https://www.nature.com/articles/nrmicro2832.

[29] Macgregor PF. Gene expression in cancer: the application of microarrays. Expert Review of Molecular Diagnostics. 2003;3(2):185–200. Available from: https://doi.org/10.1586/14737159.3.2.185.

[30] Burgess JK. Gene Expression Studies Using Microarrays. Clinical and Experimental Pharmacology and Physiology. 2001;28(4):321–328. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1440-1681.2001.03448.x.

[31] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009 January;10:57–63. Available from: https://www.nature.com/articles/nrg2484.

[32] Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nature Reviews Genetics. 2019;20(11):631–656. Available from: https://doi.org/10.1038/s41576-019-0150-2.

[33] Schmidt A, Forne I, Imhof A. Bioinformatic analysis of proteomics data. BMC Systems Biology. 2014;8(2):S3. Available from: https://doi.org/10.1186/1752-0509-8-S2-S3.

[34] Chowdhury HA, Bhattacharyya DK, Kalita JK. (Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020;17(4):1154–1173. Available from: https://ieeexplore.ieee.org/document/8613814.

[35] Voigt A, Nowick K, Almaas E. A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. PLoS computational biology. 2017 September;13:e1005739. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005739.

[36] Cuevas DA, Edirisinghe J, Henry CS, Overbeek R, O'Connell TG, Edwards RA. From DNA to FBA: How to Build Your Own Genome-Scale Metabolic Model. Frontiers in Microbiology. 2016;7. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2016.00907.

[37] Blazeck J, Alper H. Systems metabolic engineering: Genome-scale models and beyond [Review]. Biotechnology Journal. 2010 July;5(7):647–659. Available from: https://doi.org/10.1002/biot.200900247.

[38] D'Ari R, Casadesús J. Underground metabolism. BioEssays. 1998;20(2):181–186. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291521-1878%28199802%2920%3A2%3C181%3A%3AAID-BIES10%3E3.0.CO%3B2-0.

[39] Rosenberg J, Commichau FM. Harnessing Underground Metabolism for Pathway Development. Trends in Biotechnology. 2019 January;37(1):29–37. Available from: https://doi.org/10.1016/j.tibtech.2018.08.001.

[40] Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology. 2012;10(4):291–305. Available from: https://doi.org/10.1038/nrmicro2737.

[41] Chen Y, Li F, Nielsen J. Genome-scale modeling of yeast metabolism: retrospectives and perspectives. FEMS Yeast Research. 2022 January;22(1). Foac003. Available from: https://doi.org/10.1093/femsyr/foac003.

[42] Konopka A. What is microbial community ecology? The ISME Journal. 2009;3(11):1223–1230. Available from: https://doi.org/10.1038/ismej.2009.88.

[43] Tian J, Ge F, Zhang D, Deng S, Liu X. Roles of Phosphate Solubilizing Microorganisms from Managing Soil Phosphorus Deficiency to Mediating Biogeochemical P Cycle. Biology. 2021;10(2). Available from: https://www.mdpi.com/2079-7737/10/2/158.

[44] Oyeleke S, Okusanmi T. Isolation and characterization of cellulose hydrolysing microorganism from the rumen of ruminants. African Journal of Biotechnology. 2008;7(10). Available from: https://www.ajol.info/index.php/ajb/article/view/58700/47026.

[45] De Wit R, Bouvier T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environmental Microbiology. 2006;8(4):755–758. Available from: https://ami-journals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1462-2920.2006.01017.x.

[46] Becking LGMB. Geobiologie of inleiding tot de milieukunde. 18-19. WP Van Stockum & Zoon; 1934.

[47] Ribeiro KF, Duarte L, Crossetti LO. Everything is not everywhere: a tale on the biogeography of cyanobacteria. Hydrobiologia. 2018;820(1):23–48. Available from: https://doi.org/10.1007/s10750-018-3669-x.

[48] Kent AD, Yannarell AC, Rusak JA, Triplett EW, McMahon KD. Synchrony in aquatic microbial community dynamics. The ISME Journal. 2007;1(1):38–47. Available from: https://doi.org/10.1038/ismej.2007.6.

[49] Lotka AJ. Elements of mathematical biology. New York: Dover; 1956.

[50] MacArthur RH. The theory of island biogeography. vol. 1 of Monographs in population biology. Princeton, N.J: Princeton University Press; 1967.

[51] Andrews JH, Harris RF. r- and K-Selection and Microbial Ecology. In: Marshall KC, editor. Advances in Microbial Ecology. Boston, MA: Springer US; 1986. p. 99–147. Available from: https://doi.org/10.1007/978-1-4757-0611-6_3.

[52] Diwan AD, Harke SN, Gopalkrishna, Panche AN. Aquaculture industry prospective from gut microbiome of fish and shellfish: An overview. Journal of Animal Physiology and Animal Nutrition. 2022;106(2):441–469. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jpn.13619.

[53] Perry WB, Lindsay E, Payne CJ, Brodie C, Kazlauskaite R. The role of the gut microbiome in sustainable teleost aquaculture. Proceedings Biological sciences. 2020 May;287:20200184. Available from: https://doi.org/10.1098/rspb.2020.0184.

[54] Defoirdt T. Implications of Ecological Niche Differentiation in Marine Bacteria for Microbial Management in Aquaculture to Prevent Bacterial Disease. PLoS pathogens. 2016 November;12:e1005843. Available from: https://doi.org/10.1371/journal.ppat.1005843.

[55] Llewellyn MS, Boutin S, Hoseinifar SH, Derome N. Teleost microbiomes: the state of the art in their characterization, manipulation and importance in aquaculture and fisheries. Frontiers in microbiology. 2014;5:207. Available from: https://doi.org/10.3389/fmicb.2014.00207.

[56] Attramadal KJK, Øie G, Størseth TR, Alver MO, Vadstein O, Olsen Y. The effects of moderate ozonation or high intensity UV-irradiation on the microbial environment in RAS for marine larvae. Aquaculture. 2012;330-333:121–129. Available from: https://www.sciencedirect.com/science/article/pii/S0044848611009318.

[57] Attramadal KJK, Øien JV, Kristensen E, Evjemo JO, Kjørsvik E, Vadstein O, et al. UV treatment in RAS influences the rearing water microbiota and reduces the survival of European lobster larvae (Homarus gammarus). Aquacultural Engineering. 2021;94:102176. Available from: https://www.sciencedirect.com/science/article/pii/S0144860921000327.

[58] Vestrum RI, Attramadal KJK, Winge P, Li K, Olsen Y, Bones AM, et al. Rearing Water Treatment Induces Microbial Selection Influencing the Microbiota and Pathogen Associated Transcripts of Cod (Gadus morhua) Larvae. Frontiers in Microbiology. 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2018.00851.

[59] Skjermo J, Salvesen I, Øie G, Olsen Y, Vadstein O. Microbially matured water: a technique for selection of a non-opportunistic bacterial flora in water that may improve performance of marine larvae. Aquaculture International. 1997;5(1):13–28. Available from: https://doi.org/10.1007/BF02764784.

[60] Badiola M, Mendiola D, Bostock J. Recirculating Aquaculture Systems (RAS) analysis: Main issues on management and future challenges. Aquacultural Engineering. 2012;51:26–35. Available from: https://www.sciencedirect.com/science/article/pii/S014486091200060X.

[61] Attramadal KJK, Salvesen I, Xue R, Øie G, Størseth TR, Vadstein O, et al. Recirculation as a possible microbial control strategy in the production of marine larvae. Aquacultural Engineering. 2012;46:27–39. Available from: https://www.sciencedirect.com/science/article/pii/S0144860911000689.

[62] Lidicker J William Z. A Clarification of Interactions in Ecological Systems. BioScience. 1979 August;29(8):475–477. Available from: https://doi.org/10.2307/1307540.

[63] Graham DW, Knapp CW, Van Vleck ES, Bloor K, Lane TB, Graham CE. Experimental demonstration of chaotic instability in biological nitrification. The ISME Journal. 2007;1(5):385–393. Available from: https://doi.org/10.1038/ismej.2007.45.

[64] Wiechmann A, Ciurus S, Oswald F, Seiler VN, Müller V. It does not always take two to tango: "Syntrophy" via hydrogen cycling in one bacterial cell. The ISME Journal. 2020;14(6):1561–1570. Available from: https://doi.org/10.1038/s41396-020-0627-1.

[65] De Vos P, Thompson F, Thompson C, Swings J. Chapter 2 - A Flavor of Prokaryotic Taxonomy: Systematics Revisited. In: Kurtböke I, editor. Microbial Resources. Academic Press; 2017. p. 29–44. Available from: https://www.sciencedirect.com/science/article/pii/B9780128047651000023.

[66] Parker CT, Tindall BJ, Garrity GM. International Code of Nomenclature of Prokaryotes [Journal Article]. International Journal of Systematic and Evolutionary Microbiology. 2019;69(1A):S1–S111. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.000778.

[67] Zuo G, Qi J, Hao B. Polyphyly in 16S rRNA-based LVTree Versus Monophyly in Whole-genome-based CVTree. Genomics, proteomics & bioinformatics. 2018 October;16:310–319. Available from: https://doi.org/10.1016/j.gpb.2018.06.005.

[68] Woese CR. Interpreting the universal phylogenetic tree. Proceedings of the National Academy of Sciences. 2000;97(15):8392–8396. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.97.15.8392.

[69] Cohan FM. What are Bacterial Species? Annual Review of Microbiology. 2002;56(1):457–487. PMID: 12142474. Available from: https://doi.org/10.1146/annurev.micro.56.012302.160634.

[70] Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America. 1977 November;74:5088–90. Available from: https://doi.org/10.1073/pnas.74.11.5088.

[71] Kitahara K, Miyazaki K. Revisiting bacterial phylogeny. Mobile Genetic Elements. 2013;3(1):e24210. PMID: 23734299. Available from: https://doi.org/10.4161/mge.24210.

[72] Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. Journal of clinical microbiology. 2007 September;45:2761–4. Available from: https://doi.org/10.1128/JCM.01228-07.

[73] Schütte UME, Abdo Z, Bent SJ, Shyu C, Williams CJ, Pierson JD, et al. Advances in the use of terminal restriction fragment length polymorphism (T-RFLP) analysis of 16S rRNA genes to characterize microbial communities. Applied microbiology and biotechnology. 2008 September;80:365–80. Available from: https://doi.org/10.1007/s00253-008-1565-4.

[74] Muyzer G, Dewaal E, Uitterlinden A. Profiling Of Complex Microbial-Populations By Denaturing Gradient Gel-Electrophoresis Analysis Of Polymerase Chain Reaction-Amplified Genes-Coding For 16s Ribosomal-Rna [Article]. Applied and Environmental Microbiology. 1993

March;59(3):695–700. Available from: https://doi.org/10.1128/AEM.59.3.695-700.1993.

[75] Heijs SK, Haese RR, van der Wielen PWJJ, Forney LJ, van Elsas JD. Use of 16S rRNA gene based clone libraries to assess microbial communities potentially involved in anaerobic methane oxidation in a Mediterranean cold seep. Microbial ecology. 2007 April;53:384–98. Available from: https://doi.org/10.1007/s00248-006-9172-3.

[76] Bartram AK, Lynch MDJ, Stearns JC, Moreno-Hagelsieb G, Neufeld JD. Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. Applied and Environmental Microbiology. 2011;77(11):3846–3852. Available from: https://journals.asm.org/doi/abs/10.1128/AEM.02772-10.

[77] Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal. 2012;6(8):1621–1624. Available from: https://doi.org/10.1038/ismej.2012.8.

[78] Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proceedings of the National Academy of Sciences. 2006;103(32):12115–12120. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.0605127103.

[79] Whitman WB, editor. Bergey’s manual of systematics of archaea and bacteria. Bergey’s Manual Trust; 2015. Available from: https://doi.org/10.1002/9781118960608.

[80] Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, Krichevsky MI, et al. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics [Journal Article]. International Journal of Systematic and Evolutionary Microbiology. 1987;37(4):463–464. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-37-4-463.

[81] Stackebrandt E, Goebel BM. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology [Journal Article]. International Journal of Systematic and Evolutionary Microbiology. 1994;44(4):846–849. Available from: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-44-4-846.

[82] Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2006 November;361:1929–40. Available from: https://doi.org/10.1098/rstb.2006.1920.

[83] Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol. 1985;39:321–46. Available from: https://doi.org/10.1146/annurev.mi.39.100185.001541.

[84] Nguyen NP, Warnow T, Pop M, White B. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. NPJ Biofilms Microbiomes. 2016;2:16004. Available from: https://doi.org/10.1038/npjbiofilms.2016.4.

[85] Gundersen MS, Morelan IA, Andersen T, Bakke I, Vadstein O. The effect of periodic disturbances and carrying capacity on the significance of selection and drift in complex bacterial communities. ISME Communications. 2021;1(1):53. Available from: https://doi.org/10.1038/s43705-021-00058-4.

[86] Sanschagrin S, Yergeau E. Next-generation sequencing of 16S ribosomal RNA gene amplicons. Journal of visualized experiments : JoVE. 2014 August;Available from: https://doi.org/10.3791/51709.

[87] Vestrum RI, Attramadal KJK, Vadstein O, Gundersen MS, Bakke I. Bacterial community assembly in Atlantic cod larvae (Gadus morhua): contributions of ecological processes and metacommunity structure. FEMS Microbiology Ecology. 2020 September;96(9). Fiaa163. Available from: https://doi.org/10.1093/femsec/fiaa163.

[88] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 October;26(19):2460–1. Available from: https://doi.org/10.1093/bioinformatics/btq461.

[89] Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Ocean plankton. Determinants of community structure in the global plankton interactome. Science. 2015 May;348(6237):1262073. Available from: https://doi.org/10.1126/science.1262073.

[90] Mathew KA, Ardelan MV, Villa Gonzalez S, Vadstein O, Vezhapparambu VS, Leiknes Ø, et al. Temporal dynamics of carbon sequestration in coastal

North Atlantic fjord system as seen through dissolved organic matter characterisation. Science of The Total Environment. 2021;782:146402. Available from: https://www.sciencedirect.com/science/article/pii/S0048969721014704.

[91] Gause GF. Competition for common food in Protozoa. In: The struggle for existence. Baltimore, The Williams & Wilkins company, 1934; 1934. p. 83–104. Https://www.biodiversitylibrary.org/bibliography/4489. Available from: https://www.biodiversitylibrary.org/item/23409.

[92] Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 502–505. Available from: https://doi.org/10.1007/978-0-387-32833-1_379.

[93] Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLOS Computational Biology. 2012 September;8(9):1–11. Available from: https://doi.org/10.1371/journal.pcbi.1002687.

[94] Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol. 2012;8(7):e1002606. Available from: https://doi.org/10.1371/journal.pcbi.1002606.

[95] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Frontiers in Microbiology. 2017;8:2224. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2017.02224.

[96] Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015 June;31(19):3172–3180. Available from: https://doi.org/10.1093/bioinformatics/btv349.

[97] Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics. 2015 June;31(20):3322–3329. Available from: https://doi.org/10.1093/bioinformatics/btv364.

[98] Deng Y, Jiang YH, Yang Y, He Z, Luo F, Zhou J. Molecular ecological network analyses. BMC Bioinformatics. 2012;13(1):113. Available from: https://doi.org/10.1186/1471-2105-13-113.

[99] Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology. 2015 May;11(5):1–25. Available from: https://doi.org/10.1371/journal.pcbi.1004226.

[100] Lotka AJ. Analytical Note on Certain Rhythmic Relations in Organic Systems. Proceedings of the National Academy of Sciences. 1920;6(7):410–415. Available from: http://www.pnas.org/content/6/7/410.

[101] Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L, Gobourne A, et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. Nature. 2015 January;517(7533):205–U207. Available from: https://doi.org/10.1038/nature13828.

[102] Stein RR, Bucci V, Toussaint NC, Buffie CG, Raetsch G, Pamer EG, et al. Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PLoS Comput Biol. 2013 December;9(12). Available from: https://doi.org/10.1371/journal.pcbi.1003388.

[103] Kloppers PH, Greeff JC. Lotka-Volterra model parameter estimation using experiential data. Appl Math Comput. 2013 November;224:817–825. Available from: https://doi.org/10.1016/j.amc.2013.08.093.

[104] Bucci V, Tzen B, Li N, Simmons M, Tanoue T, Bogart E, et al. MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. Genome Biol. 2016 June;17(121). Available from: https://doi.org/10.1186/s13059-016-0980-6.

[105] Bootstrap. In: The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 51–54. Available from: https://doi.org/10.1007/978-0-387-32833-1_40.

[106] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001 August;29(4):1165–1188. Available from: http://www.jstor.org/stable/2674075.

[107] Carr A, Diener C, Baliga NS, Gibbons SM. Use and abuse of correlation analyses in microbial ecology. The ISME journal. 2019 November;13:2647–2655. Available from: https://doi.org/10.1038/s41396-019-0459-z.

[108] Faust K, Lahti L, Gonze D, de Vos WM, Raes J. Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin

Microbiol. 2015 June;25:56–66. Available from: https://doi.org/10.1016/j.mib.2015.04.004.

[109] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: https://www.R-project.org.

[110] Schwager E, Bielski C, Weingart G. ccrepe: ccrepe_and_nc.score; 2014. Available from: http://bioconductor.org/packages/release/bioc/html/ccrepe.html.

[111] Pettersen JP. A study of the interactions and dynamics of microbial communties. NTNU; 2019. Available from: https://hdl.handle.net/11250/2656641.

[112] Correlation Coefficient. In: The Concise Encyclopedia of Statistics. New York, NY: Springer New York; 2008. p. 115–119. Available from: https://doi.org/10.1007/978-0-387-32833-1_83.

[113] Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. Quaestiones Geographicae. 2011;30(2). Available from: https://content.sciendo.com/view/journals/quageo/30/2/article-p87.xml.

[114] Fortunato S. Community detection in graphs. Physics Reports. 2010;486(3):75–174. Available from: https://www.sciencedirect.com/science/article/pii/S0370157309002841.

[115] Pons P, Latapy M, Yolum P, Gungor T, Gurgen F, Ozturan C. Computing communities in large networks using random walks. Lect Notes Comput Sc. 2005;3733:284–293. Available from: https://doi.org/10.1007/11569596_31.

[116] Joe H Ward Jr . Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association. 1963;58(301):236–244. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[117] Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 2010 July;20(7):947–59. Available from: https://doi.org/10.1101/gr.104521.109.

[118] Bock C, Jensen M, Forster D, Marks S, Nuy J, Psenner R, et al. Factors shaping community patterns of protists and bacteria on a European scale.

Environ Microbiol. 2020 June;22(6):2243–2260. Available from: https://doi.org/10.1111/1462-2920.14992.

[119] Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J. 2016 July;10(7):1669–1681. Available from: https://doi.org/10.1038/ismej.2015.235.

[120] Hirano H, Takemoto K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. BMC Bioinformatics. 2019 June;20(329). Available from: https://doi.org/10.1186/s12859-019-2915-1.

[121] Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, et al. SAR11 clade dominates ocean surface bacterioplankton communities. Nature. 2002;420:806–810. Available from: https://doi.org/10.1038/nature01240.

[122] Brown MV, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, et al. Global biogeography of SAR11 marine bacteria. Molecular systems biology. 2012 July;8:595. Available from: https://doi.org/10.1038/msb.2012.28.

[123] Partensky F, Blanchot J, Vaulot D. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters: a review. Bulletin de l'Institut Oceanographique - Special issue: Marine cyanobacteria. 1999 July;19:457–476. Available from: https://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers15-02/010019788.pdf.

[124] Hardin G. The Competitive Exclusion Principle. Science. 1960;131(3409):1292–1297. Available from: https://www.science.org/doi/abs/10.1126/science.131.3409.1292.

[125] Hutchinson GE. The Paradox of the Plankton. The American Naturalist. 1961;95(882):137–145. Available from: https://doi.org/10.1086/282171.

[126] Record NR, Pershing AJ, Maps F. The paradox of the "paradox of the plankton". ICES Journal of Marine Science. 2013 June;71(2):236–240. Available from: https://doi.org/10.1093/icesjms/fst049.

[127] Hubbell SP. The unified neutral theory of biodiversity and biogeography. vol. 32 of Monographs in population biology. Princeton, N.J: Princeton University Press; 2001.

[128] Rosindell J, Hubbell SP, Etienne RS. The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol Evol. 2011 July;26(7):340–8. Available from: https://doi.org/10.1016/j.tree.2011.03.024.

[129] Buchanan AV, Sholtis S, Richtsmeier J, Weiss KM. What are genes "for" or where are traits "from"? What is the question? BioEssays. 2009;31(2):198–208. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.200800133.

[130] Pesole G. What is a gene? An updated operational definition. Gene. 2008;417(1):1–4. Available from: https://www.sciencedirect.com/science/article/pii/S0378111908001169.

[131] Kurakin A. Self-organization vs Watchmaker: stochastic gene expression and cell differentiation. Development Genes and Evolution. 2005;215(1):46–52. Available from: https://doi.org/10.1007/s00427-004-0448-7.

[132] Bos N, Pulliainen U, Sundström L, Freitak D. Starvation resistance and tissue-specific gene expression of stress-related genes in a naturally inbred ant population. Royal Society Open Science. 2016;3(4):160062. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsos.160062.

[133] Aibar S, Abaigar M, Campos-Laborie FJ, Sánchez-Santos JM, Hernandez-Rivas JM, De Las Rivas J. Identification of expression patterns in the progression of disease stages by integration of transcriptomic data. BMC Bioinformatics. 2016;17(15):432. Available from: https://doi.org/10.1186/s12859-016-1290-4.

[134] Ye SQ, Usher DC, Zhang LQ. Gene Expression Profiling of Human Diseases by Serial Analysis of Gene Expression. J Biomed Sci. 2002;9(5):384–394. Available from: https://www.karger.com/DOI/10.1159/000064547.

[135] van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene coexpression analysis for functional classification and gene–disease predictions. Briefings in Bioinformatics. 2017 January;19(4):575–592. Available from: https://doi.org/10.1093/bib/bbw139.

[136] Crick F. Central Dogma of Molecular Biology. Nature. 1970;227(5258):561–563. Available from: https://doi.org/10.1038/227561a0.

[137] Strasser BJ; Geneva Med Sch Inst History Med & Hlth; Dept Med Genet & Dev. A world in one dimension: Linus Pauling, Francis Crick and the central dogma of molecular biology [Article; Proceedings Paper]. History and Philosophy of the Life Sciences. 2006;28(4):491–512. Conference on History of Central Dogma of Moleculary Biology and Its Epistemological Status TodayWorkshop on the 50th Anniversary of the Central Dogma, Geneva, SWITZERLANDGeneva, SWITZERLAND, FEB 22-23, 2007FEB, 2007. Available from: http://www.jstor.org/stable/23334182.

[138] Lopez E, Madero L, Melen GJ, Ramirez M, Roukos DH, Cho WC. Clinical Proteomics in Cancer Research [Article]. Current Proteomics. 2013 July;10(2):179–191. Available from: https://doi.org/10.2174/1570164611209990001.

[139] Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422(6928):198–207. Available from: https://doi.org/10.1038/nature01511.

[140] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1):559. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559.

[141] de la Fuente A. From 'differential expression' to 'differential networking' – identification of dysfunctional regulatory networks in diseases. Trends in Genetics. 2010;26(7):326–333. Available from: https://www.sciencedirect.com/science/article/pii/S0168952510000879.

[142] Helland MO. Implementation and Application of Method for Differential Correlation Network Analysis. [Master thesis]. NTNU – Norwegian University of Science and Technology; 2017. Available from: http://hdl.handle.net/11250/2465378.

[143] Eddelbuettel D, François R. Rcpp: Seamless R and C++ Integration. Journal of Statistical Software. 2011;40(8):1–18. Available from: https://www.jstatsoft.org/v40/i08/.

[144] Eddelbuettel D. Seamless R and C++ Integration with Rcpp. New York: Springer; 2013. ISBN 978-1-4614-6867-7. Available from: https://doi.org/10.1007/978-1-4614-6868-4.

[145] Eddelbuettel D, Balamuta JJ. Extending R with C++: A Brief Introduction to Rcpp. The American Statistician. 2018;72(1):28–36. Available from: https://doi.org/10.1080/00031305.2017.1375990.

[146] Chapman B, Jost G, van der Pas R. Using OpenMP: Portable Shared Memory Parallel Programming. Scientific and Engineering Computation. Cambridge: MIT Press; 2007. Available from: https://ieeexplore.ieee.org/book/6267237.

[147] Lawrence M, Morgan M. Scalable Genomics with R and Bioconductor. Statistical Science. 2014;29(2):214 – 226. Available from: https://doi.org/10.1214/14-STS476.

[148] Gulla M. An integrated systems biology approach to investigate transcriptomic data of thyroid carcinoma. NTNU; 2019. Available from: http://hdl.handle.net/11250/2621725.

[149] Morselli Gysi D, de Miranda Fragoso T, Zebardast F, Bertoli W, Busskamp V, Almaas E, et al. Whole transcriptomic network analysis using Co-expression Differential Network Analysis (CoDiNA). PLOS ONE. 2020 October;15(10):1–28. Available from: https://doi.org/10.1371/journal.pone.0240523.

[150] Tang J. Microbial metabolomics. Current genomics. 2011 September;12:391–403. Available from: https://doi.org/10.2174/138920211797248619.

[151] Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BO. Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology. 2009;7(2):129–143. Available from: https://doi.org/10.1038/nrmicro1949.

[152] Králová B. Electrophoretic methods for the isolation and characterization of enzymes. Analytica Chimica Acta. 1999;383(1):109–117. Available from: https://www.sciencedirect.com/science/article/pii/S0003267098004929.

[153] Garcia S, Thompson RA, Giannone RJ, Dash S, Maranas CD, Trinh CT. Development of a Genome-Scale Metabolic Model of Clostridium thermocellum and Its Applications for Integration of Multi-Omics Datasets and Com-

putational Strain Design. Frontiers in Bioengineering and Biotechnology. 2020;8. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00772.

[154] Domenzain I, Sánchez B, Anton M, Kerkhoven EJ, Millán-Oropeza A, Henry C, et al. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nature communications. 2022 June;13:3766. Available from: https://doi.org/10.1038/s41467-022-31421-1.

[155] Boogerd FC, Bruggeman FJ, Richardson RC. Mechanistic Explanations and Models in Molecular Systems Biology. Foundations of Science. 2013;18(4):725–744. Available from: https://doi.org/10.1007/s10699-012-9302-y.

