Modeling Metabolic Networks and their Environment Interaction

Marko Budinich Abarca

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Marko BUDINICH ABARCA

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique et applications, section CNU 27
Spécialité : Bioinformatique
Unité de recherche : Laboratoire d'informatique de Nantes-Atlantique (LINA)

Soutenue le 28 avril 2017

Modeling Metabolic Networks and their Environment Interaction

JURY

Président :	**M. Fabien JOURDAN**, Directeur de Recherche, INRA Toulouse
Rapporteur :	**M. Olivier BERNARD**, Directeur de Recherche, INRIA de Sophia-Antipolis
Examinateurs :	**M^me^ Monique ZAGOREC**, Directeur de Recherche, ONIRIS NANTES
	M^me^ Laurence GARCZAREK, Directeur de Recherche, CNRS Station Biologique Roscoff
	M. Bruno SAINT-JEAN, Chargé de Recherche, IFREMER Nantes
Directeur de thèse :	**M. Jérémie BOURDON**, Professeur, Université de Nantes
Co-directeur de thèse :	**M. Damien EVEILLARD**, Maître de Conférences, Université de Nantes

Résumé

Ces travaux de thèse portent sur le développement et l'application des méthodes pour la modélisation des réseaux métaboliques. De manière plus précise, cette thèse se focalise sur les approches de modélisation par contraintes (Constraint Based Methods en anglais ou CBMs). Dans le cadre de cette modélisation, les réseaux métaboliques sont représentés par les interactions entre les metabolites (*i.e.*, variables) et l'environnement (*i.e.*, paramètres). Pour modéliser ces systèmes vivants, les CBMs possèdent de nombreux avantages par rapport aux autres méthodes de modélisation. Premièrement, la méthode de modélisation repose sur des techniques d'optimisation, par définition sous contraintes, qui sont actuellement en plein essor et pour lesquels de nombreux algorithmes existent. En pratique, ces algorithmes peuvent traiter avec efficacité des problèmes pour des réseaux métaboliques de taille réaliste, ce qui est d'une grande importance pour accompagner l'inexorable augmentation des données d'origine génomique et métabolomique.

Ce manuscrit de thèse est divisé en deux parties. La première partie est consacrée à la modélisation des interactions entre le réseau métabolique d'un micro-organisme et son environment. Dans un premier temps, nous détaillerons les concepts mathématiques sous-jacents aux CBMs.

Le modèle mathématique pour les CBMs est obtenu directement à partir des réactions biochimiques identifiées à partir des annotations fonctionnelles des gènes codant pour des enzymes. Ces réactions décrivent les transformations d'espèces chimiques ; ou substrats ; en d'autres ; ou produits. Dans ce contexte, les espèces chimiques sont appelées aussi métabolites. Au delà de la simple description des réactions, il est également possible de déterminer les contraintes stœchiométriques requises pour l'ensemble des substrats d'une réaction ; les coefficients stœchiométriques.

Dans le cadre des CBMs, l'ensemble de toutes ces contraintes sont stockées dans une matrice stœchiométrique S, dans laquelle chaque réaction correspond à une colonne et chaque métabolite correspond à une ligne. Les coefficients stockés dans cette matrice seront ceux associées aux métabolites impliqués dans les différentes réactions. Par convention, le coefficient sera négatif si le metabolite correspondant est un substrat ; il sera positif si c'est un produit et nul si il ne participe pas à la réaction.

Cette modélisation du réseau métabolique permet de mettre en oeuvre deux analyses fondamentales : les Analyses des Flux Equilibrés (i.e., Flux Balance Analysis - FBA) et l'analyse de variabilité des flux (i.e., Flux Variability Analysis - FVA).

En utilisant cette matrice S, il est en effet possible de considérer l'équilibre d'action de masse pour le système, tel que $\frac{dK}{dt} = Sv$, où K est le vecteur des concentrations des métabolites du système, et v le vecteur des vitesses des réactions ou appelé également flux. Sous l'hypothèse d'état stationnaire du système dynamique, l'équilibre d'action de masse peut être simplifié à $Sv = 0$. Ainsi l'espace des solutions de ce système peut être étudié par des techniques développées de description de l'espace nul de la matrice S, comme les modes élémentaires (Elementary Modes). Néanmoins, cette description reste combinatoire et difficile à résumer du fait du nombre de solutions qui augmente drastiquement avec la taille des réseaux métaboliques.

Ainsi, une autre approche consiste à explorer l'espace des solutions en utilisant des techniques qui considèrent un autre critère d'optimisation. Il faut pour cela utiliser une "fonction objective", qui biologiquement représente le taux de synthèse de biomasse ou taux de croissance microbienne. Cette fonction est une combinaison linéaire des différents flux dans le réseau. La maximisation d'une telle fonction constitue l'hypothèse centrale de la FBA.

Mathématiquement, la valeur maximale de cette fonction est unique, mais malheureusement, la combi-

naison des flux permettant l'obtention de cette valeur ne l'est pas. Pour analyser l'ensemble des solutions qui satisfont une croissance maximale ou la production optimale de biomasse, il faut appliquer la technique de FVA qui identifie les valeurs maximales et minimales de chaque flux autour de la valeur optimale de la fonction objective.

De part la description des précédentes techniques, il est clair que l'optimisation mathématique est un élément central des formulations des CBMs. Par exemple, il a été récemment proposé une CBMs appelée "Stoichiometric Capacitance" (Larhlimi et al., 2012a) qui permet d'évaluer la capacité théorique maximale d'un certain réseau métabolique. La formulation mathématique de ce problème est une optimisation pour laquelle certaines des variables sont restreintes sur le domaine des valeurs entières. L'application de ce principe d'optimisation identifie certaines capacitances stœchiométrique (SC, en anglais) telles que le réseau métabolique soit forcé à produire certaines substances d'intérêt, comme l'éthanol ou certains acides aminés. Au-delà des applications bio-technologiques de ces travaux, l'élément central de cette modélisation est le fait que la maximisation de la fonction objective force tant l'utilisation que la non-utilisation de certaines réactions métaboliques (i.e., certaines réactions deviennent obligatoires ou non). Ce phénomène se traduit par une reconfiguration interne du métabolisme qui peut mener à l'excrétion de certains produits sous certaines conditions environnementales.

Dans un deuxième temps, nous avons étudié si les réactions obligatoires ou non-obligatoires pouvaient être liées à un contexte évolutif. En effet, une théorie écologique, l'hypothèse de la reine rouge (Red Queen Hypothesis, RQH), postule que, dans certaines niches écologiques, l'augmentation de la « fitness » d'un individu est compensée pour une perte de fitness chez les autres participants de la communauté. Afin de maintenir un équilibre évolutif, les autres organismes vont alors avoir tendance à évoluer de manière à maintenir la fitness originale ce qui peut se manifester par une évolutions des organismes pour maintenir une fitness, et ce à des échelles moléculaires et génomiques. Récemment, afin de complémenter la RQH, une autre théorie, appelée hypothèse de la reine noire (Black Queen Hypothesis, BQH), postule que la perte de fonctions peut être aussi est une stratégie évolutive pour peu que la fonction perdue soit procurée par l'environnement. D'un point de vue modélisation, cela implique que les effets de l'évolution ne soient pas équivalents, que l'on s'intéresse aux groupes de réactions obligatoires ou non-obligatoires. Ce phénomène évolutif est étudié chez *Pseudomonas fluorescens*, en utilisant FBA et FVA afin de déterminer le caractère obligatoire ou non-obligatoire des réactions et par transitivité des gènes liées au réseau métabolique et ce afin d'observer la quantité de mutations accumulée au cours du temps pour chacun de ces deux groupes de gènes. Les expériences qui suivent les analyses du modèle montrent que les deux groupes ont des comportements différents, étant probablement dues à l'application de processus liés à la RQH chez les gènes obligatoires et BQH sur les non-obligatoires.

Cependant, au-delà de la simple anticipation des résultats expérimentaux par une modélisation métabolique, nous avons identifié un problème complémentaire. Grâce au lien entre l'optimisation et les CBMs, il est possible de chercher les conditions expérimentales qui maximisent le nombre de réactions d'un type donné ; comme non-obligatoire pour tester expérimentalement l'hypothèse de la BQH. Nous avons formalisé ce problème, qui n'est pas aujourd'hui soluble par les méthodes standard de résolution des contraintes. En effet, cette formulation de problème est associée aux problèmes d'énumération des modes élémentaires (EM), une connexion inattendue mais complexe à résoudre. Ainsi, on notera comme certains problèmes biologiques qui ont des applications pratiques comme l'élaboration des méthodes de cultures microbiennes, peut motiver une recherche plus fondamentale en Informatique (Budinich et al., 2015).

La deuxième partie de cette thèse propose le développement d'une modélisation dédiée aux réseaux métaboliques d'une communauté bactérienne, dans laquelle plus de deux types de bactéries interagissent. En effet, si la plupart des CBMs reposent sur la modélisation d'un système cellulaire, dans un environnement naturel, les conditions sont rarement axéniques. Bien au contraire, les microorganismes forment des communautés via la mise en place d'interaction interspécifiques. Cette observation est la principale motivation au développement de CBMs dédiés à la modélisation du réseau métabolique des communautés microbiennes.

De manière historique, la méthode proposée a été d'envisager schématiquement l'union de toutes les

réactions connues des micro-organismes en présence au sein d'une seule matrice stœchiométrique, et ce sans faire de distinction des organismes. Cette méthode est connue sous le terme de "Lump" ou soupe pour représenter un écosystème. Par ailleurs, d'autres méthodes modélisent le problème cette fois ci en considérant tous les micro-organismes et ce via l'identification d'un réseau métabolique pour chaque phénotype (aussi appelés « guildes »), formant ainsi un enchevêtrement de compartiments spatialement séparés mais en interaction via le milieu environnemental. Chaque guilde est représentée par une matrice stœchiométrique propre.

Pour dissocier les deux hypothèses de modélisation, nous avons comparé les solutions des deux approches de modélisation. Plus particulièrement, nous avons comparé les espaces de solutions respectifs et ce pour deux applications biologiques de tailles différentes, l'une composée de trois organismes différents (*Synechococcus spp.*, phototrophes anoxygeniques filamenteuses et bactéries réductrices de sulfure, (Taffs et al., 2009)) et un système producteur du méthane (*Desulfovibrio vulgaris* et *Methanococcus maripaludis*, (Stolyar et al., 2007)). En utilisant la méthode de Flux Modules (Müller and Bockmayr, 2013), les espaces de solutions sont décomposés et comparés. En général, les deux espaces sont différents et pointent des differences attendues sur les résultats de FBA et FVA. Ces résultats sont en accordance avec ceux déjà observés par Klitgord and Segrè (2009), et mettent en valeur la nécessité d'une CBM dédiée pour simuler des communautés bactériennes qui tienne compte (i) de différents compartiments pour différents organismes et (ii) avec des objectifs propres à chaque compartiment pour rendre compte de la croissance de chaque phénotype au sein d'une communauté.