[156] Machado D, Costa RS, Ferreira EC, Rocha I, Tidor B. Exploring the gap between dynamic and constraint-based models of metabolism. Metabolic Engineering. 2012;14(2):112–119. Available from: https://www.sciencedirect.com/science/article/pii/S1096717612000043.

[157] Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current opinion in biotechnology. 2003 October;14:491–6. Available from: https://doi.org/10.1016/j.copbio.2003.08.001.

[158] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010 March;28(3):245–8. Available from: https://www.nature.com/articles/nbt.1614.

[159] Varma A, Boesch BW, Palsson BO. Biochemical production capabilities of escherichia coli. Biotechnology and Bioengineering. 1993;42(1):59–73. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.260420109.

[160] Fell DA, Small JR. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. The Biochemical journal. 1986 September;238:781–6. Available from: https://doi.org/10.1042/bj2380781.

[161] Tyson JJ, Novak B. Control of cell growth, division and death: information processing in living cells. Interface Focus. 2014;4(3):20130070. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsfs.2013.0070.

[162] Wintermute EH, Lieberman TD, Silver PA. An objective function exploiting suboptimal solutions in metabolic networks. BMC Systems Biology. 2013 October;7(1):98. Available from: https://doi.org/10.1186/1752-0509-7-98.

[163] Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology. 2010 July;6:390. Available from: https://doi.org/10.1038/msb.2010.47.

[164] Höffner K, Harwood SM, Barton PI. A reliable simulator for dynamic flux balance analysis. Biotechnology and Bioengineering. 2013;110(3):792–802. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.24748.

[165] Gomez JA, Höffner K, Barton PI. DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. BMC Bioinformatics. 2014;15(1):409. Available from: https://doi.org/10.1186/s12859-014-0409-8.

[166] Herrmann HA, Dyson BC, Vass L, Johnson GN, Schwartz JM. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. npj Systems Biology and Applications. 2019;5(1):32. Available from: https://doi.org/10.1038/s41540-019-0109-0.

[167] Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic engineering. 2003 October;5:264–276. Available from: https://doi.org/10.1016/j.ymben.2003.09.002.

[168] Gudmundsson S, Thiele I. Computationally efficient flux variability analysis. BMC Bioinformatics. 2010;11(1):489. Available from: https://doi.org/10.1186/1471-2105-11-489.

[169] Antoniewicz MR. A guide to metabolic flux analysis in metabolic engineering: Methods, tools and applications. Metabolic Engineering. 2021;63:2–12. Tools and Strategies of Metabolic Engineering. Available from: https://www.sciencedirect.com/science/article/pii/S1096717620301683.

[170] Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries.

Current Opinion in Biotechnology. 2018;51:103–108. Systems biology • Nanobiotechnology. Available from: https://www.sciencedirect.com/science/article/pii/S0958166917302045.

[171] Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, et al. MediaDB: A Database of Microbial Growth Conditions in Defined Media. PLOS ONE. 2014 August;9(8):1–10. Available from: https://doi.org/10.1371/journal.pone.0103548.

[172] Burgard A, Pharkya P, Maranas C. OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and Bioengineering. 2003 December;84(6):647–657. Available from: https://doi.org/10.1002/bit.10803.

[173] Zhang C, Hua Q. Applications of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine. Frontiers in Physiology. 2016;6. Available from: https://www.frontiersin.org/articles/10.3389/fphys.2015.00413.

[174] Alter TB, Ebert BE. Determination of growth-coupling strategies and their underlying principles. BMC Bioinformatics. 2019;20(1):447. Available from: https://doi.org/10.1186/s12859-019-2946-7.

[175] Godara A, Kao KC. Adaptive laboratory evolution for growth coupled microbial production. World Journal of Microbiology and Biotechnology. 2020;36(11):175. Available from: https://doi.org/10.1007/s11274-020-02946-8.

[176] Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America. 2002 November;99:15112–7. Available from: https://doi.org/10.1073/pnas.232349399.

[177] Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. Proceedings of the National Academy of Sciences of the United States of America. 2005 May;102:7695–700. Available from: https://doi.org/10.1073/pnas.0406346102.

[178] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature protocols. 2007;2:727–38. Available from: https://doi.org/10.1038/nprot.2007.99.

[179] Bärwald G, Fischer A. Crabtree effect in aerobic fermentations using grape juice for the production of alcohol reduced wine. Biotechnology Letters. 1996;18(10):1187–1192. Available from: https://doi.org/10.1007/BF00128590.

[180] Hagman A, Säll T, Piškur J. Analysis of the yeast short-term Crabtree effect and its origin. The FEBS journal. 2014 November;281:4805–4814. Available from: https://doi.org/10.1111/febs.13019.

[181] Nilsson A, Nielsen J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. Scientific Reports. 2016;6(1):22264. Available from: https://doi.org/10.1038/srep22264.

[182] Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, et al. SABIO-RK–database for biochemical reaction kinetics. Nucleic Acids Res. 2012 January;40(Database issue):D790–6. Available from: https://doi.org/10.1093/nar/gkr1046.

[183] Schomburg I, Chang A, Hofmann O, Ebeling C, Ehrentreich F, Schomburg D. BRENDA: a resource for enzyme data and metabolic information. Trends Biochem Sci. 2002 January;27(1):54–6. Available from: https://doi.org/10.1016/s0968-0004(01)02027-8.

[184] Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 2002 January;30(1):47–9. Available from: https://doi.org/10.1093/nar/30.1.47.

[185] Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proceedings of the National Academy of Sciences of the United States of America. 2016 March;113:3401–6. Available from: https://doi.org/10.1073/pnas.1514240113.

[186] Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature Communications. 2018;9(1):5252. Available from: https://doi.org/10.1038/s41467-018-07652-6.

[187] Heckmann D, Campeau A, Lloyd CJ, Phaneuf PV, Hefner Y, Carrillo-Terrazas M, et al. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proceedings

of the National Academy of Sciences of the United States of America. 2020 September;117:23182–23190. Available from: https://doi.org/10.1073/pnas.2001562117.

[188] Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. Nature Catalysis. 2022;5(8):662–672. Available from: https://doi.org/10.1038/s41929-022-00798-z.

[189] Li G, Hu Y, Zrimec J, Luo H, Wang H, Zelezniak A, et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. Nature Communications. 2021;12(1):190. Available from: https://doi.org/10.1038/s41467-020-20338-2.

[190] Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási AL, et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci USA. 2007 July;104(31):12663–8. Available from: https://doi.org/10.1073/pnas.0609845104.

[191] Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol. 2017 August;13(8):935. Available from: https://doi.org/10.15252/msb.20167411.

[192] Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters. PLOS Computational Biology. 2012 July;8(7):1–9. Available from: https://doi.org/10.1371/journal.pcbi.1002575.

[193] Bekiaris PS, Klamt S. Automatic construction of metabolic models with enzyme constraints. BMC Bioinformatics. 2020 January;21(1):19. Available from: https://doi.org/10.1186/s12859-019-3329-9.

[194] Meadows AL, Karnik R, Lam H, Forestell S, Snedecor B. Application of dynamic flux balance analysis to an industrial Escherichia coli fermentation. Metabolic Engineering. 2010;12(2):150–160. Metabolic Flux Analysis for Pharmaceutical Production Special Issue. Available from: https://www.sciencedirect.com/science/article/pii/S1096717609000627.

[195] Flassig RJ, Fachet M, Höffner K, Barton PI, Sundmacher K. Dynamic flux balance modeling to increase the production of high-value compounds

in green microalgae. Biotechnol Biofuels. 2016;9:165. Available from: https://doi.org/10.1186/s13068-016-0556-4.

[196] Mahadevan R, Edwards JS, Doyle FJ 3rd. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophys J. 2002 September;83(3):1331–40. Available from: https://doi.org/10.1016/S0006-3495(02)73903-9.

[197] Hobbs JK, Jiao W, Easter AD, Parker EJ, Schipper LA, Arcus VL. Change in Heat Capacity for Enzyme Catalysis Determines Temperature Dependence of Enzyme Catalyzed Rates. ACS Chem Biol. 2013;8(11):2388–2393. Available from: https://doi.org/10.1021/cb4005029.

[198] van der Kamp MW, Prentice EJ, Kraakman KL, Connolly M, Mulholland AJ, Arcus VL. Dynamical origins of heat capacity changes in enzyme-catalysed reactions. Nature Communications. 2018;9(1):1177. Available from: https://doi.org/10.1038/s41467-018-03597-y.

[199] Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nature communications. 2019 August;10:3586. Available from: https://doi.org/10.1038/s41467-019-11581-3.

[200] Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. Biochemical Society Transactions. 2018;46(4):931 – 936. Available from: https://search.ebscohost.com/login.aspx?direct=true&db=fsr&AN=132451967&site=ehost-live.

[201] Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010 January;5(1):93–121. Available from: https://www.nature.com/articles/nprot.2009.203.

[202] van den Berg NI, Machado D, Santos S, Rocha I, Chacón J, Harcombe W, et al. Ecological modelling approaches for predicting emergent properties in microbial communities. Nature Ecology & Evolution. 2022;6(7):855–865. Available from: https://doi.org/10.1038/s41559-022-01746-7.

[203] Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biology. 2019;20(1):158. Available from: https://doi.org/10.1186/s13059-019-1769-1.

[204] Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Res. 2018 September;46(15):7542–7553. Available from: `https://doi.org/10.1093/nar/gky537`.

[205] Martyushenko N, Almaas E. ModelExplorer - software for visual inspection and inconsistency correction of genome-scale metabolic reconstructions. BMC Bioinformatics. 2019 January;20(1):56. Available from: `https://doi.org/10.1186/s12859-019-2615-x`.

[206] Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: the universal protein knowledgebase in 2021 [Article]. Nucleic Acids Research. 2021 January;49(D1):D480–D489. Available from: `https://doi.org/10.1093/nar/gkaa1100`.

[207] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32:1792–1797. Available from: `https://doi.org/10.1093/nar/gkh340`.

[208] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular biology and evolution. 2018 June;35:1547–1549. Available from: `https://doi.org/10.1093/molbev/msy096`.

[209] Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics (Oxford, England). 2012 July;28:1823–1829. Available from: `https://doi.org/10.1093/bioinformatics/bts252`.

[210] May M. A DIY approach to automating your lab. Nature. 2019 May;569:587–588. Available from: `https://doi.org/10.1038/d41586-019-01590-z`.

[211] Zhang C, Wijnen B, Pearce JM. Open-Source 3-D Platform for Low-Cost Scientific Instrument Ecosystem. SLAS Technology. 2016;21(4):517–525. Special Issue: Collaborative 3D Printing Technology. Available from: `https://www.sciencedirect.com/science/article/pii/S2472630322014121`.

[212] Wong BG, Mancuso CP, Kiriakov S, Bashor CJ, Khalil AS. Precise, automated control of conditions for high-throughput growth of yeast and bacteria with eVOLVER. Nature Biotechnology. 2018;36(7):614–623. Available from: `https://doi.org/10.1038/nbt.4151`.

[213] Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, et al. Cohort Profile: The HUNT Study, Norway. International Journal of Epidemiology. 2013 August;42(4):968–977. Available from: https://doi.org/10.1093/ije/dys095.

[214] Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59. Available from: https://doi.org/10.1038/s43586-021-00056-9.

[215] Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research. 2012 January;40:D700–5. Available from: https://doi.org/10.1093/nar/gkr1029.

[216] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2013 January;41(D1):D590–D596. Available from: https://doi.org/10.1093/nar/gks1219.

[217] King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Research. 2015 October;44(D1):D515–D522. Available from: https://doi.org/10.1093/nar/gkv1049.

[218] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018. Available from: https://doi.org/10.1038/sdata.2016.18.

[219] Koltay T. Data literacy for researchers and data librarians. Journal of Librarianship and Information Science. 2017;49(1):3–14. Available from: https://doi.org/10.1177/0961000615616450.

[220] Nature Journal Editorial policies. Reporting standards and availability of data, materials, code and protocols; 2022. Available from: https://www.nature.com/nature/editorial-policies/reporting-standards.

[221] The Research Council of Norway. Sharing research data; 2022. Available from: https://www.forskningsradet.no/en/research-policy-strategy/open-science/research-data/.

[222] Hucka M, Finney A, Sauro H, Bolouri H, Doyle J, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models [Article]. Bioinformatics. 2003 March;19(4):524–531. Available from: https://doi.org/10.1093/bioinformatics/btg015.

[223] Hucka M, Bergmann FT, Chaouiya C, Draeger A, Hoops S, Keating SM, et al. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2. Journal of Integrative Bioinformatics. 2019 June;16(2, SI). Available from: https://doi.org/10.1515/jib-2019-0021.

[224] König M, Watanabe LH, Grzegorzewski J, Myers CJ. Dynamic Flux Balance Analysis Models in SBML. bioRxiv. 2022;Available from: https://www.biorxiv.org/content/early/2022/03/28/245076.

[225] Kratochvíl M, Heirendt L, Wilken SE, Pusa T, Arreckx S, Noronha A, et al. COBREXA.jl: constraint-based reconstruction and exascale analysis. Bioinformatics. 2021 November;38(4):1171–1172. Available from: https://doi.org/10.1093/bioinformatics/btab782.

Paper I

Robust bacterial co-occurence community structures are independent of r- and K-selection history

Jakob Peder Pettersen, Madeleine S. Gundersen, Eivind Almaas

Published in

Scientific Reports (2021)

scientific reports

OPEN

Robust bacterial co-occurence community structures are independent of *r*- and *K*-selection history

Jakob Peder Pettersen[1], Madeleine S. Gundersen[1] & Eivind Almaas[1,2]✉

Selection for bacteria which are *K*-strategists instead of *r*-strategists has been shown to improve fish health and survival in aquaculture. We considered an experiment where microcosms were inoculated with natural seawater and the selection regime was switched from *K*-selection (by continuous feeding) to *r*-selection (by pulse feeding) and vice versa. We found the networks of significant co-occurrences to contain clusters of taxonomically related bacteria having positive associations. Comparing this with the time dynamics, we found that the clusters most likely were results of similar niche preferences of the involved bacteria. In particular, the distinction between *r*- or *K*-strategists was evident. Each selection regime seemed to give rise to a specific pattern, to which the community converges regardless of its prehistory. Furthermore, the results proved robust to parameter choices in the analysis, such as the filtering threshold, level of random noise, replacing absolute abundances with relative abundances, and the choice of similarity measure. Even though our data and approaches cannot directly predict ecological interactions, our approach provides insights on how the selection regime affects the composition of the microbial community, providing a basis for aquaculture experiments targeted at eliminating opportunistic fish pathogens.

In aquaculture, the fish is in close contact with its environmental microbiome[1]. Fish larvae are at an especially vulnerable life stage, with high death rates causing economic problems in the aquaculture industry[2,3]. Research during the last decades has uncovered that the bacterial composition of the larval environment affects their survival, and both detrimental and favourable host-microbe interactions have been identified. This interplay strongly suggests that one may manipulate the larval environmental microbiome to improve their health and survival[4,5].

A viable approach for microbiome control is to select against opportunistic pathogens and select for favourable bacteria. This approach is based on the concept of *r*- and *K*-strategists, introduced in microbial ecology by Andrews and Harris[6]. Most opportunistic bacteria are *r*-strategists, meaning that they grow rapidly when resources are in surplus. If the opportunistic strain is pathogenic, such environments facilitate its proliferation and may subsequently lead to fish disease. However, *r*-strategists compete poorly when the environment is resource-limited. In such a competitive environment, the slow-growing *K*-strategists will quickly dominate due to their high resource-acquiring affinities and high yields[1,7]. Thus, resource availability is a crucial variable to manage in aquaculture. Stable resource availability promotes *K*-selection, whereas fluctuating availability promotes *r*-selection[7].

Since it is unclear, due to lack of experimental evidence, whether selecting for *K*-strategists will make a recurring set of bacteria co-occur or whether competition results in co-exclusion among the *K*-strategists, we wanted to investigate this problem using co-occurrence network analysis. Furthermore, we also wanted to study the extent to which *r*-strategists co-occur with *K*-strategists. Earlier studies have suggested that using similarity measures for network inference could determine bacterial niches, describe a microbial community's response to environmental disturbances, predict ecological keystone organisms, and explain changes in a microbial community over time[8–12]. We hypothesized that such a tool could partition bacteria based on their growth preferences and be useful to characterise and identify which bacteria are *r*-strategists and which are *K*-strategists. When a microbial community is subjected to external disturbances, it may change composition permanently, it may be

[1]Department of Biotechnology and Food Science, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. [2]Department of Public Health and General Practice, K.G. Jebsen Center for Genetic Epidemiology, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. ✉email: eivind.almaas@ntnu.no

resistant (insensitive to the disturbance), or it may "remember" its original state and be resilient (return to its original state after initially changing)[13]. We were interested in investigating whether *r*- and *K*-selection will give the microbial communities memory, and whether the selection regime would provide resistance or resilience against changing external factors.

Given our research questions, we found the dataset from Gundersen *et al.*[14] to be particularly useful. This is a 2 × 2 factorial crossover microcosm experiment that tested varying feeding regime and resource availability (high/low). Briefly, half of the microcosms were pulse-fed resources which promotes the growth of *r*-strategists, whereas the other half received a steady, continuous supply of nutrients promoting *K*-strategists. We hereafter refer to these selection regimes as *r*- and *K*-selection. The bacterial communities were sampled and characterised through 16S ribosomal RNA gene sequencing (16S-RNA)[9,15,16] at 18 time-points over a 50 day period.

What made the Gundersen dataset[14] particularly suited for our analysis, was that approximately halfway into the experiment (i.e. between day 28 and 29), the *r*- and *K*-selection regimes were switched such that each microcosm was subjected to both selection regimes. While the original paper on the dataset focused on studying the effect of the selection regimes and resource availability on ecological community assembly, our work uses network analysis to gain a more detailed understanding of the community dynamics than solely comparing samples can provide.

Results

To investigate the bacterial community structure and dynamics in *r*- and *K*-selected communities, we assessed the co-occurrence patterns of 1,537 operational taxonomic units[17] (OTUs) observed in the microcosm experiment. This 16S-rRNA gene dataset consisted of 202 samples from 12 microcosms cultivated over 50 days. Note that, 6 of the microcosm were *r*-selected, and 6 were *K*-selected. Between day 28 and 29, the *r*/*K*-selection regime was switched such that *r*-selected communities now were *K*-selected (the RK-group) and vice versa (the KR group). Furthermore, the microcosms varied by the amount of resources supplied, high (H) and low (L). However, exploratory analysis of the dataset did not indicate any relevant effect of the resource supply, and hereafter we will only focus on the *r*- and *K*-selection regimes.

Similarity measurements and network inference.

We assessed the co-occurrence patters between the OTUs using two similarity measurements and varying levels of random noise, OTU filtering, and type of abundance (relative/absolute). In contrast to many other 16S-rRNA microbiome datasets, we estimated the bacterial community's absolute abundances using flow cytometry.

Here, we present the results for the Spearman correlation measure with a low level of random noise, low OTU filtering threshold and absolute abundances (see the Methods' section for more details). We decided to focus on the rank-based Spearman correlation because it is widely applied for detecting associations[16,18,19]. We will later discuss the robustness of these results by contrasting and comparing with other similarity measures and parameter choices.

From an ecological perspective, an interaction between two microbes is an effect which one microorganism has on another. This includes cross-feeding, biofilm formation, and parasitism[9,20,21]. However, in further discussion, unless stated otherwise, we will use the term interaction in a network-theoretic perspective, where we apply a "guilt by association" heuristic. This means that, we define two OTUs to have a positive interaction if they co-occur in the same samples to a larger degree than expected by random chance. Conversely, we define two OTUs to have a negative interaction if they co-occur more rarely than expected by random chance[16,22,23]. Even if there cannot be any direct ecological interactions between the bacteria in different microcosms, the network concept of interactions still enables us to infer associations across samples collected from different microcosms.

We wanted to create a network of the pairwise associations between the OTUs and thus had to determine which edges to include. Selecting a hard threshold for the q-value for an interaction to be statistically significant (for instance at $q \leq 0.05$), is not an easy choice[24]. We, therefore, illustrate the number of significant interactions over a range of threshold q-values up to 0.05 (Fig. 1). From this figure, we see that there is no obvious cutoff.

There were in total 3250 interactions having $q \leq 0.05$, of which 1679 were $0.05 \geq q \geq 10^{-4}$ and 639 with $q < 10^{-10}$. Therefore, we determined 500 edges to be a reasonable balance between selecting high-significance edges and a network with lower average node connections. With this setting, the effective q-value threshold became $5.0 \cdot 10^{-13}$. The resulting network modules were labelled using the walktrap algorithm with 20 steps[25] (Fig. 2).

Phylogenetic clustering within the modules.

The co-occurance network analysis clustered the OTUs into four distinct modules (Fig. 2). Also, two OTUs were not assigned any module (shown as colourless nodes inside module 3) as they were connected to the rest of the network with negative interactions only. Module 3 and 4 stood out as the most interesting modules for two reasons. First, they had the largest number of nodes. Second, there were negative links between the modules, suggesting mutual exclusion between modules. For these two main modules we observed a large number of positive links within the modules, but negative edges between them. Therefore we wondered whether the modules were phylogenetically clustered. Indeed, we observed a clear pattern between OTU module membership and phylogenetic classification (Fig. 3 and Supplementary Table S1). Module 3 consisted primarily of *Alphaproteobacteria* (including *Rhodobacteraceae*) and *Flavobacteria*, whereas most OTUs in module 4 were *Gammaproteobacteria* (including *Colwellia* and *Vibrio*). Hence, we see that within each module, the OTUs were most often phylogenetically related.

Temporal trajectories of the microbial communities.

After having observed the network modules, we were interested in understanding the co-occurrence structures and how it influenced community dynamics.

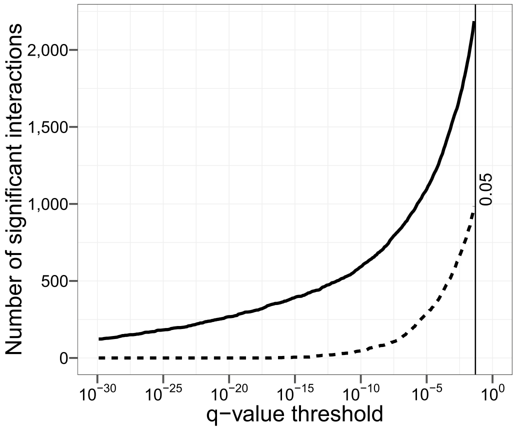


Figure 1. The cumulative number of significant interactions as a function of the critical q-value threshold considered. The solid line signifies positive interactions detected, while the dashed line represents the number of negative interactions.

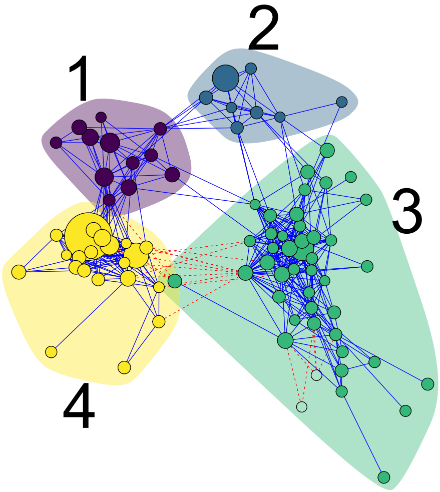


Figure 2. Module-labelled network of the 500 most significant interactions in the r/K-selection-switch dataset. Each of the 86 nodes is an OTU, while each edge corresponds to a statistically significant association between the OTUs. Blue solid edges indicate positive interactions, whereas red dashed edges indicate negative interactions. The node-sizes are scaled logarithmically according to overall mean abundance.

To investigate the community dynamics within the microbial communities, we plotted the Bray-Curtis PCoA-ordinations of the samples and observed the successional trajectories of each microcosm (Fig. 4).

From this time trajectory plot, we compared the panels diagonally and observed that microcosms undergoing K-section converged towards the upper-left area in the plot, whereas microcosms under r-section converged the middle-right area. This effect seemed independent of the experimental period and of pre-existing experimental conditions. A PERMANOVA analysis showed that the current selection regime was most important for the community composition ($R^2 = 0.344$ and $p < 10^{-6}$) compared to the minor effect of the overall selection group ($R^2 = 0.084$ and $p = 0.017$), see Methods for details. Consequently, in this respect the communities did not seem to have any memory-effect that gave rise to resistance against changes in composition. As the r/K-selection regimes resulted in clustered communities, we aimed at investigating how the network arose from these dynamics.

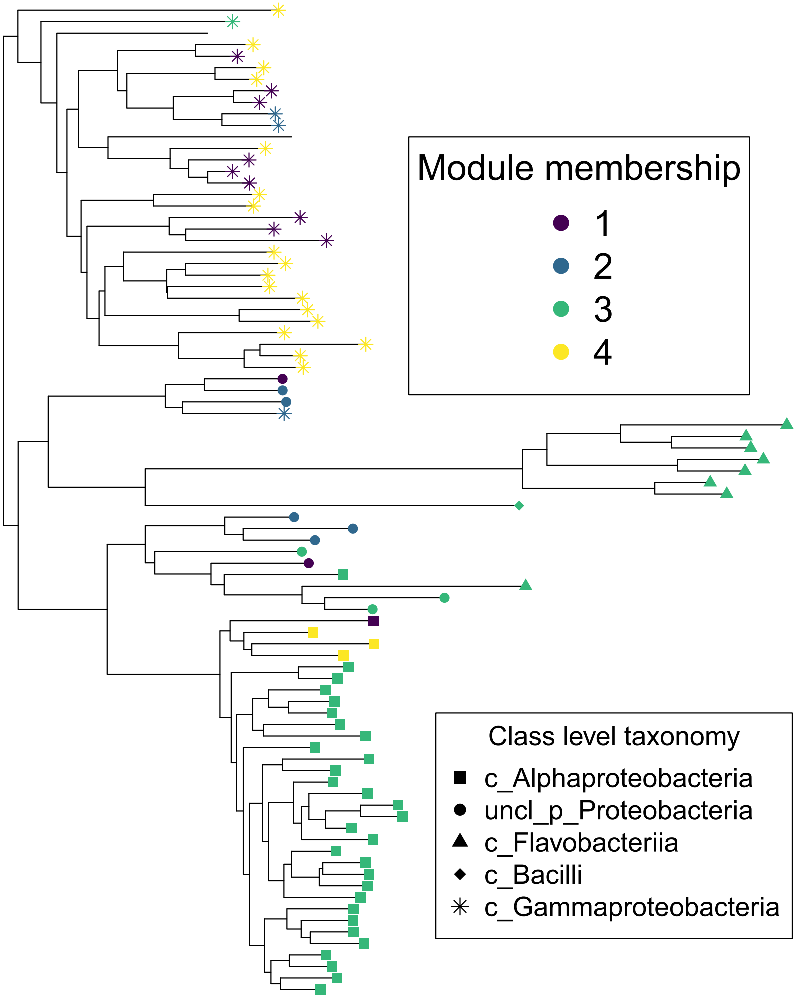


Figure 3. The phylogentic tree of the 86 OTUs from Fig. 2. together with the class level taxonomical assignment. Point colour indicates module membership, whereas the shape indicates class level taxonomical assignment. Notice that there are some inconsistencies between the phylogenetic tree and the assigned taxonomy.