Pour répondre à cette problématique, une étude de la littérature montre qu'il est possible de considérer plusieurs objectifs ; une approche multi-objectif; mais appliquée à un unique réseau métabolique. Par ailleurs, d'autres méthodes permettent de considérer différents compartiments, mais les algorithmes de résolution de ces problèmes, globalement, considèrent une unique fonction objective pour représenter l'écosystème, et ce malgré la présence de plusieurs compartiments. La difficulté pour intégrer les deux approches réside dans le fait d'identifier une solution multi-objectif.

La littérature informatique propose des méthodes qui utilisent une fonction pour décrire les objectifs multiples d'un système. Ces méthodes sont appelés "scalarization methods" et permettent de considérer plusieurs fonctions objectives et de construire une fonction qui retourne une valeur réelle, permettant l'application des algorithmes standards des problèmes mono-objectif.

Ainsi, il est possible d'interpréter le concept de maximiser plusieurs objectifs sous la forme d'un front de Pareto. Dans ce contexte, nous pouvons alors comparer des objets qui sont alors des vecteurs et non des nombres pour les approches mono-objectif. Cependant, comme dans les espaces vectoriels, il n'y aura pas de relation d'ordre total mais seulement un concept d'ordre partiel, dans lequel la notion de maximum se traduit comme une collection de valeurs possibles formant ainsi le "front de Pareto".

Motivés par ces résultats théoriques, nous proposerons une CBM basée sur le front Pareto comme étant la solution à étudier. À partir des réseaux métaboliques existants, une procédure qui permette de construire un réseau métabolique de l'écosystème microbien est proposée comme ayant un composant central, le milieu, permettant une espace d'échange des métabolites entre les différentes souches. Une fois la matrice stœchiométrique globale modélisée, nous avons proposé la mise en place d'une FBA multiobjective (MO-FBA) via l'application d'un algorithme nommé "Benson Outer Approximation" qui retourne une description géométrique du front de Pareto en utilisant ses pointes extremes. Par extension, cette même description géométrique permet de définir un FVA multiobjective (MO-FVA).

Cette méthode a été appliquée à une communauté constituée de trois souches bactériennes: *Synechococcus spp.*, des phototrophes anoxygeniques filamenteuses et bactéries réductrices de sulfure, (Taffs et al., 2009). L'application de MO-FBA et MO-FVA permet d'extraire de nombreuses informations à partir du modèle. Premièrement, le point de production maximale de la biomasse de la communauté ne correspond pas aux productions optimales de chaque organisme ; mais au contraire dans des conditions de croissance sous-optimale (par rapport à des optimum individuels). Plus particulièrement, cette modélisation permet de suivre avec précision les échanges d'azote et de carbone au sein de l'écosystème mais aussi au sein de chaque métabolisme microbien. La description du front de Pareto permet aussi d'inclure une restriction

permettant l'exploration de l'ensemble des solutions proches du front en utilisant d'autres critères. En effet, il est possible de suivre un critère de production d'entropie. Les résultats montrent alors que la production expérimentale se situe à l'interface entre la production maximale de biomasse de l'écosystème et la production d'entropie maximale, ce qui ouvre la possibilité à de nouvelles approches pour comprendre les équilibres des écosystèmes microbiens.

En conclusion, les CBMs sont capables de faire face aux défis issus de la quantité croissante d'information génétique. Par ailleurs, le lien entre la fitness écologique et les fonctions objectives des CBMs des systèmes cellulaires permet d'étendre les applications des CBMs aux problématiques écologiques. Notamment, le développement d'une méthode dédiée pour modéliser des communautés permet de relier ces deux champs scientifiques, en donnant un ensemble d'outils pour quantifier les relations au sein d'un écosystème décrit de manière holistique.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisors Damien Eveillard, Jérémie Bourdon and Bruno Saint-Jean for the continuous support of my Ph.D study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. They make me feel more of a colleague rather than a student. Besides my advisors, I would like to thank to my thesis committee: Monique Zagorec, Laurence Garczarek, Fabien Jourdan and Olivier Bernard, for their insightful comments. My sincere thanks also goes to CNRS and GRIOTE project for founding my Ph.D thesis.

I thank my fellow lab-mates at LS2N (former LINA) for the stimulating discussions and for all the fun we have had in the last four years. Also I thank my friends and former lab-mates at Laboratory of Bioinformatics and Mathematics of the Genome, at the University of Chile. In particular, I am grateful to Alejandro Maass for enlightening me the first glance of research.

A special thanks to my friends from all over the world. I met some of them in Chile, others in France and some while abroad. Shared experiences and conversations helped a lot to make this thesis, even (and perhaps specially) when we were not talking about it.

Last but not the least, I would like to thank my family: Karina and Luka; my parents, brother, sister, aunts, uncles and cousins; Karina's family and my extended family around the world, for supporting me throughout this experience and my life in general.

Contents

1

Introduction

Since the second half of 20^{th} century, biology experienced an amazing revolution, mainly propelled by the elucidation of DNA structure and its central role in biological processes. In addition, thanks to advances in biochemical techniques, the last 30 years have sought an explosion of sequence data, both in terms of number of sequences and completed genomes (Figure 1.1).