***r*/*K*-strategist network patterns.** We further investigated what influenced the dynamics of the community, and conversely the dynamics' contributions to the overall network network in Fig. 2. For this, we visualised the rank-based *z*-scores of the OTU abundances (see Methods for details) for selected days during the experiment for high resource supply (Fig. 5). For low resource supply, the results were very similar and is thus not discussed any further (see Supplementary Fig. S1 for further details).

There were some obvious patterns that were apparent when investigating the temporal networks, especially with regards to module 3 and 4. The abundance of the OTUs in module 3 increased during *K*-selection, while the ones in module 4 had the opposite trend and had high abundances during *r*-selection. Hence, within each module, the OTUs had coordinated abundance patterns leading to positive inferred interactions. On the other hand, between module 3 and 4, the abundance patterns were anti-coordinated such that we obtained negative interactions.

We expected the dataset to display two time periods of instability: The first at the start of the experiment when the microbial community would adapt to lab-culture conditions, and the second disturbance instability after switching the selection regime, after day 28. During these unsteady periods, we expected more instability and less coordination between OTUs belonging to the same module. This in turn, would contribute to negative

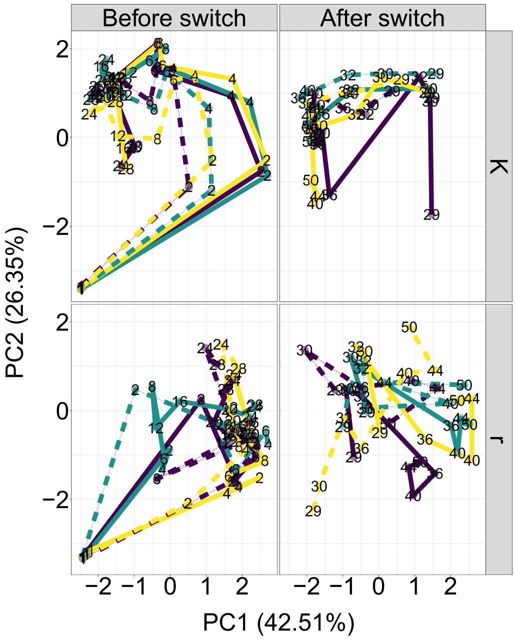


Figure 4. PCoA ordination of Bray-Curtis distances between samples showing the time trajectories for each microcosm. The single ordination was faceted vertically based on the state of selection regime at the time of sampling (being *r* or *K*-selection), and horizontally to highlight temporal trends. Solid and dotted lines indicate high (H) and low (L) resource supply, respectively. The labels indicate the day of sampling, whereas the line colours are purely to visually distinguish the replicate time series.

interactions or weaken the positive ones. However, this expectation was only partially fulfilled because we observed negative edges within modules in Fig. 5 also outside the two predicted periods of instability. One potential factor contributing to instability at the beginning of the experiment, was the fact that the oligotrophic seawater was introduced to high amount of nutrients, favouring *r*-strategists to proliferate even if the feeding was continuous.

Network robustness. If our results were different when changing parameters, the conclusions would be less likely to give us any real insight into how the communities actually behave. Therefore, we checked the robustness of the chosen parameters, changing one at a time while keeping all other parameters constant. Increasing the levels of random noise from low to medium (see Methods section) did not give any substantial difference in terms of significant interactions. Some cosmetic changes were visible due to different color labeling of communities and orientations of plots (details in Supplementary Section S2). Exchanging estimated absolute abundances with relative ones gave higher proportion of negative interactions and different assignments of OTUs into modules (see Supplementary Section S3). However, the greater trends in the results stay the same, such as clustering based on phylogeny and the considerable change in the community behaviour after switching selection regime.

Selecting a more stringent OTU filtering cutoff only has a minor consequence on the results, at the level of cosmetic changes in the plots (see Supplementary Section S4 for details). On the other hand, we notice a more pronounced effect when replacing the Spearman correlation by Pearson correlation. This is not surprising, since Spearman is non-parametric and Pearson measures degree of linear co-occurrence. In this case (Pearson), we got far fewer negative significant interactions for the same *q*-value, none of which are among the 500 most significant ones. Still, modules of phylogenetically related OTUs are present, and the selection regime still seems to explain the modules (Supplementary Section S5).

Discussion

In literature, challenges of microbial datasets such as sparsity, compositionality and habitat filtering have been addressed and solutions proposed for finding ecological interactions[22,26–29]. Despite the fact that predictions from ecological interaction-inference tools have been successfully validated in some cases[30,31], any universally accepted gold standard of finding ecological microbial interactions is not yet agreed upon. Furthermore, some reviews assessing existing methods for inferring ecological interactions have demonstrated that current methods have far too low predictive power, and more refined approaches specifically designed to cope with difficulties in microbial datasets have failed to perform better than the basic ones[19,32]. Hence, we believe that our choice

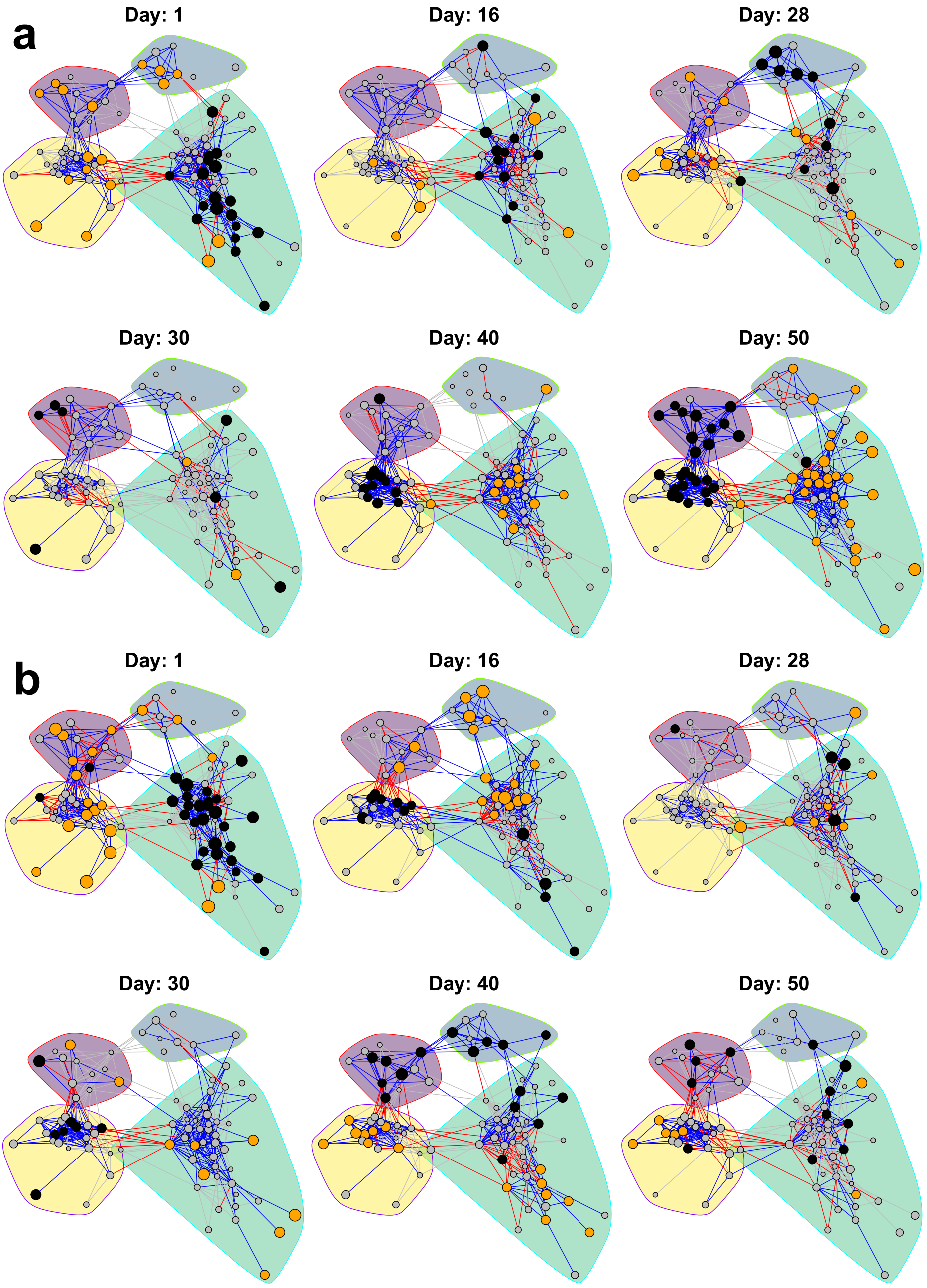


Figure 5. Dynamic visualisation of the network in Fig. 2, for (**a**) the RK selection group and (**b**) the KR selection group for high (H) resource supply. Nodes are coloured according to the corresponding OTUs' abundance compared to its overall mean for all sampling days, represented by its *z*-score. Orange, grey and black nodes mean higher, about the same or lower abundance than its mean, respectively. The edges are coloured by the product of the nodes' *z*-scores. This means that blue and red edges contribute to positive and negative association across the time series, respectively. The grey edges indicate that no major contribution to neither positive nor negative association is made. As we want to emphasize the orange and black nodes, the nodes with higher absolute *z*-scores are larger.

of using the relatively simple ReBoot[22] procedure is reasonable, even though the approach in and of itself is somewhat coarse-grained.

We observed that our correlation networks clustered the OTUs according to taxonomy and niche preferences as a result of selection for *K*- and *r*-strategists. The finding that taxonomy and niche preferences dominate co-occurrence patterns is in line with work by Chaffron *et al.*[33] who produced similar results from samples stored in a ribosomal RNA database. Along the same line, Bock *et al*[34] also noted that many of the interactions in a correlation network occur between closely related species when studying bacterial and protist communities in European lakes. Bacteria with similar niches are expected to be competitors and, hence, have negative interactions with each other. However, the effect of habitat filtering will create positive correlations between species with similar niches that are often stronger than those arising from ecological competition[10,35,36]. The same reasoning goes for taxonomical relatedness, as closely related organisms often belong to the same niche and have similar functions. This favours positive interactions within modules, whereas we, to a lesser extent got negative interactions between modules where the growth requirements are different.

Moreover, we have not undertaken any attempt to deal with indirect interactions. This means that two OTUs can appear with a strong (correlation) link even though they have no direct effect on each other, but instead interact with a third OTU. Consequently, it is challenging to determine causality when working with inferred interactions. Also, such indirect effects can be caused by environmental variables and biological entities not taken into account, such as protists and bacteriophages.

For reasons mentioned above, our results are not meant to directly represent real ecological interactions. Nevertheless, our results are interesting from a fish-health perspective, as they show that selection regime can control community composition. In terms of *r*- and *K*-selection, literature consider the orders *Alteromonadales* and *Vibrionales* represented in module 4 as *r*-strategists, whereas the *Rhodobacteraceae* in module 3 are considered *K*-strategists[37]. Additionally, *Vibrio* strains are known to cause disease in fish, whereas *Rhodobacteraceae* bacteria have been shown to protect against *Vibrio* infections through competition[38–41]. This agrees with and extends prior knowledge that *K*-selection is a potent tools for improving fish health and survival[1,7].

The long-term behaviour of the community did not appear to depend on its prehistory. Potentially, this means that changing the microbiota from a detrimental to a healthy state in a running aquaculture facility requires the same measures as ensuring a healthy microbiota for a new facility. Furthermore, the trustworthiness of the results is strengthened by their robustness to changes in parameter settings, such as filtering cut-off, amount of random noise, type of abundance, and similarity measure.

This experiment was conducted in an artificial setting without any fish of which the health could be tracked. Furthermore, we do not know whether up-scaling and broader exposure could change the workings of the microbial community. Therefore, follow-up studies could be implemented in realistic aquaculture settings, perhaps such as a RAS facility, to investigate whether switching between *K*-selection and *r*-selection will yield the same community dynamics as described in this paper. Additionally, such an experiment would provide opportunity to investigate possible connections between the state of the microbial community and the health of the fish.

We acknowledge that there exist alternative approaches one could follow. For instance, treating the OTUs as discrete units is a bit misleading. As the results show, closely related OTUs often occur together, so it could make more sense to treat the bacteria as a taxonomical continuum. A novel approach based on amplicon sequence variants (ASVs) avoids the clustering of OTUs altogether by considering each individual unique read as an own entity[42]. The phylogenetic relatedness between ASVs could then be used as a constraint for finding co-occurrence patterns. In addition, incorporating environmental information, such as organic nutrient load, salinity, and temperature, would be useful because this allows us to better predict how the desired *K*-selection should be obtained. Joint Species Distribution Models (JSDMs)[43,44] might have this useful potential to account for both species interaction, environmental factors, and taxonomical relatedness. However, its use in microbial ecology is still in its early stages and time dynamics are not yet embedded into the framework[45,46].

Methods

Selection-switch experiment. The dataset used for this article is previously published[14], but we include a brief summary for completeness: Natural seawater was collected and used to inoculate microcosms in a 2 × 2 factorial crossover design with 3 replicates conducted for 50 days, which were sampled 18 times during the experiment. Half of the microcosms were given high (H) resource supply, whereas the other half were given low (L) resource supply. The factor of resource supply level was constant throughout the experiment. The other factor was the selection regime, which meant that the microcosms were either given continuous supply of nutrients (favouring *K*-selection, and hence the designation K) or being pulse-fed with nutrients after diluting the contents of the microcosms with growth medium (favouring *r*-selection, designated R). The active selection regime was switched at the experimental halfway point (between days 28 and 29), yielding two selection groups designated as RK and KR.

DNA was extracted from the collected samples, and the V3-V4 region of the bacterial 16S-rRNA gene was amplified with PCR using broad-coverage primers and the index sequences were ligated. The amplicon library was pooled and sequenced with two runs on an Illumina MiSeq machine. The reads are available at the European Nucleotide Archive with accession number ERS7182426-ERS7182513.

The USEARCH pipeline[47] (v11) was used to remove low-quality reads and cluster the reads into OTUs at 97% similarity level. Finally, the taxonomy of the OTUs was determined by the Sintax classifier using data from the RPD training set (v.16) where the confidence threshold was set to 80%.

Quantification of bacterial density. For each sample, the bacterial density was quantified using flow cytometry (BC Accuri C6)[14]. In brief, the bacterial communities were diluted in 0.1x TE buffer, mixed with 2x

SYBR Green II RNA gel stain (ThermoFisher Scientific) and incubated in the dark at room temperature for 15 minutes. Then, each sample was measured for 2.5 minutes at 35 μL min^{-1} with an FL1-H (533/30 nm) threshold of 3000. We gated the bacterial population as those events with an FL1-A $> 10^4$ and FSC-A $< 10^5$. The raw flow cytometry data files are available at https://doi.org/10.6084/m9.figshare.15104409.

Alignment and phylogentic tree. The selection-switch dataset was acquired directly from the authors[14]. This dataset consists of a total of 206 samples. Two of these samples were taken from the communities from which the reactors were inoculated, whereas the other samples were taken from the microcosms with 17 time points x 4 regimes x 3 replicates. We discarded the inoculum samples for further analysis. The OTU reference sequences were aligned with SINA version 1.6.1[48] using the SILVA Release 138 NR 99 SSU dataset[49]. Using this aligment, the phylogentic tree was constructed by neighbour-joining using MEGA X[50] with default parameters.

Filtering and preprocessing. The mean number of reads per sample was 63,460 with standard deviation 31,411. For our analysis, we wanted to estimate the abundance of each OTU as accurately as possible and therefore skipped any correction for unequal sequencing depth. Read counts for each OTU in each sample were divided by the total number of reads for the sample, generating relative abundances. Thereafter, all OTUs having a maximum abundance (across all samples) below a certain threshold, were removed. Three levels of filtering thresholds (as count proportions) were applied: High level at $5 \cdot 10^{-3}$, medium level at $1 \cdot 10^{-3}$ and low level at $5 \cdot 10^{-4}$. The purpose of the filtering was to remove rare OTUs in order to avoid noise and spurious correlations[11]. For obtaining estimates of absolute abundances, the relative abundances were scaled by the estimate of total bacterial cell density for each sample. The `phyloseq` package (version 1.36.0)[51] and the `R` programming language (version 4.1.1)[52] facilitated this procedure. In addition, we wrote an R-package named `micInt` (version 0.18.0, available at https://github.com/AlmaasLab/micInt) to facilitate and provide a pipeline for the analysis.

Similarity measures and addition of noise. For this study, we used two similarity measures, the Pearson correlation and the Spearman correlation. A similarity measure, as referred to in this article, can be thought of as a function $f : \mathbb{R}^n \times \mathbb{R}^n \to D$ where $D = [-1, 1]$. In this regard, $f\left(\mathbf{x}, \mathbf{y}\right)$ is the similarity of two abundance vectors $\mathbf{x}$ and $\mathbf{y}$ belonging to different OTUs, where $f\left(\mathbf{x}, \mathbf{y}\right) = 1$ indicates perfect correlation, $f\left(\mathbf{x}, \mathbf{y}\right) = 0$ indicates no correlation and $f\left(\mathbf{x}, \mathbf{y}\right) = -1$ indicates perfect negative correlation. Noise was added to distort patterns of double zeros, which otherwise could result in spurious correlations. Given two vectors $\mathbf{x}$ and $\mathbf{y}$ of abundances, normally distributed noise was added to each of the abundance vectors, and the similarity measure has invoked thereafter: Given a similarity measure *f*, the similarity between the abundance vectors after adding noise is given by:

$$f^*\left(\mathbf{x}, \mathbf{y}\right) = f\left(\mathbf{x} + \boldsymbol{\varepsilon}_{\boldsymbol{x}}, \mathbf{y} + \boldsymbol{\varepsilon}_{\boldsymbol{y}}\right), \tag{1}$$

where $\boldsymbol{\varepsilon}_{\boldsymbol{x}}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{y}}$ are random vector where all components are independent and normally distributed with mean zero and variance γ^2. The level of noise γ was determined by the smallest non-zero relative abundance x_{min} in the dataset and a fixed constant s called the *magnitude factor*, such that $\gamma = s \cdot x_{\text{min}}$. For no noise, $s = 0$, for low noise $s = 1$, for middle noise $s = 10$ and for high noise $s = 100$.

Network creation. Significance of the pairwise OTU associations were determined by the ReBoot procedure introduced by Faust *et al.*[22] and shares the underlying algorithm used in the CoNet Cytoscape package[53]. This approach accepts a dataset of microbial abundances and a similarity measure, and evaluates for each pair of OTUs in the dataset the null hypothesis H_0: "The association between the OTUs is caused by chance". By bootstrapping over the samples, the similarity score of each pair of OTUs is estimated, forming a bootstrap distribution. By randomly permuting the pairwise abundances of OTUs and finding the pairwise similarity scores, a bootstrap distribution is formed. The bootstrap and permutation distribution are then compared with a two-sided *Z*-test (based on the normal distribution) to evaluate whether the difference is statistically significant. For this, the *z*-value, *p*-value and *q*-value (calculated by the Benjamini-Hochberg-Yekutieli procedure[54]) are provided for each pair of OTUs in the dataset. Our ReBoot approach is based on the `R`-package `ccrepe` (version 1.28.0)[55], but is integrated into the `micInt` package with the following major changes:

- The original ReBoot uses renormalization of the permuted abundances to keep the sum-to-constant constraint. Whereas this is reasonable to do with relative abundances, our modified version enables turning this feature off when we analyse data with absolute abundances.
- Optimizations have been made to memory use and CPU consumption to enable analyses of large datasets.
- In contrast to the usual ReBoot procedure, networks generated by the different similarity measures are not merged by *p*-value, but kept as they are.

For our analysis the number of bootstrap and permutation iterations was set to 1000. All OTUs being absent in more than $n \cdot 10^{-\frac{4}{n}}$ samples, where n is the total number of samples, were excluded through the `errthresh` argument but still kept for renormalization (if turned on). The associations were made across all samples, even the ones belonging to a different selection group or resource supply.

Dynamic PCoA visualization. All samples in the dataset were used for PCoA ordination, where the Bray-Curtis distance metric between the samples was applied before creating the decomposition. After the ordination was computed, the samples were divided into four facets based on their combination of current selection regime and resource supply. Finally, all samples belonging to the same microcosm were connected by a line in chronological order and the line was given a separate style based on the resource supply and coloured to visually distinguish it from the two other replicate microcosm within the same facet.

Permutational multivariate analysis of variance. Sequential PERmutational Multivariate Analysis of VAriance (PERMANOVA) of the samples was conducted on the absolute abundances, where only the samples from day 28 and 50 were included. These sample points correspond to time just before the experimental selection-regime crossover and a point at the end of the experiment. These days were selected because they were the most likely to capture the composition of stable communities in contrast to transient ones. The procedure was carried out by the function `adonis` from the `R` package `vegan` (version 2.5-7) with 10^6 permutations. The dependent data given to the function was the matrix of one minus the Spearman correlation of the samples (in order to resample dissimilarity), while the independent variables were the selection group (first variable) and the current selection regime (second variable).

Network visualization. The networks were plotted by the R package `igraph` (version 1.2.6)[56]. Network modules were found by the walktrap[25] algorithm implemented in `igraph` with the setting `steps=20`, including the positive edges only. Later, the negative edges were added and the networks plotted with the community labelling.

The time dynamics of the networks were visualised by taking the former network and adjusting the node colour and size, as well as the edge colour. For this, a certain combination of selection group (i.e RK) and resource supply (i.e H) was chosen. Further, let $x_{i,j,k}$ be the abundance of OTU k at sampling day i in microcosm j. As there are three replicates, we have that $j = 1, 2, 3$. If the underlying network was created by Pearson correlation, we denote the day mean $x_{i,\cdot,k}$ as the average over the replicates, this is:

$$x_{i,\cdot,k} = \frac{x_{i,1,k} + x_{i,2,k} + x_{i,3,k}}{3}. \tag{2}$$

The time series mean of OTU k, $x_{\cdot,\cdot,k}$ is the mean of these daily means over all sampling days,

$$x_{\cdot,\cdot,k} = \frac{\sum_{i=1}^{N} x_{i,\cdot,k}}{N}, \tag{3}$$

where N denotes the number of sampling days. Furthermore, we have the associated standard deviation σ_k as given by:

$$\sigma_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_{i,\cdot,k} - x_{\cdot,\cdot,k}\right)^2}. \tag{4}$$

The z-value of the abundance of OTU k at day i is then:

$$z_{i,k} = \frac{x_{i,\cdot,k} - x_{\cdot,\cdot,k}}{\sigma_k}. \tag{5}$$

This value is used in the mapping of the node sizes and colours. The node for OTU k at sampling day i has the size $a + b \cdot \left|z_{i,k}\right|$, where a and b are constants. Furthermore, the same node is coloured:

- Black if $z_{i,k} < -1$. This indicates that the OTU that day had a lower abundance than the average.
- Grey if $-1 \leq z_{i,k} \leq 1$. This indicates that the OTU that day had about the same abundance as the average.
- Orange if $z_{i,k} > 1$. This indicates that the OTU that day had a higher abundance than the average.

Furthermore, the edge colour are dependent on the product of the two participating nodes. Hence, the edge between OTU k and OTU l at day i will have the colour:

- Red if $z_{i,k} \cdot z_{i,l} < -0.3$. This shows a contribution to a negative interaction.
- Gray if $-0.3 \leq z_{i,k} \cdot z_{i,l} \leq 0.3$. This shows no major contribution of neither a positive nor negative interaction.
- Blue if $z_{i,k} \cdot z_{i,l} > 0.3$. This shows a contribution to a positive interaction.

Our approach is motivated by the fact that the Pearson correlation $\rho_{k,l}$ of the day means of OTU k and OTU l is given by:

$$\rho_{k,l} = \frac{1}{N}\sum_{i=1}^{N} z_{i,k} \cdot z_{i,l}. \tag{6}$$

For the Spearman correlation, the visualization is based on the rank of each of the OTU abundance values in a sample. Hence, instead of using the raw abundances $x_{i,j,k}$ in the calculation of the day mean, the ranks $r_{i,j,k}$

are used instead, and all subsequent calculations and mappings are the same. In a scenario when there is only one replicate, the quantity $\rho_{k,l}$ would then be the Spearman correlation of the abundances of OTU k and OTU l.

Data availability

The `micInt` package used for the analysis is available at GitHub https://github.com/AlmaasLab/micInt. The raw flow cytometry data files are available at https://doi.org/10.6084/m9.figshare.15104409 The scripts used the analyse the data and create the figures are available on GitHub at https://github.com/yaccos/Microbial-co-occurence.

Received: 4 August 2021; Accepted: 19 November 2021
Published online: 06 December 2021

References

1. De Schryver, P. & Vadstein, O. Ecological theory as a foundation to control pathogenic invasion in aquaculture. *ISME J.* **8**, 2360–2368. https://doi.org/10.1038/ismej.2014.84 (2014).
2. Hjeltnes, B., Bang-Jensen, B., Bornø, G., Haukaas, A. & Walde, C. S. The health situation in norwegian aquaculture 2018 (2019).
3. Macpherson, H. L., Bergh, O. & Birkbeck, T. H. An aerolysin-like enterotoxin from vibrio splendidus may be involved in intestinal tract damage and mortalities in turbot, *Scophthalmus maximus* (l.), and cod, *Gadus morhua* l., larvae. *J. Fish Dis.* **35**, 153–167 (2012).
4. Derome, N., Gauthier, J., Boutin, S. & Llewellyn, M. *Bacterial Opportunistic Pathogens of Fish* 81–108 (Springer, Cham, 2016).
5. May, T. *et al.* Reducing mortality associated with opportunistic infections in atlantic salmon salmo salar fry using hydrogen peroxide and peracetic acid. *Aquacul. Res.* (2021).
6. Andrews, J. H. & Harris, R. F. r-selection and k-selection and microbial ecology. *Adv. Microbial. Ecol.* **9**, 99–147 (1986).
7. Vadstein, O., Attramadal, K. J. K., Bakke, I. & Olsen, Y. K-selection as microbial community management strategy: A method for improved viability of larvae in aquaculture. *Front. Microbiol.*https://doi.org/10.3389/fmicb.2018.02730 *(2018).*
8. Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–351. https://doi.org/10.1038/ismej.2011.119 (2012).
9. Faust, K. & Raes, J. Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**, 538–550. https://doi.org/10.1038/nrmicro2832 (2012).
10. Röttjers, L. & Faust, K. From hairballs to hypotheses-biological insights from microbial networks. *FEMS Microbiol. Rev.* **42**, 761–780. https://doi.org/10.1093/femsre/fuy030 (2018).
11. Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: Unraveling microbial community dynamics. *Curr. Opin. Microbiol.* **25**, 56–66. https://doi.org/10.1016/j.mib.2015.04.004 (2015).
12. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* **15**, R89. https://doi.org/10.1186/gb-2014-15-7-r89 (2014).
13. Shade, A. *et al.* Fundamentals of microbial community resistance and resilience. *Front. Microbiol.* **3**, 417. https://doi.org/10.3389/fmicb.2012.00417 (2012).
14. Gundersen, M. S., Morelan, I. A., Andersen, T., Bakke, I. & Vadstein, O. The effect of periodic disturbances and carrying capacity on the significance of selection and drift in complex bacterial communities. *ISME Commun.* **1**, 53. https://doi.org/10.1038/s43705-021-00058-4 (2021).
15. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557. https://doi.org/10.1126/science.1107851 (2005).
16. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol.* **25**, 217–228. https://doi.org/10.1016/j.tim.2016.11.008 (2017).
17. Nguyen, N.-P., Warnow, T., Pop, M. & White, B. A perspective on 16s rrna operational taxonomic unit clustering using sequence similarity. *npj Biofilms Microbiomes*https://doi.org/10.1038/npjbiofilms.2016.4 *(2016).*
18. Springer, *Spearman Rank Correlation Coefficient*, 502–505 (Springer, 2008).
19. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–1681. https://doi.org/10.1038/ismej.2015.235 (2016).
20. Braga, R. M., Dourado, M. N. & Araújo, W. L. Microbial interactions: Ecology in a molecular perspective. *Braz. J. Microbiol.* **47**, 86–98. https://doi.org/10.1016/j.bjm.2016.10.005 (2016).
21. Tshikantwa, T. S., Ullah, M. W., He, F. & Yang, G. Current trends and potential applications of microbial interactions for human welfare. *Front. Microbiol.* **9**, 1156. https://doi.org/10.3389/fmicb.2018.01156 (2018).
22. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.*https://doi.org/10.1371/journal.pcbi.1002606 *(2012).*
23. Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front. Microbiol.*https://doi.org/10.3389/fmicb.2014.00219 *(2014).*
24. Betensky, R. A. The p-value requires context, not a threshold. *Am. Stat.* **73**, 115–117. https://doi.org/10.1080/00031305.2018.1529624 (2019).
25. Pons, P. *et al.* Computing communities in large networks using random walks. *Lect. Notes Comput. Sci.* **3733**, 284–293 (2005).
26. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, 1–11. https://doi.org/10.1371/journal.pcbi.1002687 (2012).
27. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.*https://doi.org/10.3389/fmicb.2017.02224 *(2017).*
28. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226. https://doi.org/10.1371/journal.pcbi.1004226 (2015).
29. Brisson, V., Schmidt, J., Northen, T. R., Vogel, J. P. & Gaudin, A. A new method to correct for habitat filtering in microbial correlation networks. *Front. Microbiol.* **10**, 585. https://doi.org/10.3389/fmicb.2019.00585 (2019).
30. Stein, R. R. *et al.* Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. *PLoS Comput. Biol.*https://doi.org/10.1371/journal.pcbi.1003388 *(2013).*
31. Lima-Mendez, G. *et al.* Ocean plankton determinants of community structure in the global plankton interactome. *Science (New York, N.Y.)* **348**, 1262073. https://doi.org/10.1126/science.1262073 (2015).
32. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinform.*https://doi.org/10.1186/s12859-019-2915-1 *(2019).*
33. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–959. https://doi.org/10.1101/gr.104521.109 (2010).
34. Bock, C. *et al.* Factors shaping community patterns of protists and bacteria on a European scale. *Environ. Microbiol.*https://doi.org/10.1111/1462-2920.14992 *(2020).*

35. Fisher, C. K. & Mehta, P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One*https://doi.org/10.1371/journal.pone.0102451 *(2014).*
36. Cazelles, K., Araujo, M. B., Mouquet, N. & Gravel, D. A theory for species co-occurrence in interaction networks. *Theor. Ecol.* **9**, 39–48. https://doi.org/10.1007/s12080-015-0281-9 (2016).
37. Lauro, F. M. *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 15527–15533. https://doi.org/10.1073/pnas.0903507106 (2009).
38. D'Alvise, P. W., Melchiorsen, J., Porsby, C. H., Nielsen, K. F. & Gram, L. Inactivation of vibrio anguillarum by attached and planktonic roseobacter cells. *Appl. Environ. Microbiol.* **76**, 2366–2370. https://doi.org/10.1128/AEM.02717-09 (2010).
39. Planas, M. *et al.* Probiotic effect in vivo of roseobacter strain 27–4 against vibrio (listonella) anguillarum infections in turbot (scophthalmus maximus l.) larvae. *Aquaculture* **255**, 323–333. https://doi.org/10.1016/j.aquaculture.2005.11.039 (2006).
40. D'Alvise, P. W., Lillebø, S., Wergeland, H. I., Gram, L. & Bergh, Ø. Protection of cod larvae from vibriosis by phaeobacter spp.: A comparison of strains and introduction times. *Aquaculture* **384–387**, 82–86. https://doi.org/10.1016/j.aquaculture.2012.12.013 (2013).
41. Porsby, C., Nielsen, K. & Gram, L. Phaeobacter and ruegeria species of the roseobacter clade colonize separate niches in a danish turbot (scophthalmus maximus)-rearing farm and antagonize vibrio anguillarum under different growth conditions. *Appl. Environ. Microbiol.* **74**, 7356–7364. https://doi.org/10.1128/AEM.01738-08 (2008) (**Cited By 112**).
42. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643. https://doi.org/10.1038/ismej.2017.119 (2017).
43. Warton, D. I. *et al.* So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.* **30**, 766–779. https://doi.org/10.1016/j.tree.2015.09.007 (2015).
44. Ovaskainen, O. *et al.* How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* **20**, 561–576. https://doi.org/10.1111/ele.12757 (2017).
45. Björk, J. R., Hui, F. K. C., O'Hara, R. B. & Montoya, J. M. Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Mol. Ecol.* **27**, 2714–2724. https://doi.org/10.1111/mec.14718 (2018).
46. Leite, M. F. & Kuramae, E. E. You must choose, but choose wisely: Model-based approaches for microbial community analysis. *Soil Biol. Biochem.* **151**, 108042. https://doi.org/10.1016/j.soilbio.2020.108042 (2020).
47. Edgar, R. C. Search and clustering orders of magnitude faster than blast. *Bioinformatics* **26**, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461 (2010).
48. Pruesse, E., Peplies, J. & Glöckner, F. O. Sina: Accurate high-throughput multiple sequence alignment of ribosomal rna genes. *Bioinformatics (Oxford, England)* **28**, 1823–1829. https://doi.org/10.1093/bioinformatics/bts252 (2012).
49. Quast, C. *et al.* The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596. https://doi.org/10.1093/nar/gks1219 (2013).
50. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. Mega x: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549. https://doi.org/10.1093/molbev/msy096 (2018).
51. McMurdie, P. J. & Holmes, S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
52. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
53. Faust, K. & Raes, J. Conet app: Inference of biological association networks using cytoscape. *F1000Res* **5**, 1519. https://doi.org/10.12688/f1000research.9050.2 (2016).
54. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
55. Schwager, E., Bielski, C. & Weingart, G. *ccrepe: ccrepe_and_nc.score*, r package version 1.14.0 edn. (2014).
56. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* 1695 (2006).

Acknowledgements

Thanks to Olav Vadstein for being helpful with explaining the experimental procedures and supply input on how the data could be interpreted and processed. J.P.P. and E.A. are grateful to ERA CoBioTech project CoolWine and Norwegian Research Council grant 283862 for funding.

Author contributions

E.A. and J.P.P. conceived the project. M.S.G provided the data and comments to how to interpret it. J.P.P wrote the software, did the analysis, created visualisations and wrote the first draft of the paper. All authors contributed to and accepted the final version of the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03018-z.

Correspondence and requests for materials should be addressed to E.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Paper II

csdR, an R package for differential co-expression analysis

Jakob Peder Pettersen, Eivind Almaas

Published in

BMC Bioinformatics (2022)

BMC Bioinformatics

SOFTWARE Open Access

csdR, an R package for differential co-expression analysis

Jakob P. Pettersen[1] and Eivind Almaas[1,2*]

*Correspondence:
eivind.almaas@ntnu.no
[1] Department of Biotechnology and Food Science, NTNU- Norwegian University of Science and Technology, Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: Differential co-expression network analysis has become an important tool to gain understanding of biological phenotypes and diseases. The CSD algorithm is a method to generate differential co-expression networks by comparing gene co-expressions from two different conditions. Each of the gene pairs is assigned conserved (C), specific (S) and differentiated (D) scores based on the co-expression of the gene pair between the two conditions. The result of the procedure is a network where the nodes are genes and the links are the gene pairs with the highest C-, S-, and D-scores. However, the existing CSD-implementations suffer from poor computational performance, difficult user procedures and lack of documentation.

Results: We created the R-package `csdR` aimed at reaching good performance together with ease of use, sufficient documentation, and with the ability to play well with other tools for data analysis. `csdR` was benchmarked on a realistic dataset with 20,645 genes. After verifying that the chosen number of iterations gave sufficient robustness, we tested the performance against the two existing CSD implementations. `csdR` was superior in performance to one of the implementations, whereas the other did not run. Our implementation can utilize multiple processing cores. However, we were unable to achieve more than ~2.7 parallel speedup with saturation reached at about 10 cores.

Conclusion: The results suggest that `csdR` is a useful tool for differential co-expression analysis and is able to generate robust results within a workday on datasets of realistic sizes when run on a workstation or compute server.

Keywords: R, Genome-scale, Co-expression, Gene network, Network

Introduction

Experimental high-throughput techniques, such as microarray and RNA sequencing, allow for large-scale assays of gene expressions. Correlation-based network approaches have been used for analysing a wide variety of gene-expression data in humans, identifying both individual genes and clusters of genes with prominent relationships to the phenotype (disease) in question [1–4]. More recently, there has been a realization that differential co-expression analyses, i.e. the study of changes in the correlations rather than just a test for their presence or absence in the conditions, may identify important

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

genes [5, 6]. This may be of interest for the study of diseases, as a central goal is to identify genes contributing to differences between sick patients and healthy controls.

There exist multiple methods for differential co-expression analysis [7, 8]. Some make separate co-expression networks for both conditions and compare the networks in order to score differential expressed genes [9–11]. Another major approach is based on scoring gene pairs directly based on their differential expression between different conditions [12–15]. The CSD approach [7] is of the second type and explicitly distinguishes between three different kinds of differential co-expression, that of Conserved (C), Specific (S), and Differentiated (D), hence its name. Each pair of genes will have a score for each of these three categories.

Previously, two implementations of CSD have been written. The first one (https://github.com/andre-voigt/CSD) was written as part of the original CSD work [7]. It is implemented in a combination of C++ and Python and is not focused on performance and user-friendliness. The other implementation (https://github.com/magnusolavhelland/CSD-Software) is written in C++ and is fine-tuned for performance [16]. However, practical experience has shown it difficult to use due to its strict and obscure requirements for input data format. `CoDiNa` [17] is an R package which implements a procedure similar to CSD and allows for comparing data from more than two environments. On the other hand, `CoDiNa` does not account for the variability in co-expression within an environment.

Implementation

We will assume that the expression vectors of two genes A and B have Spearman correlations of ρ_1 and ρ_2 in the first and second condition respectively. Furthermore, we define σ_1 and σ_2 as the corresponding standard deviations of the aforementioned Spearman correlations, estimated by resampling. The values for C, S and D are then defined by:

$$C = \frac{|\rho_1 + \rho_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \tag{1}$$

$$S = \frac{||\rho_1| - |\rho_2||}{\sqrt{\sigma_1^2 + \sigma_2^2}}, \tag{2}$$

$$D = \frac{||\rho_1| + |\rho_2| - |\rho_1 + \rho_2||}{\sqrt{\sigma_1^2 + \sigma_2^2}}. \tag{3}$$

The CSD algorithm consists of three principal parts:

1 Calculation of the Spearman correlation between each pair of genes for each of the two datasets individually. This is conducted with resampling to provide an estimate of the variance of the correlation.

2 Comparison of the values of mean correlation and standard deviation from the two conditions, allowing the computation of Conserved, Specific, and Differentiated scores.

3 Selection of the gene pairs with the highest values for C, S and D, and the generation of a network from them. In typical disease-network analyses, this network is studied further with tools such as module finding and enrichment analysis.

An example CSD network containing both C-, S- and D-links is shown in Fig. 1. A link in a CSD-network indicates a relation between the genes across the two conditions and is likely to be due to regulatory effects which are the same or different in the two conditions. With this in mind, we can consider the CSD network a product of the underlying gene regulatory network. This allows us to suggest regulatory mechanisms which are the same for both conditions in addition to mechanisms which are different in the two conditions. Hence, CSD can be used as a tool to point to possible gene-phenotype relationships underlying the condition in question. In turn, the results from CSD can be integrated with prior knowledge to shed more light on the genetic basis for the condition and serve as a starting point for follow-up experiments.

`csdR` is an R [18] package which implements this procedure and is written to achieve good performance, be well documented and user-friendly, and provide seamless integration with other tools in the R ecosystem. The source code is available on GitHub (https://github.com/AlmaasLab/csdR). Parts of it are written with `Rcpp` [19–21] in order to boost performance. The package is designed to utilize multicore processors and processor SIMD (Single Instruction, Multiple Data) instructions through its usage of openMP [22]. In addition, the package is available in Bioconductor release 3.14.

The data provided to the package must be numerical data organized in matrices by sample and gene. In theory, any numerical measure of gene expression could be used. In practice, normalized read counts from RNA-seq or proteomics studies are the most relevant to use. Note that imputation of missing values is not implemented in the package. If missing values are present in the raw data, an error message will be reported to the user.

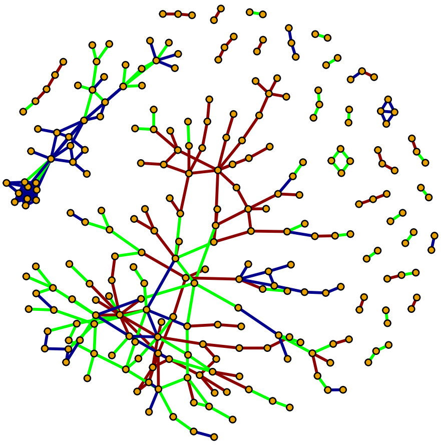


Fig. 1 An example CSD network taken from the `csdR` vignette. The nodes are genes and the edges C- (dark blue), S- (green) and D-links (dark red)

In the original implementation [7] and the CSD C++ implementation, the resampling is done through an ad-hoc method termed independent subsampling, meaning that no points are sampled together more than once and each subsample has a fixed size. Instead, our implementation uses bootstrapping [23] which is a more common statistical practice. This means that the data points are drawn with replacement, and each bootstrap sample contains as many points as the original sample. Consequently, a data point is likely to be picked more than once in the same sample or not being in the sample at all. Compared to indendent subsampling, bootstrapping may be conducted an arbitrary number of times, which ensures stable results given a sufficiently large number of iterations. In addition, bootstrapping is easier to implement and allows for faster computations.

As part of computing the Spearman correlation, the observational ranks of the genes in each sample must be computed. In the original implementation, this rank is re-computed for every gene pair. `csdR` optimizes this approach by first finding the ranks of all genes before computing the all-to-all Pearson correlation of the ranks. For this computation, the efficient `WGCNA` version of `cor` is used [1, 24]. Internally, `WGCNA::cor()` uses matrix multiplication handled by BLAS (Basic Linear Algebra Subprograms). Because this step is the major performance bottleneck, linking R against an optimized BLAS library, such as OpenBLAS (http://www.openblas.net/), is strongly recommended.

In order to ensure numerically stable computation of the variance of the co-expressions, Welford's algorithm [25] is applied. For the final step of selecting edges with the largest values of C, S and D, past implementations used random sampling to find the importance cutoff. Our implementation however, uses the more direct approach of partial sorting through the C++ STL functions `std::nth_element` and `std::sort`.

Results and discussion

For small datasets (order of 100 samples and 100 genes), all implementations are so fast that the runtime is of no practical importance. We benchmarked the different implementation on a realistic dataset derived from RNA-seq of thyroid glands. The data for the patients with thyroid cancer (case) consisted of 504 samples, while the control dataset consisted of 399 samples. These datasets are the full versions of the down-scaled datasets `sick_expression` and `normal_expression` provided in the package. See https://github.com/AlmaasLab/csdR/blob/main/inst/script/download_preprocess.md for more details on how the data were obtained and pre-processed. There were a total of 20, 645 genes being compared, which resulted in 213, 097, 690 different gene pairs. We ran the three implementations with importance level set to $p = 10^{-6}$. This resulted in C-, S- and D-networks with 213 edges each. For the two first implementations, the number of random selections was kept to 10^4, and the size of the subsamples set to 10. All benchmarked software was compiled with GCC version 9.3.0 using compiler flags `-O3 -march=native` and run on 10 virtual 2.4 GHz Intel Broadwell processors. For `csdR`, the benchmarks were conducted using R version 4.1.0 linked against libopenblas version 0.3.15.

In order to determine the number of iterations for `csdR`, we investigated the robustness of the highest ranking links across different random seeds. We ran 10 parallels with different random seeds over 1000 iterations, identified the intersection of the highest

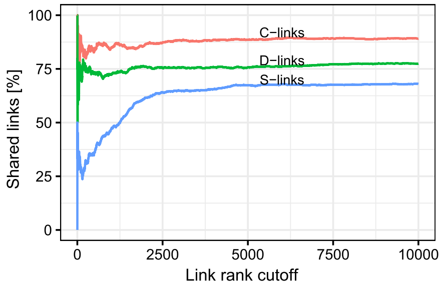


Fig. 2 Cumulative overlap of the highest ranking links from `csdR`. The shared links are the intersection of 10 parallels run over 1000 iterations with different random seeds

Table 1 Running time (s) for CSD on the large datasets with 1000 iterations

Implementation	Cores	Step 1	Step 2	Step 3	Total
Original	1	1,261,628	4900	116,051	1,382,579
csdR	1	41,387	79	11	41,477
csdR	5	20,737	50	15	20,802
csdR	10	15,488	45	13	15,546
csdR	15	15,192	50	12	15,254

ranking gene pairs between these 10 parallels, and finally calculated the proportion these shared gene pairs made up (Fig. 2). For all three link types, the recall across all 10 parallels stabilised at approximately 70%, 80%, and 90% for the S-, D and C-links, respectively, when the number of selected links exceeded 2500. For smaller numbers of links, there are more random fluctuations. We observe that the S-links have the lowest rate of recall. This observation can be attributed to the fact that gene pairs with large S-values have low levels of co-expression in one of the conditions and their scores are therefore more susceptible to random noise. We are of the opinion that the robustness at 1000 iterations is sufficient for practical use, and this choice was therefore used in the benchmarking process. Better robustness may be obtained by increasing the number of iterations at the expense of run time.

For this benchmark, we were not able to use CSD-C++, as we were unable to reshape the data into a format the program would accept. A custom format repair tool (https://github.com/lars-as/csd_cs_ged_tools) was attempted, but did not resolve the issues. For the other two implementations, the results are shown in Table 1. We notice that `csdR` is much faster than the original implementation even on a single core. The original implementation is single-threaded and can thus not take advantage of multiple cores. For `csdR`, running the algorithm on 5 cores instead of one reduces the time spent approximately by a factor two. Doubling the core count to 10 provides another reduction factor of ~25% of the time. For 15 cores however, no performance gain beyond the margin of error was observed. We suspect that the algorithm's failure to scale to such a large number of cores is due to the system's memory bandwidth being exhausted. Another result worth noticing is the fact that only the first step in the CSD procedure determines the performance in practise. The contributions from step 2 and 3 were negligible except for step 3 in the original implementation, which consumes up 8.4% of the overall time. In terms of memory usage, `csdR` consumed approximately 30 GB of RAM on the benchmarked datasets. Due to the fact that most laptops and many desktop computers have less memory than this, `csdR` is more suited for powerful workstations or compute servers.

Conclusions

We have shown that `csdR` is reasonably fast even for large datasets and provides sufficiently robust results. In addition, it is more accessible to the common user and better documented than the previous CSD implementations.

Availability and requirements

Project name: csdR
Project home page: https://github.com/AlmaasLab/csdR
Operating systems: Cross-platform
Programming language: R, C++11
Other requirements: R(>= 4.1.0), R packages `WGCNA`, `glue`, `matrixStats`, `RhpcBLASctl` and `Rcpp`
License: GNU General Public License v3.0
Any restrictions to use by non-academics: The terms of the GPL-3 license must be respected.

Abbreviations

BLAS: Basic linear Algebra Subprograms; CSD: Conserved, Specific, and Differentiated; SIMD: Single instruction, multiple data; WGCNA: Weighted correlation network analysis.

Acknowledgements

Thanks for to Kristin Salvesen for testing the software and André Voigt for helpful discussions.

Authors' contributions

J.P.P. developed, tested, and benchmarked the software and wrote the first draft of the paper. E.A. supervised the project. Both authors read and approved the final manuscript.

Funding

This work is funded by ERA CoBioTech project CoolWine and the Norwegian Research Council grant 283862.

Availability of data and materials

The scripts and datasets used for benchmarking are available for download at Figshare https://doi.org/10.6084/m9.figshare.16713121

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Biotechnology and Food Science, NTNU- Norwegian University of Science and Technology, Trondheim, Norway. [2]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and General Practice, NTNU- Norwegian University of Science and Technology, Trondheim, Norway.

Received: 30 September 2021 Accepted: 7 February 2022
Published online: 19 February 2022

References

1. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 2008;(1), 559.
2. Najafzadeh L, Mahmoudi M, Ebadi M, Dehghan Shasaltaneh M, Masoudinejad A. Co-expression network analysis reveals key genes related to ankylosing spondylitis arthritis disease: computational and experimental validation. Iran J Biotechnol. 2021;19(1):74–85. https://doi.org/10.30498/IJB.2021.2630.
3. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380–4. https://doi.org/10.1038/nature10110.
4. Miller JA, Horvath S, Geschwind DH. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci. 2010;107(28):12698–703. https://doi.org/10.1073/pnas.0914257107.
5. de la Fuente A. From 'differential expression' to 'differential networking' -identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26(7):326–33. https://doi.org/10.1016/j.tig.2010.05.001.
6. Chowdhury HA, Bhattacharyya DK, Kalita JK. (Differential) co-expression analysis of gene expression: a survey of best practices. IEEE-ACM Trans Comput Biol Bioinform. 2020;17(4):1154–73. https://doi.org/10.1109/TCBB.2019.289317.
7. Voigt A, Nowick K, Almaas E. A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. PLoS Comput Biol. 2017;13(9):1–34. https://doi.org/10.1371/journal.pcbi.1005739.
8. Kakati T, Bhattacharyya DK, Barah P, Kalita JK. Comparison of methods for differential co-expression analysis for disease biomarker prediction. Comput Biol Med. 2019;113:103380. https://doi.org/10.1016/j.compbiomed.2019.103380.
9. Reverter A, Ingham A, Lehnert SA, Tan S-H, Wang Y, Ratnakumar A, Dalrymple BP. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics. 2006;22(19):2396–404. https://doi.org/10.1093/bioinformatics/btl392.
10. Wu C, Zhu J, Zhang X. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. BMC Bioinform. 2012;13(1):182. https://doi.org/10.1186/1471-2105-13-182.
11. Choi JK, Yu U, Yoo OJ, Kim S. Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics. 2005;21(24):4348–55. https://doi.org/10.1093/bioinformatics/bti722.
12. Yu H, Liu B-H, Ye Z-Q, Li C, Li Y-X, Li Y-Y. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinform. 2011;12(1):315. https://doi.org/10.1186/1471-2105-12-315.
13. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol. 2013;9(3):1–15. https://doi.org/10.1371/journal.pcbi.1002955.
14. Gao X, Arodz T. Detecting differentially co-expressed genes for drug target analysis. Procedia Comput Sci. 2013;18:1392–401. https://doi.org/10.1016/j.procs.2013.05.306.
15. Fukushima A. Diffcorr: An r package to analyze and visualize differential correlations in biological networks. Gene. 2013;518(1):209–14. https://doi.org/10.1016/j.gene.2012.11.028.
16. Helland MO. Implementation and application of method for differential correlation network analysis. Master's thesis, NTNU - Norwegian University of Science and Technology. 2017. http://hdl.handle.net/11250/2465378
17. Morselli Gysi D, de Miranda Fragoso T, Zebardast F, Bertoli W, Busskamp V, Almaas E, Nowick K. Whole transcriptomic network analysis using co-expression differential network analysis (codina). PLoS ONE. 2020;15(10):1–28. https://doi.org/10.1371/journal.pone.0240523.
18. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. R Foundation for Statistical Computing. https://www.R-project.org/
19. Eddelbuettel D, François R. Rcpp: seamless R and C++ integration. J Stat Softw. 2011;40(8):1–18. https://doi.org/10.18637/jss.v040.i08.
20. Eddelbuettel D. Seamless R and C++ Integration With Rcpp. Springer, New York, 2013. https://doi.org/10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7
21. Eddelbuettel D, Balamuta JJ. Extending R with C++: a brief introduction to Rcpp. Am Stat. 2018;72(1):28–36. https://doi.org/10.1080/00031305.2017.1375990.
22. Chapman B, Jost G, van der Pas R. Using OpenMP: portable shared memory parallel programming. Scientific and Engineering Computation. MIT Press, Cambridge. 2007. Books24x7, Inc
23. Bootstrap. Springer, New York, NY, 2008, pp. 51–54. https://doi.org/10.1007/978-0-387-32833-1_40.
24. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw. 2012;46(11):1–17.
25. Welford BP. Note on a method for calculating corrected sums of squares and products. Technometrics. 1962;4(3):419–20. https://doi.org/10.1080/00401706.1962.10490022.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper III

Parameter inference for enzyme and temperature constrained genome-scale models

Jakob Peder Pettersen, Eivind Almaas

Under review in

Scientific Reports

Parameter inference for enzyme and temperature constrained genome-scale models

Jakob Peder Pettersen[1] **and Eivind Almaas**[1,2,*]

[1]Department of Biotechnology and Food Science, NTNU- Norwegian University of Science and Technology, Trondheim, Norway
[2]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and General Practice, NTNU - Norwegian University of Science and Technology, Trondheim, Norway
[*]Corresponding author: eivind.almaas@ntnu.no

ABSTRACT

The metabolism of all living organisms is dependent on temperature, and therefore, having a good method to predict temperature effects at a system level is of importance. A recently developed Bayesian computational framework for enzyme and temperature constrained genome-scale models (etcGEM) predicts the temperature dependence of an organism's metabolic network from thermodynamic properties of the metabolic enzymes, markedly expanding the scope and applicability of constraint-based metabolic modelling.