Among others, computer sciences plays a pivotal role in this revolution. For example, in 1960's, Margaret Oakley Dayhoff wrote a series of FORTRAN programs running in an IBM 7090 computer, which determined all consistent amino-acid sequences from overlapping partial digestion peptides (Hagen, 2000). Later on, computational methods were used also to cover data representation (Lipman and Pearson, 1985), similarity search (Altschul et al., 1990) and data storage (Hagen, 2011). These works highlight the importance of interdisciplinarity in solving open questions in life sciences.

From the interaction between **computer sciences**, **mathematics** and **molecular biology** three new subdisciplines are recognized nowadays, with a high overlap between them: *Bioinformatics* focuses mostly in the process and analysis of high volumes of information, such as provided by high throughput experiments in transcriptomics, metagenomics or metabolomics. *Computational Biology* deals with the development of models to study biological systems (Huerta et al., 2000). *Systems Biology* takes an interest in system level behavior of biological processes. Such systems are described by the interaction of their molecular constituents, conforming complex biological networks (Kitano, 2002a,b). Here, "complex networks" must be understood as networks that present emergent properties, *i.e.,* unexpected behaviors that stem from interactions between their components and the environment (Johnson, 2006).

Systems Biology has been successful into analyzing heterogenous data sets at molecular scales, providing insights into underlying processes (Kitano, 2002a,b). Increase in computer power and data availability enabled Systems Biology to move forward from small size networks to whole microorganism systems (Joyce and Palsson, 2006). As microorganisms are the most diverse and abundant cellular life forms on Earth, with estimates from 25% to 50% of earth total biomass (Whitman et al., 1998; Kallmeyer et al., 2012; Rinke et al., 2014), their holistic study remains an active research field.

In natural environments, microorganisms form communities with complex interactions, playing crucial roles from biogeochemical cycles (Lin et al., 2000; Rullkötter, 2006) to human health (Vieira-Silva et al., 2016). Therefore, studying microbial ecosystems appears as a natural progression for Systems Biology. In particular, *Network Inference* methods have been used in microbial ecology context. Generally speaking, Network Inference methods use a *"who is there and who is not"* rationale: Relative abundance data (or presence/absence information) is used to construct co-occurrence matrices, indicating when two agents of the system appear together in a sampling experiment (Raes et al., 2011). Next, tools and analysis from graph theory are used to extract relations between different agents, under the hypothesis that strong non-

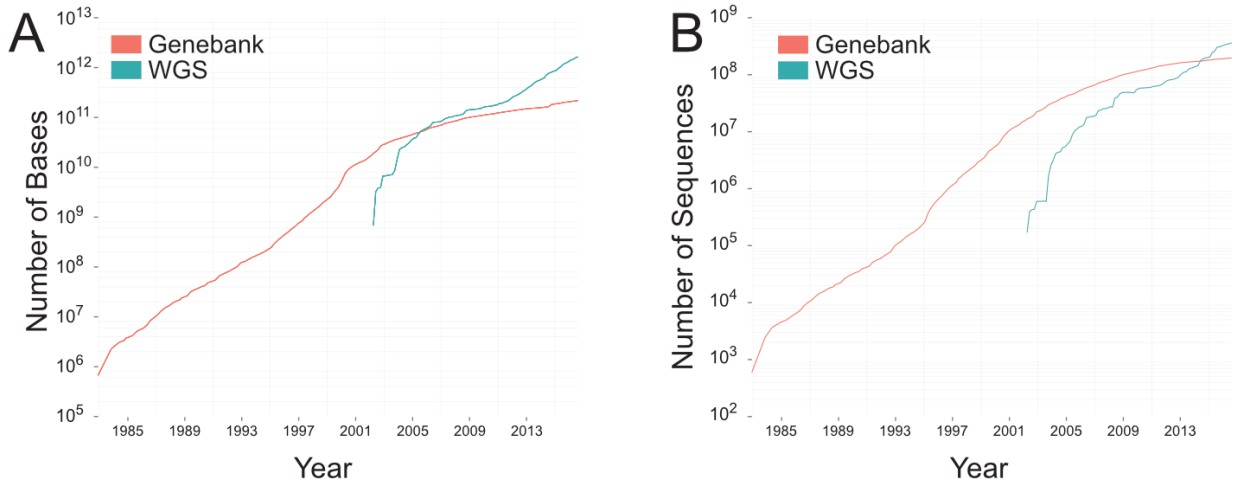


Figure 1.1 – Accumulation of genomic knowledge in NCBI Genbank and in Whole Genome Sequences (WGS). A) Number of Bases per year B) Number of Sequences per year . *Source:* NCBI

randomly correlations are due to biological reasons (Zhang and Horvath, 2005). As result, subsets of tightly related agents are obtained. When applied to microbial ecosystems, agents correspond to species or families of microorganisms (Faust and Raes, 2012). Linking environmental variables with such results highlights influence of groups of microorganisms in observed phenomena or vice-versa (Guidi et al., 2016).

However, while Network Inference methods are extremely useful to uncover relations between community members and their interactions with environmental variables, they are not well suited to quantitatively predict ecosystem functions. Building models able to quantitatively predict community functions and dynamics has been pointed as a key challenge on microbial ecology (Widder et al., 2016). Worth noticing, quantitative modeling in microbial ecology exist prior to Systems Biology era and a substantial amount of work has been done in this line. A revision of principles for building models in microbial ecology is then necessary to understand current approaches and needs of the field.

As studies in microbial ecology focus on populations or communities interacting in dynamics of ecosystems, models used in this field aim to quantify population changes according to environmental variables. To set up a model, different equations from exact sciences (*e.g.*, physics and chemistry) as well as natural sciences (*e.g.*, biology and ecology) are selected. Usually, relationships between variables are characterized by a series of parameters that are *a priori* not know, requiring successive steps of estimation and validation using experimental data. For instance, when modeling a microbial culture, organism dynamics are characterized by the specific growth rate μ, which quantifies the grams of biomass produced by 1 gram of biomass in 1 hour, and microbial metabolism is summarized by cellular yield $Y_{X/S}$ (biomass production / substrate consumption) and maintenance coefficient m (quantifying amount of substrate consumption not related to growth). Next, these quantities are linked to state variables, *i.e.*, a set of variables that describe the system such as total biomass, substrate concentration, metabolite production and total volume. These state variables are usually deduced from balance equations, according the particular process under modeling (Poggiale et al., 2014).

A second type of models rather determines relationships between specific growth rate μ and environmental factors. For instance, under limiting substrate condition, μ is often characterized by Monod's law (Monod, 1949). In Droop (1968) model, μ is related to an internal quota Q, a new state variable, which represents the amount of limiting substrate inside the cell after consumption. This approach is well suited to represent the substrate accumulation by microorganisms.

To further represent microbial communities, the effects of one population upon another should be included. The usual approach consists into describing the specific growth rate of a population as a function of densities of other populations. The Lotka-Volterra model (Lotka, 1925; Volterra, 1926) and models based in the competitive exclusion principle (Gause, 1934) are, among others, well known examples.

Other types of modeling include spatial representation, which are useful to represent phenomena such as biofilm formation. As the complexity of the biological system increases, it is often not possible to define behavior through simple equations. This problem is generally overcome by introducing "Individual Based Models" in which each cell is located in spatial coordinates and their behavior depends on an algorithm that make decisions based in the volume it occupies (Picioreanu et al., 1998, 2004); next, simulations are run iteratively to observe if from these set of rules more complex behaviors arise (Grimm et al., 2006). In the case of biogeochemical models, the objective is to describe the dynamics of one or more elements (carbon, nitrogen, phosphate, sulfur, etc.) for a given ecosystem. Usually, different organisms are classified in functional groups and then flows of elements of interest are represented by adequate state variables. Environmental parameters are then introduced (Poggiale et al., 2014). In this context, "Trait Based Models" link specified traits, *i.e.*, properties at an individual scale (*e.g.*, size and concentration) to ecological function, such as energy and/or matter flux, primary production, acid production, etc. (Krause et al., 2014).

From this overview, the role of parameters such as specific growth rate, cellular yield, substrate consumption and traits, for example, are central in current modeling approaches. Unfortunately, these parameters are not available for all microorganisms and usually need to be calculated using extensive experimental data and/or validated by experts in a particular application. Furthermore, values obtained *in-vitro* can differ from *in-vivo* conditions. In addition, recent sequence data are not integrated explicitly in the models.

To overcome these issues the use of "Genome Scale Models" has gained a lot of interest recently. In this framework, models containing detailed genomic information are used to describe organism physiology. Then, chemical species are defined and information about their exchange with the media (for example, maximal uptake rate) is included (Hanemaaijer et al., 2015; Biggs et al., 2015; Perez-Garcia et al., 2016). Genome Scale Models are able to take into account full genomic descriptions of microorganisms and can readily be exploited by several techniques. In particular, since its popularization 25 years ago (Varma and Palsson, 1994a,b), Constraint Based Models (CBM) had become a widely used and flexible approach to exploit Genome Scale Models for the sake of microorganisms physiological characteristics exploration. CBMs techniques enable calculation of specific growth rates, substrate consumption, metabolite production and cellular yields in several conditions without using additional parameters (Orth et al., 2011; Schellenberger et al., 2011). This embeds CBM with the ability to make predictions both into genomic and cellular scale using the same framework. Furthermore, a majority of related methodologies have a strong mathematical basis with standard computational methods implemented in several platforms (Bordbar et al., 2014). With the accumulation of information in several databases, along with an increasing computing power, today we are in a flourishing time for genome-wide modeling of microorganisms (Kim et al., 2012; Bordbar et al., 2014). CBM hold great promises in microbial modeling at several scales, providing a way to perform data integration and a mathematical description suitable for numerical simulations (Hanemaaijer et al., 2015).

The present thesis is framed in this interdisciplinarity between biology, mathematics and computer sciences. In particular, our main objective was to apply and develop a Systems Biology approach based on CBM to understand the relation between environmental factors and metabolic responses in microorganisms, linking molecular and environmental processes. Methods and analysis will focus on the metabolism as its constitutes the first and most direct layer interacting with the media (Varma and Palsson, 1994a,b). In particular, interpretation of specific growth rate as a fitness measure, will enable connections between CMB and other concepts in ecology. Additionally, as CBMs contains comprehensive descriptions of genes involved, analysis of genome-scale datasets will be guided by CBM, uncovering novel relations based in real chemical connections.