Here, we show that the Bayesian calculation method for inferring parameters for an etcGEM is unstable and unable to estimate the posterior distribution. The Bayesian calculation method assumes that the posterior distribution is unimodal, and thus fails due to the multimodality of the problem. To remedy this problem, we developed an evolutionary algorithm which is able to obtain a diversity of solutions in this multimodal parameter space.

We quantified the phenotypic consequences on six metabolic network signature reactions of the different parameter solutions resulting from use of the evolutionary algorithm. While two of these reactions showed little phenotypic variation between the solutions, the remainder displayed huge variation in flux-carrying capacity. This result indicates that the model is under-determined given current experimental data and that more data is required to narrow down the model predictions. Finally, we made improvements to the software to reduce the running time of the parameter set evaluations by a factor of 8.5, allowing for obtaining results faster and with less computational resources.

Introduction

Temperature is a key effector of life, which is partially due to the consequence that temperature has on catalytic properties of enzymes. For a long time, it has been known that enzymatic reactions slow down at low temperatures, whereas high temperatures destroy the enzymes, rendering them non-functional. In recent research[1,2], it has also been acknowledged that enzymes have lower catalytic rates at high temperatures due to changes in heat capacity. The effect of temperature on the behaviour of microorganisms as a whole is evident. Freezing food stops spoilage by inactivating microorganisms, whereas cooking kills them. In between these temperature extremes, there are observable effects which can be utilized commercially. One example of this is yeast production of aroma compounds, which has been shown to depend on temperature[3,4], a finding with potentially great impact on wine and beer brewing.

Until recently, no attempt has been made to computationally explain the temperature dependence of microorganismal phenotypes by propagating the temperature dependence of metabolic enzymes to the entire metabolic network of an organism. However, Gang Li *et al.*[5] came up with an extension of an enzyme-constrained genome-scale metabolic model (ecGEM) which can capture the temperature dependence of metabolism. This model is thus called an *enzyme and temperature constrained GEM* (etcGEM). As most other models, the one by Li *et al.*[5] is based on sets of assumptions and parameters. In particular, this model is based on model ecYeast7.6[6] (*Saccharomyces cerevisiae* strain S288C) and contains 764 metabolic enzymes and $2,292$ parameters associated with enzymatic activity. For each of the enzymes, the following parameters must be determined: (1) T_m, the melting temperature; (2) T_{opt}, the temperature optimum; and (3) $\Delta C_p^{\ddagger}$, the change of heat capacity from the ground state to the transitional state.

Given these parameters, it was demonstrated how an enzyme's maximum catalytic rate $k_{cat}(T)$ at a certain temperature T could be estimated[5]. These temperature-dependent maximal catalytic rates were then fed into the enzyme-constrained genome-scale model[6,7], and the metabolic flux rates were predicted using Flux Balance Analysis (FBA)[8,9]. Furthermore, Li and coworkers[5] used a Bayesian approach to infer the enzymatic parameters mentioned above from sets of training data: (1) Maximal growth rates in aerobic batch cultivations[10]; (2) Maximal growth rates in anaerobic batch cultivations[11]; (3) Chemostat cultivations which include measurements of exchange fluxes of carbon dioxide, ethanol and glucose[12].

In the training data, the experimentally determined exchange fluxes were recorded for a range of temperatures, thus generating a set of growth scenarios. The performance of a parameter set was assessed by predicting the flux rates in these growth scenarios and comparing these fluxes rates with the experimental results. Hence, the R^2 score between the experimental and modelled fluxes were used to assess the model's goodness of fit.

For defining the Bayesian model, prior parameters have to be chosen for the enzymes. Li *et al.*[5] did this through a custom heuristic which was partially based on measured temperature optima and denaturation temperatures for enzymes. By training the model with the Sequential Monte Carlo based Approximate Bayesian calculation method, an estimate of the posterior distribution of parameter sets was found[5]. However, Li *et al.* did not systematically investigate the stability of this calculation method, nor whether it suffered from identifiability issues. Thus, it is unclear how metabolic flux results from the etcGEM can be interpreted.

In this paper, we investigated the stability of the Approximate Bayesian calculation method algorithm by choosing multiple different random seeds over different priors. We found that the Bayesian calculation method is inherently numerically unstable, and thus, is unable to provide reliable results given its model assumptions. To rectify this problem, we implemented an evolutionary search algorithm that is not built upon any assumption of structure of the underlying data. Finally, we improved the execution time of the software package by almost a factor 10, making it feasible to execute on smaller-scale computational infrastructures. However, for the available data there is an identifiability problem, in which solutions that equally match the experimental data still differ in terms of fluxes through key metabolic reactions. We believe that the evolutionary algorithm will resolve the identifiability problem if more experimental data, in particular data regarding internal fluxes, are included.

Results

Improvements to the running time of the algorithm

Running the Bayesian calculation method once for 500 iterations with the chosen hyperparameters consumed approximately 17,000 CPU hours (approximately corresponding to two weeks on a 48 core computer) using the implementation from Li *et al.* Profiling showed that the particle evaluation procedure (see Methods) was the performance bottleneck, and excess time consumption was caused by COBRApy's[13] internal routines to modify metabolic models prior to solving them. Hence, preparing the models for optimization consumed far more time than the optimization proper. For this reason, we modified the implementation to use the ReFramed package (`https://github.com/cdanielmachado/reframed`) for handling the genome-scale model. We benchmarked the two versions on a computer running Intel Core i7-8565U using a single core (Table 1). With our code improvements, the performance was boosted by factor of 8.5. As a consequence, the results of the Bayesian calculation method could be obtained the day after starting it when running on a compute server. Still, only about 20% of the particle evaluation time was spent on optimization, so improvements within the ReFramed package has the potential for increasing performance even more.

Assessing the stability of the Bayesian calculation method

In order to investigate the stability of the Bayesian calculation method for stochastic effects, we ran the Bayesian calculation method with two different random seeds on the three training datasets. These runs of the Bayesian calculation method are referred to as Bayesian simulation 1 and 2. Also, in addition to using the priors selected by Ref.[5], we created three randomized priors by permutation (see Methods for details) and repeated the process of assessing stability given these priors. Thus in total, we ran the Bayesian calculation method eight times, yielding eight different populations of estimated posterior distributions. The permuted priors yielded approximately the same rate of fitness convergence as the unpermuted priors (Figure 1 **A** and **B** and Supplementary Figure S1). Between simulation 1 and 2 for the same priors, the differences were negligible.

Having observed that all priors do indeed result in parameter sets with high fitness, we next investigated whether these solutions were similar. For this, we created a Principal Component Analysis[14] (PCA) plot of the parameter sets obtained under estimation (Figure 1 **C** and **D** with a more complete overview in Supplementary Figure S2). The estimated posterior distributions, defined as the collection of particles having $R^2 > 0.9$, were different for every simulation even though the

Table 1. Comparing benchmark results for a full evaluation of a particle using all three conditions (aerobic, anaerobic and chemostat). The numbers were averaged over 10 iterations.

Implementation	Overall time [s]	Percentage time in optimization [%]
Original version from Li *et al.*	106	0.837
Updated version with ReFramed	13.4	19.7

convergence properties were similar. This means that the Bayesian calculation method is unstable for all four priors and that there are identifiability issues causing the calculation method to converge at different locations in the parameter space.

We also discovered that the calculation of R^2 values for the chemostat dataset suffered from numerical instabilities unrelated to the Bayesian calculation method. For the same particle and software version, the Gurobi solver could sometimes judge the model infeasible given the parameters and sometimes it could find a feasible solution. However, given that a solution was found, the results were consistent up to expected numeric accuracy. We therefore suspected that the inherent numerical instability in calculation of the R^2 value for chemostat data in turn had made the Bayesian calculation method unstable. Given this concern, we also ran the Bayesian calculation method without the chemostat data. We only used the priors suggested by Li *et al.* and ran four different simulations with differing random seed. The results for this setup (Supplementary Figure S3 and S4) were similar to the simulations including the chemostat dataset. Hence, the Bayesian calculation method was unstable also when withholding the chemostat dataset.

Assessing enzyme-level stability of the Bayesian calculation method

To systematically study the phenotypic behaviour of particles in the estimated posterior distributions, we performed Flux Variability Analysis (FVA)[15], see Methods for more details. While it is impossible to lock down a specific flux distribution due to an infinite number of alternative optimal solutions, FVA uncovers the flux bounds of each individual reaction capable of supporting optimal metabolic behaviour, in this case maximizing the growth rate given the model parameters. We decided to focus the investigation of FVA results on six reactions that have important biochemical roles in the metabolic network:

- Pyruvate dehydrogenase: A key reaction in connecting glycolysis to the TCA cycle and fatty acid synthesis
- Fructose-bisphosphate aldolase: An intermediate reaction in glycolysis
- Ferrocytochrome-c:oxygen oxidoreductase: The oxygen consuming reaction in the respiratory electron transport chain
- Phosphoserine phosphatase: Reaction producing the amino acid serine from intermediates in the glycolysis
- Shikimate kinase: Intermediate reaction in synthesis of folate and aromatic amino acids (phenylalanine, tyrosine and tryptophan)
- Growth: The growth (biomass) reaction is included for comparison with the other reactions. The calculated growth should ideally be identical to the experimental ones, but some deviations occurred because the posterior particles did not in general provide a perfect fit to the data.

We chose to focus on three simulations; Bayesian simulation 1 and 2 with original priors and Bayesian simulation 1 with permuted prior set 1 (Figure 2). First, we observe that the flux through shikimate kinase had a narrow flux range and was highly coupled with growth. This reaction is a part of the shikimate pathway for producing folate and aromatic amino acids. We suspect that the resulting compounds have no functionality in the model except for being part of the biomass reaction. As no alternative pathways for producing these compounds exist, the flux through the shikimate kinase reaction is thus locked at a certain fraction of the growth rate. For the other reactions, there exists more variability among the solutions. Fructose-bisphosphate aldolase and Ferrocytochrome-c:oxygen oxidoreductase are for some particles used extensively, but in other cases not at all, still giving rise to approximately the same growth rates regardless. This means that the metabolic model uses alternative pathways depending on the choice of enzyme thermodynamic parameters. The fluxes for Pyruvate dehydrogenase and Phosphoserine phosphatase generally follow the trends of the growth curve, as for shikimate kinase. However, there are outliers deviating from this pattern, again most likely due to availability of alternative pathways.

The results for the anaerobic dataset (Supplementary Figure S5) were similar, except for the fact that there was no flux through the Ferrocytochrome-c:oxygen oxidoreductase reaction as there was *no* oxygen available to be consumed. We also ran FVA on the results omitting the chemostat dataset when running the Bayesian calculation method (Supplementary Figure S6 and S7). These results also showed large variability within simulation results and across simulations.

A bimodal toy example

Given the observation that the Bayesian calculation method returned different parameter sets with high fitness, we suspected that the fitness landscape of the temperature parameters was multimodal. Hence, we decided to test the Bayesian calculation method on a toy problem with two parameters to infer; x and y.

We defined the fitness function as:

$$R^2 = f(x,y) = \frac{1}{1 + \left((x-1)^2 + (y-1)^2\right) \cdot \left((x+1)^2 + (y+1)^2\right)} \tag{1}$$

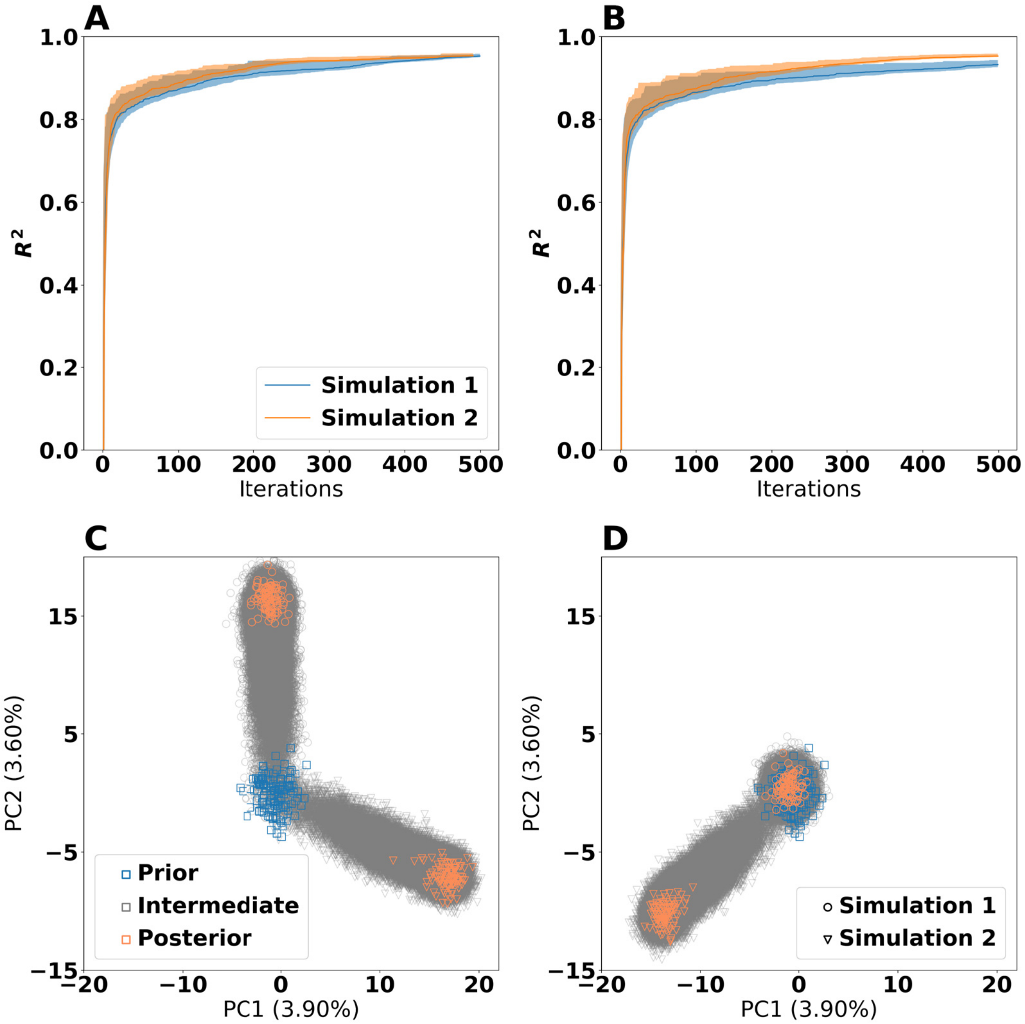


Figure 1. Comparison of the effect of prior and random seed on the Bayesian calculation method. Panel **A** and **B** show the training R^2 values for the unpermuted priors and permuted prior 1, respectively. An R^2 value of 1 corresponds to exact correspondence between the training data and the model predictions. The shaded regions indicate the 5th and 95th percentiles, whereas the solid lines indicate the median (50th percentile). Panel **C** and **D** show Principal Component Analysis (PCA) plots of the parameter sets from the unpermuted priors and permuted prior 1, respectively. Each point is a candidate parameter set. The prior points are the ones which served as a starting point for the calculation method, the estimated posterior points are the ones which had $R^2 > 0.9$, whereas all other points are intermediate points stemming from the simulations. The axes are identical for both panels and use the same ordination, making the panels directly comparable.

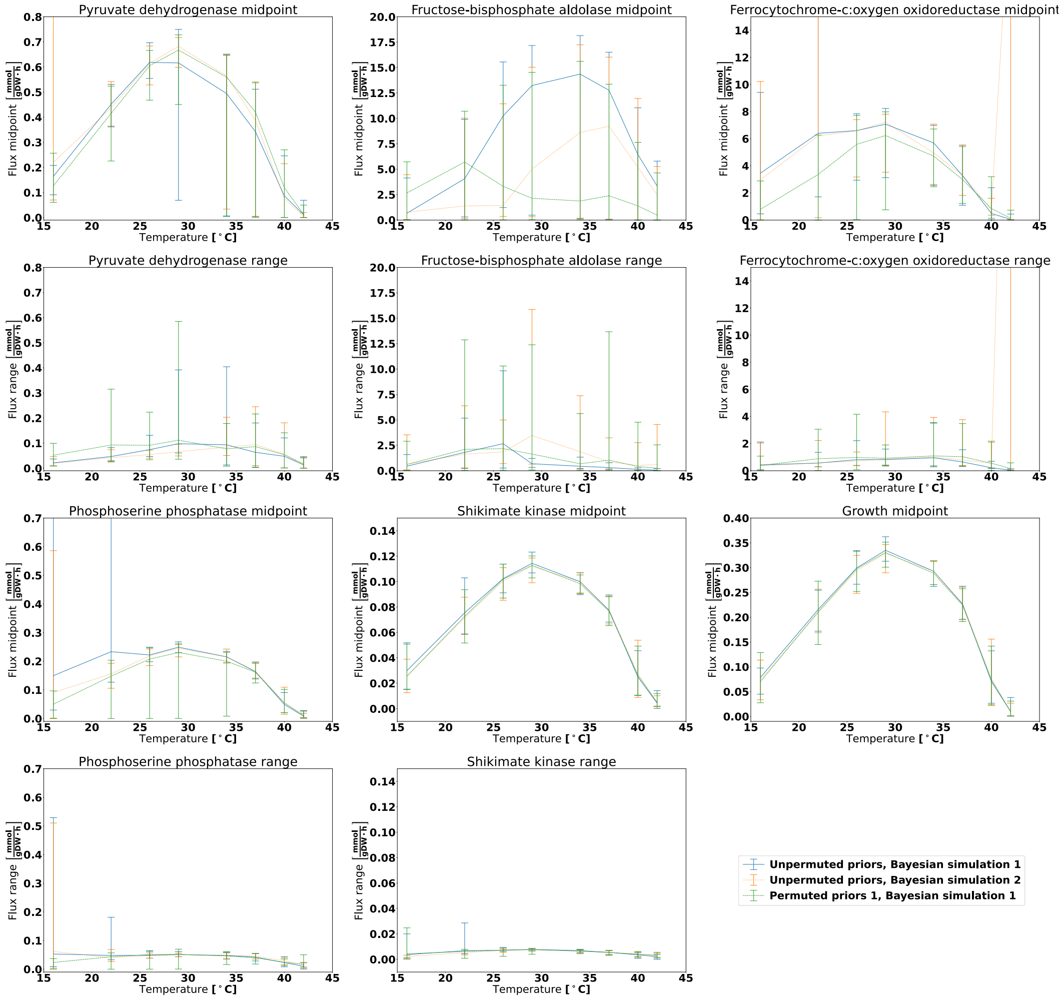


Figure 2. FVA analysis for the aerobic dataset for six reactions and varying temperature when using estimated posterior distributions obtained for the Bayesian calculation method. The midpoint panels show the FVA flux midpoint, this is: The average of the maximum and minimum attainable flux given the optimization objective. The range panels show the absolute difference between the maximum and minimum flux. The lines denote the mean midpoint or range value, whereas the error bars span from the lowest to the highest observed value. The growth reaction is included for reference, and it will always display an FVA range of zero as it is the optimization target.

This fitness function assumes values from 0 to 1 where $R^2 = 1$ is only attained in the global maxima $(x,y) = (1,1)$ and $(x,y) = (-1,-1)$. The function does not have any additional extrema, but it has a saddle point at $(x,y) = (0,0)$.

As our priors, we assumed that x and y were independent and identically normally distributed with mean zero and standard deviation 0.2. This is:

$$x,y \sim N\left(0,0.2^2\right) \tag{2}$$

Due to the symmetry of this problem, the true posterior distribution is equally centred around the two optima and we would therefore expect the Bayesian calculation method to replicate this symmetric distribution. We ran the Bayesian calculation method on this toy problem with a population size of 32 over 200 iterations with four replicates having different random seeds. When plotting the final generation of particles (Figure 3 **A**), we realized that the Bayesian calculation method clustered all of its points in a very small space close to one of the optima. Which of these two optima this was, varied based on the random seed, but the same simulation never yielded points near both of the optima. In addition, only one of the four simulations actually reached an optimum, whereas the three other simulations suffered from genetic bottlenecking, meaning that the variability in the population of particles disappeared and caused premature convergence[16].

Evolutionary algorithm

Given the instability of Bayesian calculation method and its inability to cope with multimodal fitness landscapes, we constructed an evolutionary algorithm[17–20] as an alternative for inferring parameters. More specifically, we used a variation of CrowdingDE[21] which is designed to find alternative optima in a multimodal distribution[22] (see Methods for details). Our choice of an evolutionary method was motivated by how an evolutionary algorithm searches the parameter space and its ability to combine existing solutions to create improved solutions[23]. CrowdingDE has two major hyperparameters, the scaling factor F and the crossover probability CF which both determine how crossover between individuals is done.

For testing the performance of the evolutionary algorithm, we used the previously mentioned toy example with the same priors. As for the Bayesian calculation method, the population size was set to 32 and four replicate simulations were run for 200 generations. The scaling factor was set to 0.5, the crossover probability was 0.5 and 16 new children were born per generation. (Figure 3 **B**) As opposed to the Bayesian calculation method, the evolutionary algorithm diverged into two subpopulations closing on the two optima, each consisting of approximately half the individuals. This shows that the chosen evolutionary algorithm is able to find multiple optima during the same simulation. Note however, that the two subpopulations had a considerable variability after 200 generation and thus did not suffer from genetic bottlenecking.

Encouraged by the results from the toy example, we applied the evolutionary algorithm on the problem of finding enzyme parameters. We used the same prior as suggested by Li *et al.* and chose to discard the chemostat dataset for these simulations due to its associated instability. The population size was set to 128, the children born per generation was 64 and the simulation were run for 1000 generations. We varied the hyperparameters scaling factor F and crossover probability CF. Simulations were conducted in replicate, meaning that for any combination of scaling factor and crossover probability, two simulations were executed with differing random seeds.

All simulations produced particles with $R^2 > 0.9$ by 1000 generations (Supplementary Figure S8). However, there were considerable variability in fitness among the particles in each simulation and only in two of the simulations (the ones with $F = 0.5$ and $CR = 0.99$), the median population fitness exceeded $R^2 = 0.9$. Still, having a large span of fitness values inside the same simulation is not a major disadvantage per se as one can selectively pick the individuals with high R^2. At the same time, having a large variability among the solutions is preferable to avoid genetic bottlenecking. The choice of hyperparameters also affected the rate of convergence. From what we can assess, $F = 0.5$ and $CR = 0.99$ yielded the best effect in this case (Figure 4 **A**), and we proceeded with the results from this hyperparameter combination. However, this does not necessarily mean that better choices for the control hyperparameter do not exist, nor does it mean that these hyperparameter values are appropriate given different experimental data sets[24].

We further extracted the particles from these two simulations having $R^2 > 0.98$ and created a PCA ordination (Figure 4 **B**). From this ordination, we observed that the particles ended up in distinct clusters which we believe to be local optima of the fitness function, similar to the situation in Figure 3 **B**. Furthermore, hierarchical clusters (Supplementary Figure S9) display the particles in these discrete optima. Each simulation found a number of these optima, but the same optimum was not found by both of the simulations. Still, there is no evident distinction between the populations from the two simulations as a whole. This observation is most likely due to the fact that there are so many optima that it is not feasible for the evolutionary algorithm to find all of them.

FVA analysis of the same particles (Figure 5 and Supplementary Figure S10) revealed that the populations of particles from the two simulations did not show any large systematic differences. However, *within* the same population of solutions, there were considerable variation and outliers. This points to usage of alternative pathways which the experimental data could

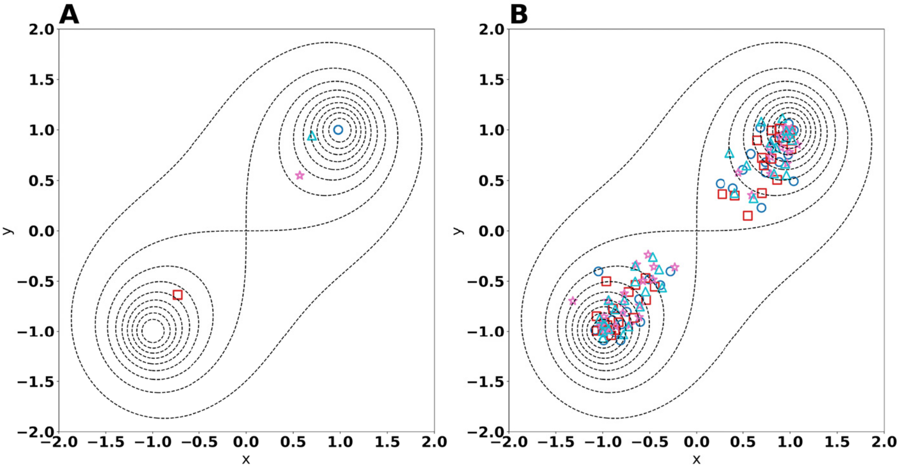


Figure 3. Results of parameter inference on the toy example. Panel **A** and **B** show the final generation of particles for the Bayesian calculation method and evolutionary algorithm, respectively. Each point represents an individual in the final population. Each shape and colour (each shape is associated to exactly one colour) represents one of the four replicate simulation. The global fitness optima of the problem are the points $(-1,-1)$ and $(1,1)$, and are marked by dotted contours. For Panel **A**, the particles from each simulation are so close that they are visually indistinguishable and therefore appear as a single point.

not lock down based on the growth rates alone. The results for Ferrocytochrome-c:oxygen oxidoreductase under aerobic condition illustrate this case; at temperatures below 37 °C there were moderate levels of agreement between the different particles. However, at higher temperatures, the coupling disappeared, meaning that alternative pathways could take over and attain approximately the same fitness.