This manuscript is divided in two sections. Section I focuses on impacts to and from environment while considering single metabolic networks. First, CBM formalisms will be introduced in chapter 2 by detailing their construction principles, how this modeling relies genes to fitness and how one can mathematically represent it. Relevant ideas of the field will be introduced by describing two of the most widely used tools based on CBM: Flux Balance Analysis and Flux Variability Analysis. Next, in chapters 3, 4 and 5, applications of CBM in different contexts will be discussed. In chapter 3 an application of a recently described

CBMs, called Stoichiometric Capacitance, will be used to design synthetic strains by inserting genes such as their specific growth rate (*i.e.*, fitness) increases. While the main focus will be biotechnological applications, such as ethanol or amino acid production, this chapter will illustrate CBM flexibility to capture the effects of genome insertion on the organism fitness; in particular, how a metabolic network reconfiguration can lead to changes in the export of metabolites. Chapters 4 and 5 will explore a different problem: How environmental conditions could alter a gene function. In chapter 4, a recent hypothesis called Black Queen Hypothesis (BQH) will be analyzed experimentally in *Pseudomonas fluorescens* and results will be compared with predictions made by Flux Balance Analysis and Flux Variability Analysis given a set of culture conditions. Based in those predictions, genes related to this metabolic network will be classified by their relevance to fitness function; genes not relevant to fitness function are expected to accumulate mutations in contrast to those who are critical to fitness. In the same context, a new CBM formulation aiming to find culture conditions such as number of non-relevant genes will be maximal, in order to increase observation of BQH at gene level, will be developed in chapter 5.

Section II is motivated by the specific need of expanding CBM to model microbial communities. First, in chapter 6, two modeling assumptions about how different agents should be considered within CBM will be compared. Indeed, either (i) all genes of each species is considered in a single entity or compartment and (ii) genes are considered spatially separated in multiple compartments, each compartment representing one species. Both modeling scheme will be applied to two different microbial ecosystem models. The corresponding set of all possible values (a mathematical space called "solution space") to each modeling approach (*i.e.*, single and multiple compartments, respectively) will be compared. By the use of dedicated techniques, it will be shown that the structures of respective solution spaces differ if one or multiple compartments are considered, confirming previous results in the literature (Klitgord and Segrè, 2009). By these observations, it is concluded that the use of several compartments should be a key element in modeling microbial communities. Motivated by this, chapter 7 will discuss several approaches considering multiple compartments in CBM. As each compartment is considered as a single organism, a particular focus will be on how objectives corresponding to each compartment are considered within each CBM. Finally, in chapter 8, a new CBM framework to model microbial ecosystems will be developed. By considering a set of metabolic networks, an ecosystem model will be systematically built. Next, a CBM considering all objectives simultaneously will be proposed, showing solutions along a geometrical description in the space of objective functions, called "Pareto Front". To compute this Pareto Front, a recent computational algorithm will be used. As the Pareto Front corresponds to a set of optimal values, it will be shown how this optimal space can be explored using thermodynamic criteria to give insights on the role of additional physical principles in ecosystem diversity.

Modeling Single Microorganisms: Effects from and to the Environment

2

Constraint Based Models for Metabolic Modeling

As discussed previously, Constraint Based Models (CBM) is a suitable modeling approach to cope with increasing genomic information. The present chapter presents a wide view of the CBM state of the art. First, principles involved in modeling metabolic reactions will be discussed in detail. Next, a review of the main hypotheses and representative methodologies to analyze corresponding models will be presented.

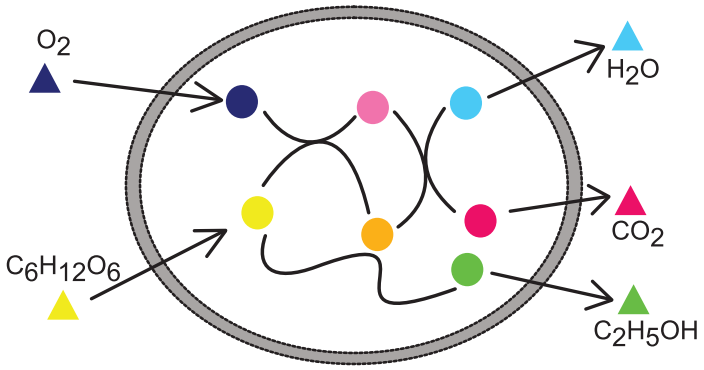
2.1 Metabolic Modeling of Reactions

One of the defining characteristics of living organisms is their capacity to use chemical species as available in their environment in order to reproduce themselves. Furthermore, presence or absence of chemical species are highly correlated with phenotypes and ecotypes observed. For example, *Escherichia coli* cultivated under aerobic conditions (*i.e.* high O_2) transforms glucose and oxygen into carbon dioxide and water; likewise, under anaerobic conditions (*i.e.* low O_2), it transforms glucose into carbon dioxide and ethanol (among other fermentations products, neglected in this example). Both reactions are illustrated in Figure 2.1, where $C_6H_{12}O_6$, O_2, CO_2, H_2O and C_2H_5OH are the chemical formulae for glucose, oxygen, carbon dioxide, water and ethanol, respectively. Numbers before chemical formula are called **stoichiometric coefficients** and indicate the amount of moles involved in such transformation; usually, stoichiometric coefficients equal to 1 are not explicitly written. Chemical compounds consumed during the reaction, called substrates, are placed left to the symbol $\rightarrow$ and chemicals originated by the reaction, called products, are placed right.

The set of chemical transformations within an organism is called *metabolism*. The set of reactions constitutes a *metabolic network*, where products of some reactions are used as substrates of others. In general, all small compounds involved in this metabolic network are called *metabolites*.

Most of reactions are catalyzed by enzymes encoded in the microbial genome. Therefore, a reasonable approximation of an organism metabolism is given by the set of enzymatic reactions, complemented with other known spontaneous reactions. For several organisms, this knowledge is extracted directly from their genome annotation. Information about reactions themselves is stored in specialized databases (Kanehisa et al., 2013; Caspi et al., 2014). Very often, this information is represented as a graph, where metabolites and reactions are described by nodes and edges, respectively (Acuña et al., 2009; Bordbar et al., 2014).

Metabolic networks can be described also trough their stoichiometric matrix $\mathbf{S}$. To this end, a matrix is constructed where metabolites are represented as rows of the matrix, reactions as columns and stoichio-



$$1\ C_6H_{12}O_6 + 6\ O_2 \longrightarrow 6\ CO_2 + 6\ H_2O \qquad (r_1)$$
$$1\ C_6H_{12}O_6 \longrightarrow 2\ C_2H_5OH + 2\ CO_2 \qquad (r_2)$$

Figure 2.1 – Toy metabolic Network, illustrating aerobic and anaerobic utilization of glucose. Grey ring represents cell membrane. Circles represent chemical species inside the microorganism, whereas triangles represent chemical species outside cell membrane. Different colors represent different chemical species. Arrows indicate transport from media to intracellular space and vice versa, whereas chemical reactions are represented by curved lines.

metric coefficients as entries. By convention, coefficients of substrates are taken as negative and products positive. If a metabolite is not used by the reaction, then it has a 0 coefficient. For example, for reactions r_1 and r_2 in Figure 2.1, their corresponding stoichiometric matrix is given by

$$\mathbf{S} = \begin{array}{r} \\ C_6H_{12}O_6 \\ O_2 \\ CO_2 \\ H_2O \\ C_2H_5OH \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{r_1} & \mathbf{r_2} \end{array} \\ \begin{bmatrix} -1 & -1 \\ -6 & 0 \\ 6 & 2 \\ 6 & 0 \\ 0 & 2 \end{bmatrix} \end{array}$$

In general, we are interested in a description of how metabolites are exchanged and transformed. For a general chemical equation

$$a\ M_1 + b\ M_2 + \ldots \longrightarrow \ldots + p\ M_{k-1} + q\ M_k$$

where M_1 to M_k represents k different chemical species. The velocity at the reaction takes place, or reaction rate, is defined as

$$v = -\frac{1}{a}\frac{d[M_1]}{dt} = -\frac{1}{b}\frac{d[M_2]}{dt} = \ldots = \frac{1}{p}\frac{d[M_{k-1}]}{dt} = \frac{1}{q}\frac{d[M_k]}{dt}$$

where $[M_i]$ is the concentration of specie M_i. In general, rate of reactions depends on the concentration of species and conditions such as temperature, so they usually take the form of functions like $v = k(T)[M_1]^{\alpha}[M_2]^{\beta}\ldots[M_k]^{\gamma}$ where T is the temperature. $k(T)$, α, β, ... , γ are known as kinetic parameters. If these parameters were known, time evolution of M_i concentration ($[M_i]$) could be determined as a function of time by resolving the resulting differential equation. Unfortunately, determining these parameters and even the function of reaction rate are complex experimental tasks. Moreover, these parameters are in general very sensitive to biochemical conditions, such as pH and cytoplasm ionic strength, so *in vitro* determinations may not correspond with *in vivo* values (Edwards and Palsson, 2000).

In a metabolic network context, temporal evolution of $[M_i]$ is determined by all reactions taking place at a given moment. In general, if they are r_1 to r_n reactions involving metabolite M_i with a_{i1} to a_{in} stoichiometric coefficients, then the mass balance equation for $[M_i]$ is described by

$$\frac{d[M_i]}{dt} = a_{i1}v_1 + a_{i2}v_2 + \ldots + a_{in}v_n = \sum_{j=1\ldots n} a_{ij}v_j$$

Using vector notation and defining column vectors $\mathbf{K} = ([M_1], [M_2], \ldots, [M_n])$,for concentrations of metabolites, and $\mathbf{v} = (v_1, v_2, \ldots, v_n)$, for reaction velocities, it is possible to write succinctly

$$\frac{d\mathbf{K}}{dt} = \mathbf{Sv}$$

In this context, components v_i of $\mathbf{v}$ are called **fluxes** and $\mathbf{v}$ is called **flux vector**. To complete the system description, it is necessary to fix a control volume where this mass balance is being applied. It is common practice to pick this volume as the cell itself and their immediate surroundings. Space out of this control volume is identified with the external environment and metabolites beyond system boundaries are called external metabolites. External metabolites are taken in or out of the system trough exchange reactions of the form "$M_{i\,\mathbf{ext}} \longrightarrow M_i$" or "$M_i \longrightarrow M_{i\,\mathbf{ext}}$" for *in* and *out* metabolites, respectively. The **ext** subscript indicates that if M_i is located in an external compartment. Rates of exchange reactions are interpreted as the uptake or excretion rates from the medium, and they can be inferred from chemostat or batch experiments.

Exchange reactions are then included in the stoichiometric matrix. For example, to complete the small model of *E. coli* containing r_1 and r_2, adding exchange reactions $\mathbf{r_{ex1}}$ to $\mathbf{r_{ex5}}$ (*i.e.*, one for each metabolite), results in the following system:

$$\frac{d\mathbf{K}}{dt} = \frac{d}{dt}\begin{pmatrix} [C_6H_{12}O_6] \\ [O_2] \\ [CO_2] \\ [H_2O] \\ [C_2H_5OH] \\ [C_6H_{12}O_6]_{\mathbf{ext}} \\ [O_2]_{\mathbf{ext}} \\ [CO_2]_{\mathbf{ext}} \\ [H_2O]_{\mathbf{ext}} \\ [C_2H_5OH]_{\mathbf{ext}} \end{pmatrix} = \begin{array}{c} \begin{matrix} \mathbf{r_1} & \mathbf{r_2} & \mathbf{r_{ex1}} & \mathbf{r_{ex2}} & \mathbf{r_{ex3}} & \mathbf{r_{ex4}} & \mathbf{r_{ex5}} \end{matrix} \\ \begin{bmatrix} -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -6 & 0 & 0 & 1 & 0 & 0 & 0 \\ 6 & 2 & 0 & 0 & -1 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{bmatrix} v_1 \\ v_2 \\ v_{ex1} \\ v_{ex2} \\ v_{ex3} \\ v_{ex4} \\ v_{ex5} \end{bmatrix} = \mathbf{Sv}$$

assuming that $C_6H_{12}O_6$ and O_2 are up-taken and CO_2, H_2O and C_2H_5OH excreted (Figure 2.1).