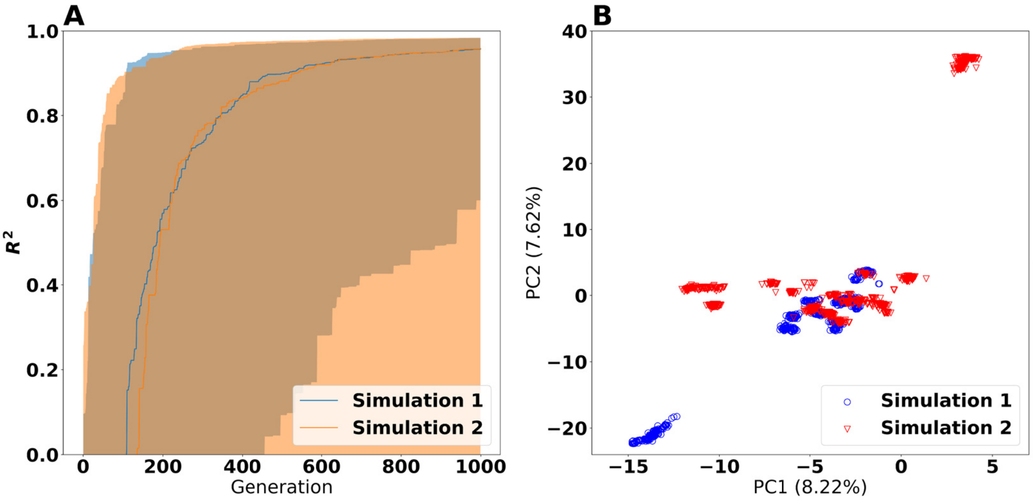


Figure 4. Results from the evolutionary algorithm with $F = 0.5$ and $CR = 0.99$. Panel **A** shows the training R^2 values. An R^2 value of 1 corresponds to exact correspondence between the training data and the model predictions. The shaded regions indicate the 5^{th} and 95^{th} percentiles, whereas the solid lines indicate the median (50^{th} percentile). Panel **B** shows the Principal Component Analysis (PCA) plot of the particles having $R^2 > 0.98$.

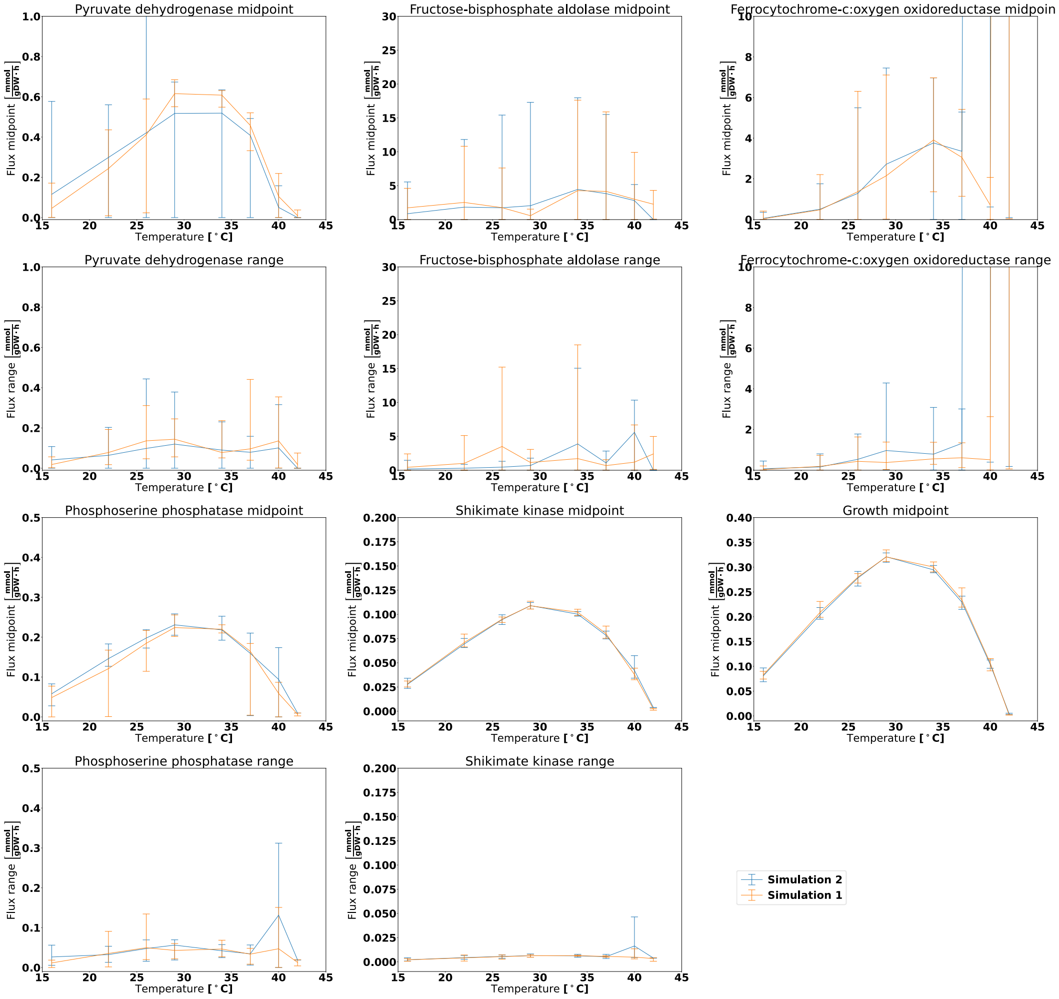


Figure 5. FVA analysis on the results from the evolutionary algorithm under aerobic conditions. The particles selected for this analysis stem from the two simulations with $F = 0.5$ and $CR = 0.99$, considering only the particles with $R^2 > 0.98$. The midpoint panels show the FVA flux midpoint, ie. the average of the maximum and minimum attainable flux given the optimization objective. The range panels show the absolute difference between the maximum and minimum flux. The lines denote the mean midpoint or range value, whereas the error bars span from the lowest to the highest observed value. The growth reaction is included for reference, and it will always display an FVA range of zero as it is the optimization target.

Discussion

We observed that computing R^2 values for the chemostat dataset resulted in numerical instability, while this problem was not present for the other two datasets. This is likely due to the the three-stage procedure of first locking the growth rate of the model to the dilution rate, then minimizing glucose uptake and setting it as a constraint for the model, before finally minimizing the protein usage and then reporting the fluxes. Even if the resulting problem is mathematically solvable, the sharp constraints still cause problems for the Gurobi LP solver, which for the same particle sometimes managed to find a feasible solution to the problem, and sometimes not. Potentially, this challenge could be mitigated by reformulating the optimization problem to obtain a growth rate and glucose uptake rate as close as possible (but not necessarily equal) to the target values[25]. Also, Gurobi supports directly setting lexicographic objectives solved in sequence, an approach which hopefully does not possess the aforementioned problem.

Our results point out that the outcome of the Bayesian calculation method is unstable and its result indeed depends on the choice of random seed. It is important to note that this instability has nothing to do with the instability of R^2 computations for the chemostat dataset. As illustrated by the toy example, even simple bimodal fitness functions can cause the Bayesian simulations to suffer from genetic bottlenecking and failing to estimate the posterior distribution. Given the strong indications that the fitness landscape for the thermodynamic enzyme parameters is multimodal, our results imply that the Bayesian calculation method failed to converge to the true theoretical posterior distribution. This is a serious problem, as the usual statistical interpretations of the Bayesian approach will lead to erroneous conclusions if applied to the results.

Each Bayesian simulation converged to a point cloud with no apparent higher-dimensional structure. This observation makes sense, considering that the Bayesian calculation method creates new particles by sampling each parameter independently from a normal distribution where the mean and standard deviation is determined by the past generation. Thus, the estimated posterior points are likely to cluster in a high-dimensional cloud where the density of each parameter is normally distributed and the spatial density is the product of the marginal densities of the parameters. Hence, the Bayesian calculation method assumes an unimodal posterior distribution and will therefore fail when applied to a problem with a multimodal posterior distribution. The failure of the chosen Bayesian calculation method does not imply that a Bayesian approach for etcGEMs necessarily is bad, but it would require considerable refinements to the Bayesian calculation method to work with multimodal posterior distributions[26,27].

The evolutionary algorithm produced results which were more robust to the choice of random seed. Simulations which were run for the same combination of hyperparameters had similar development of R^2 values. Although the choice of hyperparameters had large effects on the rate of convergence, we were only able to evaluate a small number of hyperparameters due to the large computational burden associated with running the evolutionary algorithm on the problem in question. We may therefore have missed out on more favourable hyperparameter combinations. Consequently, we see potential in using a self-adaptive Differential Evolution algorithm which does not need predetermining niching hyperparameters[28].

For our preferred hyperparameter set $F = 0.5$ and $CR = 0.99$, we obtained a large variety of solutions and fitness values. The particles with the highest fitness values were dispersed among many distinct optima in the fitness landscape. Due to the high number of optima being present, the evolutionary algorithm was unable to find all of them in a single simulation. Yet, unlike the Bayesian calculation method, the evolutionary algorithm did not appear to have any spatial bias with respect to where these optima were localized in the parameter space.

The great variety of different particles of the evolutionary algorithm is not a weakness per se, but rather an indication of the desired feature of exploring the parameter space and avoiding genetic bottlenecking. Still, the results reveal that the identifiability of the parameter inference problem is poor. As revealed by FVA, there exists large variability between the particles with high fitness with respect to the internal fluxes. Hence, the choice of metabolic pathways for yeast appear to be sensitive to the thermodynamic properties of the enzymes even if the growth rate, metabolic network model topology, and external conditions were kept the same. This phenotypic sensitivity to thermodynamic properties of the metabolic enzymes may be a possible explanation for cellular metabolic heterogeneity observed in yeast cultures[29,30]. Still, we believe that the main source for the lack of identifiability is a result of the external measurements being insufficient to account for the inner workings of the yeast cell. In this respect, we believe that measurements of proteomics[31] and fluxomics[32] will help narrow down the solution space and provide more accurate predictions of metabolic behaviour. For instance, if we got direct measurement of a reaction showing high variability between particles with our present data, such as Fructose-bisphosphate aldolase, we would be able to rule out the particles not satisfying the measured fluxes of this reaction.

The application of such refined approaches strongly suggests that a strategy for including different kinds of data is needed. Heckmann *et al.* inferred apparent k_{cat} values from large-scale proteomics and metabolomics data on a genome-scale level using gene knock-out strains and machine learning. As a result, more accurate prediction of *in vivo* fluxes were obtained compared to using k_{cat} values measured *in vitro*[33,34]. We believe that this strategy can be adapted to the current etcGEM framework and provide efficient integration of different kinds of data, thus allowing narrowing down the solution space and at least partially alleviate the problem of identifiability.

With our software improvements, running the inference algorithm is much faster than with the original implementation, yet the problem is still so computationally expensive that workstation or server-grade hardware is required. This computational burden is likely to increase when incorporating more experimental data to calibrate the parameters. Therefore, systematic improvements should be implemented in the framework to minimize unnecessary overhead in order to hold the computational burden at a manageable level.

Methods

Evaluation method for particles

A parameter set, referred to as a particle, is a collection of the three parameters T_m, T_{opt} and $\Delta C_p^{\ddagger}$ for the 764 metabolic enzymes of the ecYeast7.6 model. The goodness of fit for a particle was evaluated by the framework created by Li *et al.*[5]. In brief, this evaluation procedure acted as a black box taking a particle as input and returning an overall R^2 value. In the assessment, each particle was matched against experimental data from three experiments with *Saccharomyces cerevisiae*. These were:

- The aerobic dataset[10] measuring growth rates of yeast at 8 temperatures between 16 °C and 42 °C under aerobic batch fermentations.
- The anaerobic dataset[11] measuring growth rates of yeast at 13 temperatures between 5 °C and 40 °C under anaerobic batch fermentation. However, in our particle assessment, only 8 of the temperatures were used.
- The chemostat dataset[12] measuring exchange fluxes of carbon dioxide, ethanol and glucose at 6 temperatures between 30 °C and 38.5 °C in aerobic chemostats.

The parameters were used to adjust the effective catalytic rate $k_{cat}(T)$ (which includes denaturation) for each enzyme in the model and at each temperature T. In addition, the model's non-growth associated ATP maintenance (NGAM) was also adjusted according to the temperature. Further details are published in Li *et al.*[5].

For the aerobic and anaerobic datasets, the model's temperature-dependent parameters were tuned, and fluxes were predicted by Flux Balance Analysis (FBA) calculations at the specific temperatures, and the biomass (growth) function was set as the objective. For each of the aerobic and anaerobic datasets, the growth rates predicted by the model were compared with the experimental growth rates, and an R^2 value was reported, yielding R^2_{rae} and R^2_{ran}.

For the chemostat dataset, the procedure was somewhat different. Here, the temperature-dependent parameters were tuned (as with the other datasets) before the growth rate of the model was locked to the dilution rate of the chemostat. Thereafter, the model was optimized for minimum glucose uptake, and this uptake flux value was set as a constraint for the model. Finally, the model was optimized for minimum protein pool usage, and exchange fluxes of ethanol, carbon dioxide, and glucose were recorded. Once the three fluxes for all temperatures in the chemostat dataset were recorded, these values were compared to the experimental ones, and R^2_{chemo} was determined.

The overall R^2 for all datasets was calculated as the arithmetic mean of R^2_{rae}, R^2_{ran}, and R^2_{chemo}. This value was then returned as the final result of the evaluation procedure. The higher the R^2 value, the higher correspondence exists between the modelled solutions and the experimental results, where $R^2 = 1$ corresponds to the highest achievable fitness.

We optimized the evaluation procedure to use the ReFramed package instead of COBRApy[13] in order to reduce overhead related to modifying models. However, the results generated by our modified particle evaluation approach should be identical to the results generated by the original code by Li *et al.* for all simulations.

Approximate Bayesian calculation method

The framework and code for the Sequential Monte Carlo based Approximate Bayesian calculation method was taken directly from Li *et al.*[5], and we used the same hyperparameters as in the original publication. For seeding the calculation method, priors were needed for the values of T_{opt}, T_m and $\Delta C_p^{\ddagger}$ for each enzyme. We used the same priors as Li *et al.*. These priors considered the distribution of each parameter x_i to be normally distributed

$$x_i \sim N(\mu_i, \sigma_i), \tag{3}$$

and the marginal distribution of each parameter was independent. Some simulations were also run with permuted priors. This meant that the labels of the enzymes were randomly shuffled and each enzyme thus got the values of T_{opt}, T_m and $\Delta C_p^{\ddagger}$ of another enzyme before this new prior was used to seed the Bayesian calculation method. The computations were run for 500 iterations. The population size at the end of each iteration was 100. We generated 128 new particles for each iteration and evaluated them according to the description in the previous section. The new particles were generated by computing the mean and standard deviation for each parameter of the particle population and sampling new parameters from a normal distribution with the aforementioned mean and standard deviation. When creating new particles, we nevertheless made sure that they obeyed the constraint $T_m > T_{opt} > 0\,\mathrm{K}$. If this constraint was violated, the parameter was resampled. Selection of the particles was implemented through truncation selection, meaning that the 100 best particles from the previous iteration were passed to the next iteration while the rest were discarded.

Evolutionary algorithm

The evolutionary algorithm used in this paper for fitting enzyme parameters is based on the existing CrowdingDE[21] algorithm and was written from the ground up. Individuals in the evolutionary process were the parameter set particles discussed earlier. The population size for each iteration (generation) was set to 128. The initial 128 individuals of the population were generated by sampling from the same priors as those used by Li *et al.* The algorithm was run for 1000 generations. Each generation consisted of the following steps carried out in sequence:

- **Generation of children:** At the beginning of each generation, 64 children were created as a weighted difference of parent individuals. For each child, three parents P_1, P_2 and P_3 were selected at random from the population without replacement, ensuring that the parents were unique. We refer to P_1 as the primary parent and P_2, P_3 as the crossover parents. At first, each parameter for the child was initialized to the corresponding value for the primary parent, this is:

  $$\mu_{child,k} = \mu_{P_1,k} \quad \text{for } 1 \leq i \leq M, \tag{4}$$

  where M denotes the total number of enzyme parameters. For crossover, a random integer $i \in [1, M]$ was uniformly drawn. A counter variable j was thereafter initiated to zero and the following procedure was repeated: A random number $r \in (0,1)$ was uniformly drawn. If $j \geq M$ or $r > CR$, where CR is referred to as the *crossover probability*, crossover was cancelled and the algorithm advanced to commence the generation of the next child. Otherwise, crossover was performed on parameter $k = i + j \mod M$:

  $$\mu_{child,k} = \mu_{P_1,k} + F\left(\mu_{P_2,k} - \mu_{P_3,k}\right), \tag{5}$$

  $and\,j$ was incremented by one and the procedure was repeated for the next enzyme parameter. Updates to enzyme parameters which violated the constraints $T_m > T_{opt} > 0\,\mathrm{K}$ were reverted.

- **Evaluation of children:** Each child particle was evaluated through the same procedure as mentioned above and their respective R^2 values were reported. In our case, we withheld the chemostat dataset and therefore only averaged the R^2_{rae} and R^2_{ran} values.

- **Replacement:** For each child a generated in the same generation, the normalized parameter-space distance from the child to the individuals in the current population was calculated:

  $$D_{a,i} = \sum_{k=1}^{M} \left(\frac{\mu_{a,k} - \mu_{i,k}}{\sigma_k} \right), \tag{6}$$

  where $\mu_{a,k}$ and $\mu_{i,k}$ are values for the enzyme parameter k for the two individuals, M is the total number of enzyme parameters and σ_k is the empirical standard deviation of enzyme parameter k in the population at the end of the previous generation. The closest individual b to the child a was then chosen as:

  $$b = \arg\min_i D_{a,i}. \tag{7}$$

  If the individual (the child) a had a higher fitness than b, i.e. $R^2_a > R^2_b$, then b was discarded from the population and replaced with a. Otherwise, b was kept in the population and the child a was discarded.

The two tuning parameters F (scaling factor) and CF (crossover probability) were tested with values F$\in \{0.5, 1.0\}$ and CF$\in \{0.9, 0.99, 0.999\}$. For each combination of these parameters, two replicate simulations were conducted with different random seeds.

Ordinations

We used Principal Component Analyses (PCA) from scikit-learn (version 1.0)[35] to create ordinations of particles. The values for each parameter were standardized, subtracting the mean and dividing by the population standard deviation before ordination. The means and standard deviations were computed across all particles present in the ordination in question. As a result, the presented ordinations in Figure 1 and Supplementary Figure S2 are comparable across the panels in the same figure. For the Bayesian calculation method (Figure 1 and Supplementary Figure S4), all points generated during the simulations were included to make the ordinations. However, for the evolutionary algorithm (Figure 4), only the points attaining $R^2 > 0.98$ from the two simulations with $F = 0.5$ and $CR = 0.99$ were included in the ordination.

Flux variability analysis

Flux variability analysis (FVA)[15] was conducted on particles having $R^2 > 0.9$ (for the Bayesian calculation method) or $R^2 > 0.98$ (for the evolutionary algorithm) in order to ensure that these particles had high fitness. From each simulation, 20 particles were sampled randomly from the particles satisfying the aforementioned thresholds. For each sampled particle, FVA was run for the aerobic and anaerobic datasets across the same temperatures used for determining R^2 in the parameter fitting process. For the chemostat dataset, the numeric instability was too large to produce reliable results, and the chemostat dataset was thus discarded for the analyses. The temperature and the parameters of the particles were first used to fix the effective k_{cat} values. Subsequently, the metabolic model was optimized for maximal growth, and the lower bound of the growth reaction was locked to the obtained growth rate. Thereafter, maximum and minimum fluxes through each reaction in the model were found given the constraints. Instead of using the maximum and minimum fluxes directly, we converted them into flux midpoint (average of maximum and minimum) and flux ranges (absolute difference between maximum and minimum). Results with missing or infinite values were removed.

All flux ranges and flux midpoints with the same combination of simulation, reaction, dataset, and temperature were aggregated to give the mean, minimum and maximum values. Usually, there were 20 such values for each combination, as 20 particles were sampled for each of these combinations. However, this number could be smaller due to removal of missing and infinite values.

Hierarchical clustering

Agglomerative hierarchical clustering[36] was conducted on the particles from the evolutionary algorithm with $F = 0.5$ and $CR = 0.99$ that satisfied $R^2 > 0.98$, as for PCA and FVA. We standardized each parameter value by subtracting the mean value and divided by the standard deviation among the selected particles and also calculated pairwise Euclidean distances. Hierarchical clustering was conducted by single (minimum distance) linkage in order to put emphasis on the detection of discontinuities between clusters of particles. The results were presented in a dendrogram showing the particles as leafs. The branches of the dendrogram were coloured according to which simulation the downstream branches corresponded to. Branches containing particles from both simulations were left uncoloured (gray).

References

1. van der Kamp, M. W. *et al.* Dynamical origins of heat capacity changes in enzyme-catalysed reactions. *Nat. Commun.* **9**, 1177, DOI: 10.1038/s41467-018-03597-y (2018).

2. Hobbs, J. K. *et al.* Change in heat capacity for enzyme catalysis determines temperature dependence of enzyme catalyzed rates. *ACS Chem. Biol.* **8**, 2388–2393, DOI: 10.1021/cb4005029 (2013).

3. Birch, A. N., Petersen, M. A. & Åse S. Hansen. The aroma profile of wheat bread crumb influenced by yeast concentration and fermentation temperature. *LWT - Food Sci. Technol.* **50**, 480–488, DOI: https://doi.org/10.1016/j.lwt.2012.08.019 (2013).

4. Molina, A. M., Swiegers, J. H., Varela, C., Pretorius, I. S. & Agosin, E. Influence of wine fermentation temperature on the synthesis of yeast-derived volatile aroma compounds. *Appl. Microbiol. Biotechnol.* **77**, 675–687, DOI: 10.1007/s00253-007-1194-3 (2007).

5. Li, G. *et al.* Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *Nat Commun* **12**, 190, DOI: 10.1038/s41467-020-20338-2 (2021).

6. Sánchez, B. J. *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* **13**, 935, DOI: 10.15252/msb.20167411 (2017).

7. Bekiaris, P. S. & Klamt, S. Automatic construction of metabolic models with enzyme constraints. *BMC bioinformatics* **21**, 19, DOI: 10.1186/s12859-019-3329-9 (2020).

8. Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *The Biochem. journal* **238**, 781–6 (1986).

9. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. opinion biotechnology* **14**, 491–6 (2003).

10. Caspeta, L. & Nielsen, J. Thermotolerant yeast strains adapted by laboratory evolution show trade-off at ancestral temperatures and preadaptation to other stresses. *mBio* **6**, e00431, DOI: 10.1128/mBio.00431-15 (2015).

11. Zakhartsev, M., Yang, X., Reuss, M. & Pörtner, H. O. Metabolic efficiency in yeast saccharomyces cerevisiae in relation to temperature dependent growth and biomass yield. *J Therm Biol* **52**, 117–29, DOI: 10.1016/j.jtherbio.2015.05.008 (2015).

12. Postmus, J. *et al.* Quantitative analysis of the high temperature-induced glycolytic flux increase in saccharomyces cerevisiae reveals dominant metabolic regulation. *J Biol Chem* **283**, 23524–32, DOI: 10.1074/jbc.M802908200 (2008).

13. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. Cobrapy: Constraints-based reconstruction and analysis for python. *BMC Sys Bio* **7**, 74, DOI: 10.1186/1752-0509-7-74 (2013).

14. Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572, DOI: 10.1080/14786440109462720 (1901). https://doi.org/10.1080/14786440109462720.

15. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. engineering* **5**, 264–276, DOI: 10.1016/j.ymben.2003.09.002 (2003).

16. Katoch, S., Chauhan, S. S. & Kumar, V. A review on genetic algorithm: past, present, and future. *Multimed. Tools Appl.* **80**, 8091–8126, DOI: 10.1007/s11042-020-10139-6 (2021).

17. Maia, P., Rocha, I. & Rocha, M. Identification of robust strain designs via tandem pfba/lmoma phenotype prediction. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '17, 1661–1668, DOI: 10.1145/3067695.3082542 (Association for Computing Machinery, New York, NY, USA, 2017).

18. Patil, K. R., Rocha, I., Förster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinforma.* **6**, 308, DOI: 10.1186/1471-2105-6-308 (2005).

19. Rocha, M. *et al.* Natural computation meta-heuristics for the in silico optimization of microbial strains. *BMC Bioinforma.* **9**, 499, DOI: 10.1186/1471-2105-9-499 (2008).

20. Alter, T. B., Blank, L. M. & Ebert, B. E. Genetic optimization algorithm for metabolic engineering revisited. *Metabolites* **8**, 33, DOI: 10.3390/metabo8020033 (2018).

21. Thomsen, R. Multimodal optimization using crowding-based differential evolution. In *Proceedings of the 2004 Congress on Evolutionary Computation (IEEE Cat. No.04TH8753)*, vol. 2, 1382–1389 Vol.2, DOI: 10.1109/CEC.2004.1331058 (2004).

22. Wong, K.-C., Wu, C.-H., Mok, R. K., Peng, C. & Zhang, Z. Evolutionary multimodal optimization using the principle of locality. *Inf. Sci.* **194**, 138–170, DOI: https://doi.org/10.1016/j.ins.2011.12.016 (2012). Intelligent Knowledge-Based Models and Methodologies for Complex Information Systems.