2.2 Constraint Based Models for Genome Scale Models

The set of biochemical reactions encoded by a whole genome is usually called a Genome Scale Model (GEM). GEMs are usually the starting point to develop realistic metabolic models of a given organism. In general, the procedure to obtain a metabolic model is bottom-up. It starts with the knowledge as contained in a genomic sequence and follows a semi-automated protocol that ends with a metabolic network reconstruction. Then, usually there is one to multiple human expert revisions in order to select a subset of reactions (based either in evidence of enzymes found in their genome sequence and/or previous knowledge of organism physiology) that represent the synthesis of the main biological compounds (amino acids, sugars, lipids, proteins, DNA, RNA, etc...) from external nutrients (imported from external media by simple diffusion or more specialized transporter proteins). Finally, the whole set of biosynthesis processes of a new organism will be abstracted into a single pseudo-reaction, called "biomass function" (a function which

represents the growth of the organism itself). The main objective is to provide a model, understood as a set of biochemical reactions linked within a network, which represents the metabolic capabilities of a given microorganism (Thiele and Palsson, 2010).

However, solving $\frac{d\mathbf{K}}{dt} = \mathbf{Sv}$ is a daunting task for genome scale systems. For instance, recent genome scale metabolic model of *E. coli* contains around 700 metabolites and 1300 reactions (Orth et al., 2011), describing a system of at least one differential equation for each metabolite and one polynomial equation for each flux, provided that reaction rates can be described as polynomials and kinetics parameters are known. Furthermore, initial or boundary conditions for concentrations should be also specified (Varma and Palsson, 1994a).

Despite these difficulties, biological relevant cases to study remain available. In particular, organisms are known to be homeostatic, keeping internal concentration as constant as possible by means of regulation; furthermore, the changes of intracellular metabolites concentration occurs at very fast rates (Varma and Palsson, 1994a). This behavior corresponds to quasi-steady-states of the system, where $\frac{d[M_i]}{dt} = 0$ for internal metabolites. Therefore, for such conditions we have

$$\mathbf{Sv} = \left[\mathbf{S_{int}} \mid \mathbf{S_{ext}}\right] \mathbf{v} = \begin{bmatrix}\mathbf{0} \\ \mathbf{L}\end{bmatrix} = \mathbf{b}$$

where $\mathbf{L}$ represents known exchange reactions rates and prefixes **int** and **ext** represents the portion of internal and exchange reactions respectively. It is important to notice that equation $\mathbf{Sv} = \mathbf{b}$ characterizes all steady state solutions for $\mathbf{v}$, constraining their possible values. Besides, it is possible to include additional information about the fluxes. For instance, we can set boundaries for the values of a flux v_i, such as $l \leq v_i \leq u$. Then, our metabolism model is defined by a set of equations that **constrains** possible solutions. This type of model is called a **Constraint Based Model** (CBM).

In a broad sense, methods used in CBM explore the solution space of equations, *i.e.* the space containing all fluxes $\mathbf{v}$ which satisfies $\mathbf{Sv} = \mathbf{b}$. Usually this space is termed **flux space**. It is then possible to distinguish two types of methodological approaches. The first approach considers all valid solutions of the constrained problem and it focuses in describing the solution space (Acuña et al., 2009). We term this family of approaches *Flux Space Description*. The second approach focuses in fluxes that optimises a given *objective function*. The rationale behind the optimization of an objective function is based in the observation that optimal individuals possess advantages, prone to be selected by natural processes (Varma and Palsson, 1994a).

2.2.1 Flux Space Description

If no particular solution of the steady state is preferred, then all feasible solutions satisfying $\mathbf{Sv} = \mathbf{b}$ should be considered:

Definition. *Given a metabolic network, the steady-state flux space F is defined as*

$$F := \{\mathbf{v} \in \mathbb{R}^{|\mathcal{R}|}, \mathbf{Sv} = \mathbf{b}, \boldsymbol{l} \leq \mathbf{v} \leq \boldsymbol{u}\}$$

where $\mathcal{R}$ denotes the set of all reactions. In the following, we will denote by $\mathcal{M}$ the set of all metabolites.

A sensible strategy to deal with this space is to find a set of $\mathbf{v}$ that could be used as a basis to generate space F. **Elementary Modes** (Schuster and Hilgetag, 1994; Stelling et al., 2002; Klamt and Stelling, 2003; Zanghellini et al., 2013) describes such a generating set. Defining the support of a flux $\mathbf{v}$, $\mathbf{supp}(\mathbf{v}) = \{i : v_i \neq 0\}$, $\mathbf{e} \in F$ is an elementary mode if its support cannot be written as a proper superset of any other feasible mode $\mathbf{v}$, *i.e.* $\mathbf{supp}(\mathbf{e}) \not\supset \mathbf{supp}(\mathbf{v})$. With this, every steady state flux distribution can be represented as a weighted superposition of elementary modes with