23. Yu, X. & Gen, M. *Multimodal Optimization*, 165–191 (Springer London, London, 2010).

24. Eiben, A., Hinterding, R. & Michalewicz, Z. Parameter control in evolutionary algorithms. *IEEE Transactions on Evol. Comput.* **3**, 124–141, DOI: 10.1109/4235.771166 (1999).

25. Segrè, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.* **99**, 15112–15117, DOI: 10.1073/pnas.232349399 (2002). https://www.pnas.org/doi/pdf/10.1073/pnas.232349399.

26. Stephens, M. Dealing with multimodal posteriors and non-identifiability in mixture models (1999).

27. Yao, Y., Vehtari, A. & Gelman, A. Stacking for non-mixing bayesian computations: The curse and blessing of multimodal posteriors. *J. Mach. Learn. Res.* **23**, 1–45 (2022).

28. Jiang, R., Zhang, J., Tang, Y., Feng, J. & Wang, C. Self-adaptive de algorithm without niching parameters for multi-modal optimization problems. *Appl. Intell.* **52**, 12888–12923, DOI: 10.1007/s10489-021-03003-z (2022).

29. Pinheiro, S., Pandey, S. & Pelet, S. Cellular heterogeneity: yeast-side story. *Fungal Biol. Rev.* **39**, 34–45, DOI: https://doi.org/10.1016/j.fbr.2021.11.005 (2022).

30. MacGillivray, M. *et al.* Robust analysis of fluxes in genome-scale metabolic pathways. *Sci. reports* **7**, 268, DOI: 10.1038/s41598-017-00170-3 (2017).

31. Großeholz, R. *et al.* Integrating highly quantitative proteomics and genome-scale metabolic modeling to study ph adaptation in the human pathogen enterococcus faecalis. *NPJ systems biology applications* **2**, 16017–16017, DOI: 10.1038/npjsba.2016.17 (2016).

32. Winter, G. & Krömer, J. O. Fluxomics – connecting 'omics analysis and phenotypes. *Environ. Microbiol.* **15**, 1901–1916, DOI: https://doi.org/10.1111/1462-2920.12064 (2013). https://sfamjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.12064.

33. Heckmann, D. *et al.* Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc Natl Acad Sci U S A* **117**, 23182–23190, DOI: 10.1073/pnas.2001562117 (2020).

34. Li, Z. *et al.* High-throughput and reliable acquisition of in vivo turnover number fuels precise metabolic engineering. *Synth. Syst. Biotechnol.* **7**, 541–543, DOI: https://doi.org/10.1016/j.synbio.2021.12.006 (2022).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

36. Gan, G., Ma, C. & Wu, J. *7. Hierarchical Clustering Techniques*, 109–149 (Society for Industrial and Applied Mathematics, 2007). https://epubs.siam.org/doi/pdf/10.1137/1.9780898718348.ch7.

Acknowledgements

Thanks to Gang Li for troubleshooting technical problems with the simulations. Thanks to Ingelin Steinsland for consulting on Bayesian statistics. J.P.P. and E.A. are grateful to ERA CoBioTech project CoolWine and Norwegian Research Council grant 283862 for funding.

Author contributions statement

E.A. and J.P.P. conceived the project. J.P.P did the analysis, created visualisations and wrote the first draft of the paper. All authors contributed to and accepted the final version of the paper.

Additional information

Availability of source code and data

The source code for the analysis and visualisation is available at GitHub (https://github.com/AlmaasLab/BayesianGEM). Also, an archived copy of the repository is available at Figshare (https://doi.org/10.6084/m9.figshare.21436623). Simulation of genome-scale models has been carried out with ReFramed (https://github.com/cdanielmachado/reframed) using Gurobi (Gurobi Optimization, LLC) as solver. The data required for reproducing the figures in the main article and in the Supplementary material is available as a Figshare repository (https://doi.org/10.6084/m9.figshare.21436668).

Competing interests

The authors declare that there are no competing interests.

Paper IV

Genome-scale metabolic models reveal determinants of phenotypic differences in non-*Saccharomyces* yeasts

Jakob Peder Pettersen, Sandra Castillo, Paula Jouhten, Eivind Almaas

To be submitted

RESEARCH

Genome-scale metabolic models reveal determinants of phenotypic differences in non-Saccharomyces yeasts

Jakob P. Pettersen[1*]
, Sandra Castillo[3]
, Paula Jouhten[4]
and Eivind Almaas[1,2]

*Correspondence:
jakob.p.pettersen@ntnu.no
[1]Department of Biotechnology and Food Science, NTNU- Norwegian University of Science and Technology, Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: Use of alternative non-*Saccharomyces* yeasts in wine and beer brewing has gained more attention the recent years. This is both due to the desire to obtain a wider variety of flavours in the product and to reduce the final alcohol content. Given the metabolic differences between the yeast species, we wanted to account for the differences by using *in silico* models.

Results: We created and studied genome-scale metabolic models of five different non-*Saccharomyces* species. These were: *Metschnikowia pulcherrima*, *Lachancea thermotolerans*, *Hanseniaspora osmophila* and *Kluyveromyces lactis*. It was found that *Metschnikowia pulcherrima*, compared to the other species conducts more respiration and thus produces less fermentation products, a finding which agrees with experimental data. Complex I of the electron transport chain was predicted to be present in the model of *Metschnikowia pulcherrima*, but absent in the others. The importance of Complex I vanished when incorporating enzyme constraints, but still *Metschnikowia pulcherrima* consumed glucose more efficiently by respiration.

Conclusions: The results suggest that Complex I in the electron transport chain is a key differentiator between *Metschnikowia pulcherrima* and the other yeasts considered. Yet, there likely exists other features in the metabolic network which differentiates *Metschnikowia pulcherrima* from the other yeasts. Further experiments should be conducted to confirm the *in vivo* effect of Complex I in *Metschnikowia pulcherrima* and its respiratory metabolism.

Keywords: *Metschnikowia pulcherrima*; sMOMENT; genome-scale models; electron transport chain; Complex I; yeast; metabolic modelling; automated reconstructions; enzymatic constraints; alternative pathways; decFBA

Background

In recent years, there has been increased interest in using alternative non-*Saccharomyces* yeasts for beer and wine brewing[1, 2, 3, 4, 5]. In general, there are two motivation for the pull towards fermentation by non-*Saccharomyces* strains: First, wine producers want to reduce the resulting alcohol content. Second, some brewers want more complex aroma compounds in the product and thus try to mimic the rich taste of spontaneously fermented products. In this paper, our focus will be on the desire for lower alcohol content.

Climate change has resulted in warmer and sunnier summers in wine producing regions, leading to higher sugar content in ripe grapes. When the must of high sugar grapes are fermented, this leads to higher alcohol content in the product. As a consequence, alcohol content of wine has risen by approximately 1% alcohol by volume each decade since the 1980s in some wine producing regions[6, 7]. Whereas approaches such as delusion of the must, earlier harvesting of the grapes and post-fermentation removal of alcohol can bring down the resulting alcohol content, such approaches come at the expense of deterring oenological qualities and breaking with established standards for wine brewing[8].

In order to create wines with reduced alcohol content without losing the rich flavours, aeration during the fermentation process has been proposed as a solution. Unfortunately, using this approach with the canonical wine yeast *Saccharomyces cerevisiae* has proven to be challenging. First of all, the most common strains of *Saccharomyces cerevisiae* are Crabtree positive, meaning that they ferment glucose to ethanol even when oxygen is available[9, 10, 11]. Furthermore, aeration often leads to production of acetic acid which is considered an undesired by-product[11]. On the other hand, many non-*Saccharomyces* yeasts have weaker tendencies for acetate production and Crabtree effects[12, 13, 2]. Using non-*Saccharomyces* yeasts alone is usually not a good option because stuck fermentations and poor resulting wine quality is often the result. However, sequential fermentation with non-*Saccharomyces* strains followed by inoculation by *Saccharomyces cerevisiae* has proven to be a viable approach for production of wine with reduced alcohol content[14, 15, 16].

Genome-scale metabolic models (GEMs) are viewed as a useful tool to study and explain metabolism of an organism[17, 18]. For the model organism *Saccharomyces cerevisiae* there exist well curated models[19, 20] which has been used for various purposes. One of the these applications is explanation of the Crabtree effect using enzyme constrained models genome scale models (ecGEMs)[21, 22]. On the other hand, high-quality GEMs for non-*Saccharomyces* are harder to obtain. However, tools which facilitate automatic generation of models from genome sequences have been developed to ease construction of GEMs for non-model organism. A promising approach is carving[23] where models are created from a curated universal model which serves as a database from which the reactions are selected. Furthermore, for incorporation of enzymatic constraints, there currently exists tools and frameworks[24, 25] for automatically querying databases for protein masses and catalytic rates and integrate these data into an ecGEM.

In this article, we will construct GEMs for five of the most commonly applied alternative non-*Saccharomyces* yeast strains attempted for wine brewing[4]. These are: *Hanseniaspora osmophila*, *Kluveromyces lactis*, *Metschnikowia pulcherrima*, *Torulaspora delbrueckii*, and *Lachancea thermotolerans*. The models are automatically constructed from genome sequences and carved form a curated universal yeast model. Based on the GEMs, we will assess the whether physiological properties of the yeasts can be predicted *in silico*.

Results

Complex I differentiates *Metschnikowia pulcherrima* from the other organisms

Genome-scale models of the five non-*Saccharomyces* yeast strains were created. From these models, enzyme constrained (sMOMENT) models were made with

AutoPACMEN[24] (see Methods for details). Key properties of the models are summarized in Table 1. We first simulated the models with dFBA[26] over 12 hours without enzymatic constraints (Figure 1) and included the *Saccharomyces cerevisiae* model iND750[27] for reference. Here, the glucose was the sole carbon source, initially set to $10\,\mathrm{mmol\,L^{-1}}$ ($1.8\,\mathrm{g\,L^{-1}}$) and the supply of oxygen was restricted to 10 mmol/g DW Biomass/h (corresponds to 180 mg/g WD /h). We observed that the model for *Metschnikowia pulcherrima* was somewhat different than others as it produced low levels of fermentation products (ethanol and acetate) compared to the other models and had a higher final biomass yield for the same amount of glucose. We also tried initiating the simulations with $1000\,\mathrm{mmol\,L^{-1}}$ ($180\,\mathrm{g\,L^{-1}}$) glucose which is a more realistic sugar concentration in grape must (Supplementary Figure S1). This resulted in higher degree of fermentation due to the fact that balance between glucose and oxygen availability was shifted. Still, the same tendencies of *Metschnikowia pulcherrima* to produce less fermentation productions and attain higher biomass, were evident.

We first wanted of assess why the metabolism of *Metschnikowia pulcherrima* was different from that of *Kluyveromyces lactis* and the other yeast strains at infinite level of the enzyme pool. We investigated this issue by comparing the model of *Metschnikowia pulcherrima* with the model of *Kluyveromyces lactis* and used the GEM models without enzyme constraints. We considered using the *Saccharomyces cerevisiae* model for this task, but it proved to be difficult because the reactions and metabolite namespaces of the iND750 is different than for the non-*Saccharomyces* strains.

Of the reaction in the *Metschnikowia pulcherrima* model, 191 were not present in the *Kluyveromyces lactis* model. We therefore knocked out each of these *Metschnikowia pulcherrima* reactions cumulatively in sequence and optimized for biomass production given the initial nutrient concentration of the dFBA simulations. Some of these reactions were essential and therefore reinserted into the model before continuing. Of the considered reactions which were not essential, we observed two reactions which altered the growth rate: Complex I in the respiratory electron transport chain (NADH dehydrogenase) and mitochondrial Methylenetetrahydrofolate dehydrogenase (NAD+). Of these reactions, removal of Complex I alone was sufficient to produce the same growth as in *Kluyveromyces lactis*. Conversely, adding the Complex I reactions to the *Kluyveromyces lactis* model yielded the same growth rate as for *Metschnikowia pulcherrima*. On the other hand, removal of mitochondrial Methylenetetrahydrofolate dehydrogenase (NAD+) from the *Metschnikowia pulcherrima* did not have any effect on its own, nor did addition of the same reaction into the model of *Kluyveromyces lactis*. Therefore, we chose to focus on Complex I when comparing the models.

According to the reconstructions, *Metschnikowia pulcherrima* was annotated with Complex I whereas none of the other yeast strains had this reaction. *Saccharomyces cerevisiae*, *Kluveromyces lactis* and many other yeasts do not have the canonical Complex I of the electron transport chain, but instead feature alternative Type II NADH dehydrogenases which does not pump protons across the mitochondrial membrane[28]. According to the model, Complex I pumps 4 protons across the mitochondrial membrane for each molecule of NADH being reduced, whereas the alternative Type II NADH dehydrogenases do not possess the ability. This means that *Metschnikowia pulcherrima* was able to create a larger proton motive force (PMF) per mole of NADH being oxidized, which in turn increased the efficiency in generation of ATP per mole of glucose.

In order to obtain further evidence that Complex I was indeed present in *Metschnikowia pulcherrima*, we did a BLAST[29] search of the gene products in the *Metschnikowia pulcherrima* model which were annotated to be associated with Complex I functionality (EC number 7.1.1.2). Our query returned various Complex I subunits which already were annotated in the TrEMBL database[30] as *Metschnikowia pulcherrima*. However, all functional annotations of these proteins were based on homology and not curated experimental results. This means that we are confident that the presence of Complex I in *Metschnikowia pulcherrima* is not an annotation error although wet lab experiments are required in order to get a definitive answer.

We therefore suspected that by removing the advantage of protein pumping by Complex I, the metabolism of *Metschnikowia pulcherrima* would become more similar to the other yeast strains. We therefore artificially changed the stoichiometry of the reaction such that 2 or 0 protons were pumped for each molecule of NADH being consumed. We ran dFBA simulations without enzyme constraints with the same starting conditions and used *Kluyveromyces lactis* (lacking Complex I) as a baseline(Figure 2). From the results, we observed that the glucose consummation and biomass production were more or less identical for *Metschnikowia pulcherrima* and *Kluyveromyces lactis* when the proton pumping was turned off. Also, for the partially inhibited state of 2 protons pumped, the biomass yield and glucose consumption was somewhere between the wild-type and inhibited state. For the production of ethanol and acetate it was observed that the production of ethanol and acetate decreased with the number of protons pumped by Complex I, yet *Kluyveromyces lactis* still had a higher production of ethanol and acetate when no protons were pumped.

Protein constrains result in changes in metabolic pathways

Considering that Complex I was a key differentiator for *Metschnikowia pulcherrima* with infinite amounts of enzymatic protein available, we next studied how the activity of Complex I affected metabolism when the available enzyme pool was constrained.

When comparing the models of *Metschnikowia pulcherrima* and *Kluyveromyces lactis* under enzyme constraints, we found *Metschnikowia pulcherrima* to have a marginally higher growth rate than *Kluyveromyces lactis*, but yet a much higher consumption of glucose. Upon investigating this difference, we realized that this was due to a wasteful mitochondrial membrane proton leakage in the *Metschnikowia pulcherrima* model. However, when running Flux Variability Analysis (FVA)[31], we found that far more efficient pathways were possible when requiring 99% of maximal growth rate. Thus, we deemed the proton leakage to be physiologically implausible even though some unicellular eukaryotes have been shown to have proton uncoupling proteins[32]. We therefore corrected the model by knocking out the proton leakage reaction before running any further sMOMENT simulations.

When simulating the effect of Complex I stoichiometry of *Metschnikowia pulcherrima* with dynamic enzyme constrained FBA (decFBA)[21] (Figure 3), we could not observe any major effect of the stoichiometry of Complex I for low availability of enzymatic protein. However at the highest chosen protein pool, the biomass yield was higher the more protons were pumped by Complex I. This is as expected since decFBA becomes equivalent to dFBA when the available enzyme pool approaches infinity. For the second highest protein pool, the effects of stoichiometry were marginal. Most likely this means that the model chooses other pathways when availability of enzymatic protein is scarce and hence is not reliant of Complex I to the same degree. Moreover, the production of acetate and ethanol was affected by the number of protons pumped at the highest level of enzyme only.

Finally, we compared the the sMOMENT models of *Metschnikowia pulcherrima* and *Kluyveromyces lactis* as to get an overview of how the species compare when the access to enzymatic protein is restricted (Figure 4). As expected, the growth rate increased with the protein availability, but only up to a certain point where the substrate uptake rates became limiting. Also, the results show that while the growth rate and biomass yield for *Metschnikowia pulcherrima* was higher than for *Kluyveromyces lactis* at high availability of enzymatic protein. For lower levels of the enzyme pool, the two strains grew almost as quickly until glucose was exhausted. However, the biomass yield was higher for *Metschnikowia pulcherrima*, so the final biomass was higher than that for *Kluyveromyces lactis* even at the low levels of protein pool. In addition, the *Metschnikowia pulcherrima* produced less acetate than *Kluyveromyces lactis* for all levels of the enzyme pool. Respiration is energetically more efficient than fermentation in utilization of the carbon source than fermentation[22], but comes with a higher protein cost per unit of ATP consumed. For this reason, we would expect to observe less production of fermentation products at high levels of enzymatic proteins which agree with our observations.

Discussion

In our resulting models, we found the GEM of *Metschnikowia pulcherrima* to be different from the other models as it utilized glucose more efficiently, had a higher biomass yield and produced less fermentation products. To a great extent, this echoes recent research which suggest *Metschnikowia pulcherrima* as a good candidate for reducing alcohol content in wine[16, 15, 33]. The connection between the three observed effects are quite straight forward. Respiration instead of fermentation gives better energy utilization of the substrate, less fermentation products and better growth for the same amount of substrate consumed.

However, for some yeasts, having an extensive respiratory metabolism might not be evolutionary beneficial for two reasons. First, having the availability to respire is not an advantage when supply of oxygen is insufficient, which may be the case in commercial wine fermentation tanks. Second, when high concentrations of glucose are available, yeast may achieve higher production flux of ATP production through ethanol fermentation than by respiration because respiration requires more protein usage than fermentation[21, 22]. Indeed, we observed larger production of fermentation products at low protein availability. Our results show that the presumed presence of Complex I in *Metschnikowia pulcherrima* allows the organism to respire glucose more efficiently than the other yeast strains. Thus, *Metschnikowia pulcherrima* may be better adopted for respiration and will therefore prefer this mode of metabolism. However, according to the decFBA simulations, the advantage of Complex I at lower levels of the protein pool vanishes, but *Metschnikowia pulcherrima* still utilizes glucose better than *Kluyveromyces lactis* in these cases too. What causes this effect, is yet to be investigated.

Still, the results from this study should be taken with a bit of caution. First of all, errors could arise in the genome annotation and carving of the models. This could produce reactions present in the model, but not the organism, and vice versa. Furthermore, introducing enzyme constraints by AutoPACMEN was likely also a major source of error. Database k_{cat} values have been shown to be variable and often far from realistic *in vivo* values[34, 35, 36]. Additionally, the k_{cat} coverage for non-model organisms is low, such that *Saccharomyces cerevisiae* was likely the closest available candidate for picking k_{cat} values for many of the enzymes. In order to correctly calculate enzyme constraints, all reactions have to be annotated with their respective enzymes, and complexes with corresponding subunit stoichiometry have to be accounted for. In our models, such annotations were missing for some important reactions. Notably, the mitochondrial genome was not sequenced for any of the non-*Saccharomyces* yeast strains and as such, enzyme constraints were not imposed on proteins encoding by the mitochondrial genome. This is a major shortcoming given that many of the expensive proteins involved in respiration are electron transport proteins which are encoded by the mitochondrial genome.

Our choice of parameters for dFBA simulations were based on educated guesses in lack of good data for calibration. Sànchez *et al.* used a enzymatic protein pool of $P_{tot} = 0.448\,\text{g/gDW}$ and saturation factor of $\sigma = 0.5$ for their ecYeast7 model of *Saccharomyces cerevisiae*. In our case, this would correspond to a simulated protein pool of approximately $0.22\,\text{g/gDW}$ since we assumed full saturation. Still, some enzymatic reactions were not accounted for in the sMOMENT models, so we think a somewhat lower enzyme pool would make a fairer comparison. Glucose uptake has been shown to vary considerably between different species of yeast and even between different strains of *Saccharomyces cerevisiae*[37, 38]. From the available data and literature[39], we consider our chosen parameters to be within a realistic range.

Nevertheless, we acknowledge that glucose uptake and its balance to oxygen uptake is crucial to the nature of the fermentation. Less oxygen available compared to the consumption of glucose will favour fermentation at the expense of respiration. Additionally, regulatory mechanisms not accounted for by our models most likely also regulate the switching between fermentation and respiration[40]. Comparing Figure 1 and Supplementary Figure S1, we observed that the glucose concentration had a large effect on the production of fermented compounds, yet *Metschnikowia pulcherrima* still had a stronger respiratory metabolism than the other yeasts for high glucose concentrations. We did not account for the fact that supplying oxygen is harder when the biomass concentration is high, making a fixed oxygen uptake of $10\,\text{mmol/gDW}$ realistic in Figure 1, but unrealistic in Supplementary Figure S1.

Nevertheless, the model predictions for *Metschnikowia pulcherrima* should inspire to further research and investigations into the nature of its respiratory metabolism.

One of the central questions is whether our claim this yeast has Complex I is correct, and if so, which phenotypic effects this enzyme has. Rotenone is known to be an inhibitor of Complex I and would therefore be a useful tool to study the activity of Complex I[41, 42]. Furthermore, systematic studies must be conducted in order to assess how *Metschnikowia pulcherrima* behaves under varying availability of glucose and oxygen.

Methods

Creation of the yeast models

The protein sequence of the five species was obtained from the NCBI database[43, 44, 45, 46, 47, 48]. We annotated the function of the proteins with EggNog mapper V2[49] using Diamond[50] for the search of homologs in the EggNOG ortholog database version 5.

For the automatic model reconstruction, we used the software package CarveFungi[51]. CarveFungi, based on the CarveMe algorithm[23], creates a score for each reaction in a universal metabolic model linking their EC numbers to the annotation of the proteins obtained with EggNOG. The software contains a deep learning model to predict the subcellular localization of fungal proteins. This prediction contributes to the reaction score, assigning the reactions to a specific compartment in the model. The reaction scores are then used by a Mixed-integer linear programming problem (MILP) to maximize the reactions present in the universal model with a high score and to minimize the reactions with a low score while maintaining the network connectivity and the model functionality.

The universal metabolic model used for the model reconstruction was created by combining fungal reactions from public databases such as KEGG[52] and Metacyc[53] and was manually curated using literature to make it atom-balanced and simulatable adding exchanges and a biomass reaction extended from the one present in the yeast consensus model[19].

The automatically reconstructed metabolic models were produced as ensembles of up to 25 models corresponding to alternative reconstructions from the same genome. For our analysis, we turned each ensemble into a single consensus model where a reaction was included if it was present in half or more of the models in the ensemble.

Incorporation of enzymatic constraints

sMOMENT models with enzyme constraints were generated by feeding the GEMs into AutoPACMEN[24] version 0.6.1, applying default parameters. The BiGG metabolite file used by AutoPACMEN was retrieved from the BiGG[54] website (`http://bigg.ucsd.edu/data_access`, October 2020), while the BRENDA data was downloaded from the BRENDA[55] website (`https://www.brenda-enzymes.org/download.php`, October 2020). Before providing the models to AutoPACMEN, the models are augmented by Uniprot identifiers using Uniprot's API. AutoPACMEN retrieved k_{cat} values from SABIO-RK[56, 57] and protein masses from Uniprot[30] using its built-in API interface (October 2020). AutoPACMEN's model calibrator was not used.

dFBA and decFBA simulations

The models of the five non-*Saccharomyces* strains and the iND750 *Saccharomyces cerevisiae* model[27] were simulated *in silico* with dynamic FBA (dFBA)[26, 21]. The COBRApy package (version 0.25.0)[58] was used to handle the models and the resulting LP problems were solved by the Gurobi optimizer (version 9.1.2). Glucose was the sole carbon source available with a maximum uptake flux determined by Michaelis-Menten kinetics: $v_{glc} \leq \frac{V_{max,glc}[glc]}{K_{M,glc}+[glc]}$, where the maximal uptake rate $V_{max,glc} = 10\,\mathrm{mmol/gDW}$, the half-saturation constant $K_{M,glc} = 5\,\mathrm{mmol}$ and $[glc]$ was the glucose concentration in the medium which was initiated to $[glc]_0 = 10\,\mathrm{mmol\,L^{-1}}$ for all simulations expect for Supplementary Figure S1 where $[glc]_0 = 1000\,\mathrm{mmol\,L^{-1}}$. The biomass concentration was as initiated to $[\mathrm{X}]_0 = 0.1\,\mathrm{gDW/L}$. Oxygen was available at a fixed rate of $v_{\mathrm{oxygen}} \leq 10\,\mathrm{mmol/gDW}$.

Entities kept track of were the biomass, glucose, acetate, ethanol and glycerol. The latter three components were included to keep track of the accumulation of fermentation products from the yeast. However, none of the models produced any glycerol for the conditions tested, so we ignored glycerol for clarity. Also, we wanted to emphasize the combined production of ethanol and acetate. Therefore, the plots featured a panel showing the sum of ethanol and acetate in the medium, while another panel showed the acetate concentration. In order to block physiologically implausible metabolic exports, the exports reactions for lactate (both stereoisomers), dihydroxyacetate, D-ribulose and arabinitol were blocked. In order to obtain as consistent physiologically plausible results as possible, lexicographic objectives were applied when running FBA on the models in the following order:

1. Maximize production of biomass
2. Minimize consumption of glucose
3. Maximize excretion of ethanol
4. Maximize excretion of acetate
5. Maximize excretion of glycerol

The models were simulated using the static optimization approach and SciPy's `solve_ivp` function[59]. For the ODE solver, the BDF algorithm[60] was used with an absolute and relative tolerance of 10^{-2}. In cases where the optimization problem because infeasible, the simulation was terminated, but results were padded such that the final state of the system was imputed to all time-points beyond the termination. This happened only if the model has unable to grow because the carbon source (glucose) in the medium was depleted.

dFBA was run both for the original models generated with CarveFungi and the sMOMENT models processed through AutoPACMEN. Upon running decFBA with the sMOMENT models, the level of the enzyme pool was adjusted. For the simulations with sMOMENT models, three different levels of the enzymatic protein pool (0.1, 0.25, and 1.0 grams of protein per gram dry weight(g/gWD)) were chosen.

Abbreviations

FBA: Flux Balance Analysis
dFBA: dynamic Flux Balance Analysis
decFBA: dynamic enzyme-constrained FBA
MILP Mixed-Integer Linear Programming **GEM**: Genome-Scale Model

ecGEM: enzyme constrained Genome-Scale Model
LP: Linear Programming
DW: Dry Weight
PMF: Proton Motive Force

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The source code generating, augmenting and analysing the GEMs used in this paper in addition to the resulting GEMs are available at request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is funded by ERA CoBioTech project CoolWine and the Norwegian Research Council grant 283862.

Authors' contributions

S.C. and P.J. created the GEMs from the genome sequences using CarveFungi and conducted the bioinformatics analyses. J.P.P. augmented the GEMs with enzyme constraints using AutoPACMEN, studies the properties of the GEMs and wrote the first draft of the paper. E.A. supervised the project. All authors contributed to and accepted the final version of the paper.

Acknowledgements

Thanks to Pavlos Stephanos Bekiaris for providing technical support and assistance with the AutoPACMEN software.

Author details

[1]Department of Biotechnology and Food Science, NTNU- Norwegian University of Science and Technology, Trondheim, Norway. [2]K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and General Practice, NTNU- Norwegian University of Science and Technology, Trondheim, Trondheim, Norway. [3]VTT Technical Research Centre of Finland, Espoo, Finland. [4]Department of Bioproducts and Biosystems Aalto University, Espoo, Finland.

References

1. Lin, X., Tang, X., Han, X., He, X., Han, N., Ding, Y., Sun, Y.: Effect of metschnikowia pulcherrima on saccharomyces cerevisiae pdh by-pass in mixedfermentation with varied sugar concentrations of synthetic grape juice and inoculation ratios. Fermentation **8**(10) (2022). doi:10.3390/fermentation8100480
2. Contreras, A., Curtin, C., Varela, C.: Yeast population dynamics reveal a potential 'collaboration' between metschnikowia pulcherrima and saccharomyces uvarum for the production of reduced alcohol wines during shiraz fermentation. Applied Microbiology and Biotechnology **99**(4), 1885–1895 (2015). doi:10.1007/s00253-014-6193-6
3. García, M., Esteve-Zarzoso, B., Cabellos, J.M., Arroyo, T.: Sequential non-saccharomyces and saccharomyces cerevisiae fermentations to reduce the alcohol content in wine. Fermentation **6**(2) (2020). doi:10.3390/fermentation6020060
4. Jolly, N.P., Varela, C., Pretorius, I.S.: Not your ordinary yeast: non-Saccharomyces yeasts in wine production uncovered. FEMS Yeast Research **14**(2), 215–237 (2014). doi:10.1111/1567-1364.12111. https://academic.oup.com/femsyr/article-pdf/14/2/215/18113396/14-2-215.pdf
5. Ciani, M., Morales, P., Comitini, F., Tronchoni, J., Canonico, L., Curiel, J.A., Oro, L., Rodrigues, A.J., Gonzalez, R.: Non-conventional yeast species for lowering ethanol content of wines. Frontiers in Microbiology **7** (2016). doi:10.3389/fmicb.2016.00642
6. Varela, C., Dry, P.R., Kutyna, D.R., Francis, I.L., Henschke, P.A., Curtin, C.D., Chambers, P.J.: Strategies for reducing alcohol concentration in wine. Australian Journal of Grape and Wine Research **21**(S1), 670–679 (2015). doi:10.1111/ajgw.12187. https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajgw.12187
7. Alston, J., Fuller, K.B., Lapsley, J.T., Soleas, G.: Too much of a good thing? causes and consequences of increases in sugar content of california wine grapes*. Journal of Wine Economics **6**(2), 135–159 (2011)
8. Longo, R., Blackman, J.W., Torley, P.J., Rogiers, S.Y., Schmidtke, L.M.: Changes in volatile composition and sensory attributes of wines during alcohol content reduction. Journal of the Science of Food and Agriculture **97**(1), 8–16 (2017). doi:10.1002/jsfa.7757. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jsfa.7757
9. Bärwald, G., Fischer, A.: Crabtree effect in aerobic fermentations using grape juice for the production of alcohol reduced wine. Biotechnology Letters **18**(10), 1187–1192 (1996). doi:10.1007/BF00128590

10. Hammad, N., Rosas-Lemus, M., Uribe-Carvajal, S., Rigoulet, M., Devin, A.: The crabtree and warburg effects: Do metabolite-induced regulations participate in their induction? Biochimica et biophysica acta **1857**, 1139–1146 (2016). doi:10.1016/j.bbabio.2016.03.034
11. Curiel, J.A., Salvadó, Z., Tronchoni, J., Morales, P., Rodrigues, A.J., Quirós, M., Gonzalez, R.: Identification of target genes to control acetate yield during aerobic fermentation with saccharomyces cerevisiae. Microbial Cell Factories **15**(1), 156 (2016). doi:10.1186/s12934-016-0555-y
12. van Dijken, J.P., Weusthuis, R.A., Pronk, J.T.: Kinetics of growth and sugar consumption in yeasts. Antonie van Leeuwenhoek **63**, 343–352 (1993). doi:10.1007/BF00871229
13. Vicente, J., Ruiz, J., Belda, I., Benito-Vázquez, I., Marquina, D., Calderón, F., Santos, A., Benito, S.: The genus metschnikowia in enology. Microorganisms **8**(7) (2020). doi:10.3390/microorganisms8071038
14. Canonico, L., Comitini, F., Oro, L., Ciani, M.: Sequential fermentation with selected immobilized non-saccharomyces yeast for reduction of ethanol content in wine. Frontiers in Microbiology **7** (2016). doi:10.3389/fmicb.2016.00278
15. Hranilovic, A., Gambetta, J.M., Jeffery, D.W., Grbin, P.R., Jiranek, V.: Lower-alcohol wines produced by metschnikowia pulcherrima and saccharomyces cerevisiae co-fermentations: The effect of sequential inoculation timing. International Journal of Food Microbiology **329**, 108651 (2020). doi:10.1016/j.ijfoodmicro.2020.108651
16. Contreras, A., Hidalgo, C., Henschke, P.A., Chambers, P.J., Curtin, C., Varela, C.: Evaluation of non-saccharomyces yeasts for the reduction of alcohol content in wine. Applied and environmental microbiology **80**, 1670–8 (2014)
17. Passi, A., Tibocha-Bonilla, J.D., Kumar, M., Tec-Campos, D., Zengler, K., Zuniga, C.: Genome-scale metabolic modeling enables in-depth understanding of big data. Metabolites **12** (2021)
18. Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., Lee, S.Y.: Current status and applications of genome-scale metabolic models. Genome Biology **20**(1), 121 (2019). doi:10.1186/s13059-019-1730-3
19. Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P.M., Lappa, D., Lieven, C., Beber, M.E., Sonnenschein, N., Kerkhoven, E.J., Nielsen, J.: A consensus s. cerevisiae metabolic model yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nature Communications **10**(1), 3586 (2019). doi:10.1038/s41467-019-11581-3
20. Lu, H., Li, F., Yuan, L., Domenzain, I., Yu, R., Wang, H., Li, G., Chen, Y., Ji, B., Kerkhoven, E.J., Nielsen, J.: Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection. Molecular Systems Biology **17**(10), 10427 (2021). doi:10.15252/msb.202110427. https://www.embopress.org/doi/pdf/10.15252/msb.202110427
21. Moreno-Paz, S., Schmitz, J., Martins Dos Santos, V.A.P., Suarez-Diez, M.: Enzyme-constrained models predict the dynamics of saccharomyces cerevisiae growth in continuous, batch and fed-batch bioreactors. Microbial biotechnology (2022). doi:10.1111/1751-7915.13995
22. Nilsson, A., Nielsen, J.: Metabolic trade-offs in yeast are caused by f1f0-atp synthase. Scientific Reports **6**(1), 22264 (2016). doi:10.1038/srep22264
23. Machado, D., Andrejev, S., Tramontano, M., Patil, K.R.: Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research **46**(15), 7542–7553 (2018). doi:10.1093/nar/gky537. https://academic.oup.com/nar/article-pdf/46/15/7542/25689981/gky537.pdf
24. Bekiaris, P.S., Klamt, S.: Automatic construction of metabolic models with enzyme constraints. BMC Bioinformatics **21**(1), 19 (2020). doi:10.1186/s12859-019-3329-9
25. Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E.J., Nielsen, J.: Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol **13**(8), 935 (2017). doi:10.15252/msb.20167411
26. Mahadevan, R., Edwards, J.S., Doyle, F.J. 3rd: Dynamic flux balance analysis of diauxic growth in escherichia coli. Biophys J **83**(3), 1331–40 (2002). doi:10.1016/S0006-3495(02)73903-9
27. Duarte, N.C., Herrgård, M.J., Palsson, B.o.: Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. Genome research **14**, 1298–309 (2004)
28. Antos-Krzeminska, N., Jarmuszkiewicz, W.: Alternative type ii nad(p)h dehydrogenases in the mitochondria of protists and fungi. Protist **170**(1), 21–37 (2019). doi:10.1016/j.protis.2018.11.001
29. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**, 403–10 (1990)
30. Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Castro, L.G., Garmiri, P., Georghiou, G., Gonzales, L., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Jokinen, P., Joshi, V., Jyothi, D., Lock, A., Lopez, R., Luciani, A., Luo, J., Lussi, Y., Mac-Dougall, A., Madeira, F., Mahmoudy, M., Menchi, M., Mishra, A., Moulang, K., Nightingale, A., Oliveira, C.S., Pundir, S., Qi, G., Raj, S., Rice, D., Lopez, M.R., Saidi, R., Sampson, J., Sawford, T., Speretta, E., Turner, E., Tyagi, N., Vasudev, P., Volynkin, V., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A., Poux, S., Redaschi, N., Aimo, L., Argoud-Puy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M.-C., Bolleman, J., Boutet, E., Breuza, L., Casals-Casas, C., de Castro, E., Echioukh, K.C., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Keller, G., Kerhornou, A., Lara, V., Le Mercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T.B., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Pozzato, M., Pruess, M., Rivoire, C., Sigrist, C., Sonesson, K., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.-S., Zhang, J., Consortium, U.: Uniprot: the universal protein knowledgebase in 2021. NUCLEIC ACIDS RESEARCH **49**(D1), 480–489 (2021). doi:10.1093/nar/gkaa1100
31. Mahadevan, R., Schilling, C.H.: The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering **5**(4), 264–276 (2003). doi:10.1016/j.ymben.2003.09.002
32. Jarmuszkiewicz, W., Woyda-Ploszczyca, A., Antos-Krzeminska, N., Sluse, F.E.: Mitochondrial uncoupling

proteins in unicellular eukaryotes. Biochimica et Biophysica Acta (BBA) - Bioenergetics **1797**(6), 792–799 (2010). doi:10.1016/j.bbabio.2009.12.005. 16th European Bioenergetics Conference 2010

33. Canonico, L., Comitini, F., Ciani, M.: Metschnikowia pulcherrima selected strain for ethanol reduction in wine: Influence of cell immobilization and aeration condition. Foods (Basel, Switzerland) **8** (2019)
34. Heckmann, D., Campeau, A., Lloyd, C.J., Phaneuf, P.V., Hefner, Y., Carrillo-Terrazas, M., Feist, A.M., Gonzalez, D.J., Palsson, B.O.: Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proceedings of the National Academy of Sciences of the United States of America **117**, 23182–23190 (2020). doi:10.1073/pnas.2001562117
35. Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., Palsson, B.O.: Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nature Communications **9**(1), 5252 (2018). doi:10.1038/s41467-018-07652-6
36. Wendering, P., Arend, M., Razaghi-Moghadamkashani, Z., Nikoloski, Z.: Data integration across conditions improves turnover number estimates and metabolic predictions. bioRxiv (2022). doi:10.1101/2022.04.01.486742. https://www.biorxiv.org/content/early/2022/05/25/2022.04.01.486742.full.pdf
37. Does, A.L., Bisson, L.F.: Comparison of glucose uptake kinetics in different yeasts. Journal of bacteriology **171**, 1303–8 (1989)
38. Nissen, P., Nielsen, D., Arneborg, N.: The relative glucose uptake abilities of non-saccharomyces yeasts play a role in their coexistence with saccharomyces cerevisiae in mixed cultures. Applied Microbiology and Biotechnology **64**(4), 543–550 (2004). doi:10.1007/s00253-003-1487-0
39. Pizarro, F., Varela, C., Martabit, C., Bruno, C., Pérez-Correa, J.R., Agosin, E.: Coupling kinetic expressions and metabolic networks for predicting wine fermentations. Biotechnology and Bioengineering **98**(5), 986–998 (2007). doi:10.1002/bit.21494. https://onlinelibrary.wiley.com/doi/pdf/10.1002/bit.21494
40. Otterstedt, K., Larsson, C., Bill, R.M., Ståhlberg, A., Boles, E., Hohmann, S., Gustafsson, L.: Switching the mode of metabolism in the yeast saccharomyces cerevisiae. EMBO reports **5**(5), 532–537 (2004). doi:10.1038/sj.embor.7400132. https://www.embopress.org/doi/pdf/10.1038/sj.embor.7400132
41. Heinz, S., Freyberger, A., Lawrenz, B., Schladt, L., Schmuck, G., Ellinger-Ziegelbauer, H.: Mechanistic investigations of the mitochondrial complex i inhibitor rotenone in the context of pharmacological and safety evaluation. Scientific Reports **7**(1), 45465 (2017). doi:10.1038/srep45465
42. Ozay, E.I., Sherman, H.L., Mello, V., Trombley, G., Lerman, A., Tew, G.N., Yadava, N., Minter, L.M.: Rotenone treatment reveals a role for electron transport complex i in the subcellular localization of key transcriptional regulators during t helper cell differentiation. Frontiers in immunology **9**, 1284 (2018)
43. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D.: Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation **44**, 733–745. doi:10.1093/nar/gkv1189. Accessed 2022-12-11
44. Kluyveromyces Lactis, Assembly ASM251v1. National Center for Biotechnology Information (NCBI). `https://www.ncbi.nlm.nih.gov/genome/?term=ASM251v1`
45. Metschnikowia Pulcherrima, Assembly ASM421770v1. National Center for Biotechnology Information (NCBI). `https://www.ncbi.nlm.nih.gov/assembly/GCA_004217705.1/`
46. Lachancea Thermotolerans, Assembly ASM14280v1. National Center for Biotechnology Information (NCBI). `https://www.ncbi.nlm.nih.gov/genome/?term=ASM14280v1`
47. Torulaspora Delbrueckii, Assembly ASM24337v1. National Center for Biotechnology Information (NCBI). `https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000243375.1/`
48. Hanseniaspora Osmophila, Assembly ASM174704v1. National Center for Biotechnology Information (NCBI). `https://www.ncbi.nlm.nih.gov/genome/46405?genome_assembly_id=283698`
49. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., von Mering, C., Bork, P.: eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses **47**, 309–314. doi:10.1093/nar/gky1085
50. Buchfink, B., Reuter, K., Drost, H.-G.: Sensitive protein alignments at tree-of-life scale using diamond. Nature Methods **18**(4), 366–368 (2021). doi:10.1038/s41592-021-01101-x
51. Castillo, S.: CarveFungi (2021). `https://github.com/SandraCastilloPriego/CarveFungi`
52. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes **28**(1), 27–30. doi:10.1093/nar/28.1.27
53. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases **44**, 471–480. doi:10.1093/nar/gkv1164
54. King, Z.A., Lu, J., Draeger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., Lewis, N.E.: Bigg models: A platform for integrating, standardizing and sharing genome-scale models. NUCLEIC ACIDS RESEARCH **44**(D1), 515–522 (2016). doi:10.1093/nar/gkv1049
55. Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., Schomburg, D.: Brenda, the elixir core data resource in 2021: new developments and updates. NUCLEIC ACIDS RESEARCH **49**(D1), 498–508 (2021). doi:10.1093/nar/gkaa1025
56. Wittig, U., Kania, R., Golebiewski, M., Rey, M., Shi, L., Jong, L., Algaa, E., Weidemann, A., Sauer-Danzwith, H., Mir, S., Krebs, O., Bittkowski, M., Wetsch, E., Rojas, I., Mueller, W.: Sabio-rk-database for biochemical

reaction kinetics. NUCLEIC ACIDS RESEARCH **40**(D1), 790–796 (2012). doi:10.1093/nar/gkr1046

57. Wittig, U., Rey, M., Weidemann, A., Kania, R., Mueller, W.: Sabio-rk: an updated resource for manually curated biochemical reaction kinetics. NUCLEIC ACIDS RESEARCH **46**(D1), 656–660 (2018). doi:10.1093/nar/gkx1065
58. Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R.: Cobrapy: Constraints-based reconstruction and analysis for python. BMC Systems Biology **7**(1), 74 (2013). doi:10.1186/1752-0509-7-74
59. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). doi:10.1038/s41592-019-0686-2
60. Byrne, G.D., Hindmarsh, A.C.: A polyalgorithm for the numerical solution of ordinary differential equations. ACM Trans. Math. Softw. **1**(1), 71–96 (1975). doi:10.1145/355626.355636

Figures

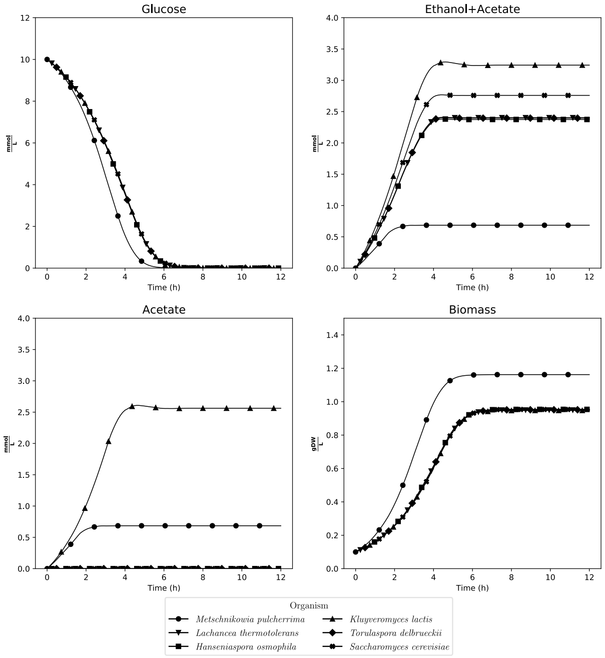


Figure 1 dFBA simulations of the models without enzymatic constraints for the six yeast strains starting with $10\,\mathrm{mmol\,L^{-1}}$ glucose. $\frac{\mathrm{gDW}}{\mathrm{L}}$: Grams of dry weight per liter.

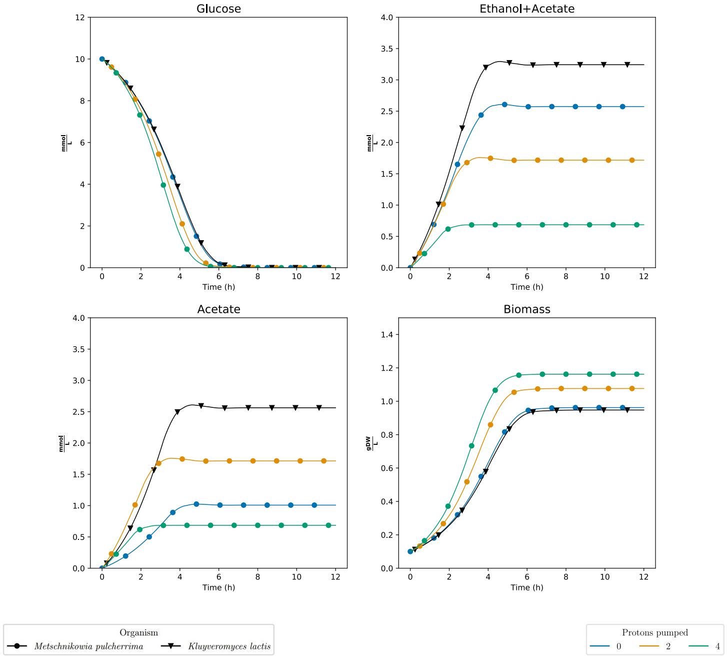


Figure 2 dFBA simulations of the models without enzymatic constraints for *Metschnikowia pulcherrima* and *Kluyveromyces lactis* when artificially changing the stoichiometry of the number of protons pumped by Complex I.

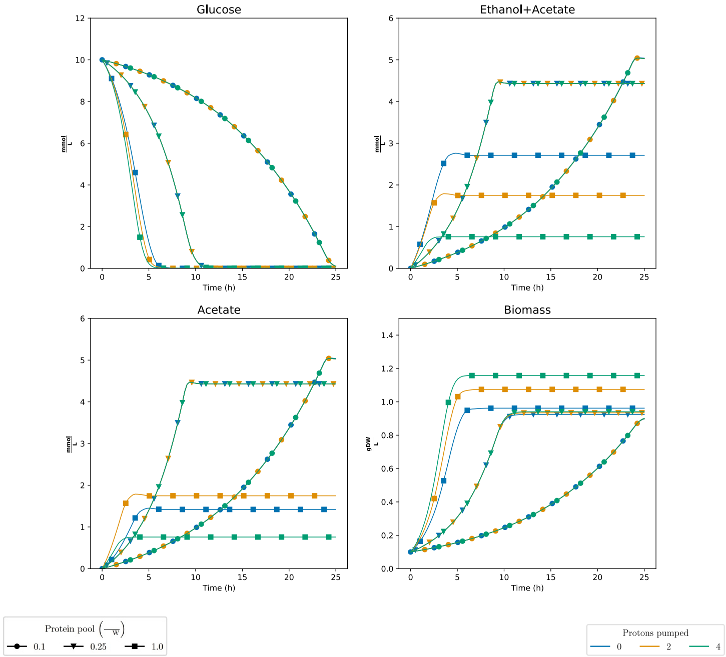


Figure 3 decFBA simulations of the sMOMENT model of *Metschnikowia pulcherrima* when artificially changing the stoichiometry of the number of protons pumped by Complex I under different levels of the protein pool. $\frac{\mathrm{gDW}}{\mathrm{L}}$: Grams of dry weight per liter.

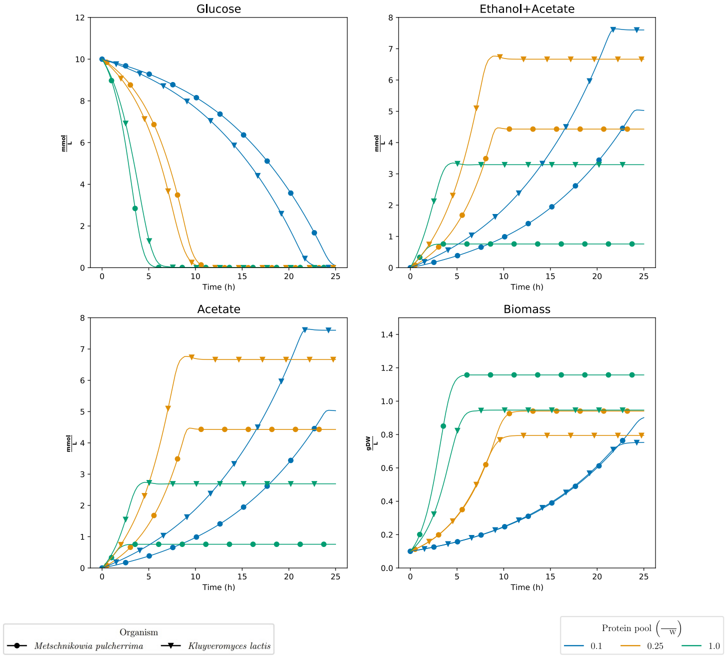


Figure 4 decFBA simulations of the sMOMENT models for *Metschnikowia pulcherrima* and *Kluyveromyces lactis* under different levels of the protein pool. $\frac{\mathrm{gDW}}{\mathrm{L}}$: Grams of dry weight per liter.

541 **Tables**

Table 1 Properties of the models studied in this paper. The model for *Saccharomyces cerevisiae* was taken from an external source[27] and did not have any corresponding enzyme constrained model.

Organism	Reactions	Reversible reactions	Metabolites	Reactions drawing from protein pool
Metschnikowia pulcherrima	2049	610	1633	1310
Lachancea thermotolerans	2049	618	1647	1319
Torulaspora delbrueckii	1876	559	1510	1163
Kluyveromyces lactis	2131	621	1774	1401
Hanseniaspora osmophila	1556	520	1218	902
Saccharomyces cerevisiae	1266	436	1059	NA

Additional Files

543 Supplementary Materials

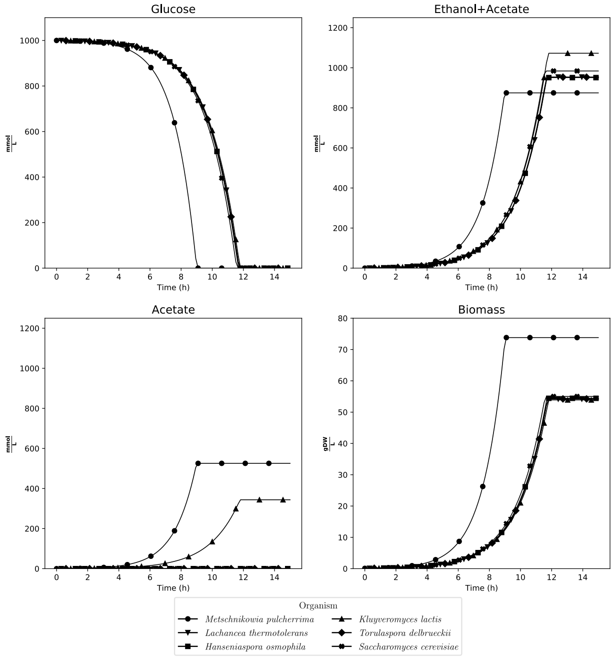


Figure S1 dFBA simulations of the models without enzymatic constraints for the six yeast strains starting with $1000\,\mathrm{mmol\,L^{-1}}$ glucose. $\frac{\mathrm{gDW}}{\mathrm{L}}$: Grams of dry weight per liter.

ISBN 978-82-326-7112-0 (printed ver.)
ISBN 978-82-326-7111-3 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology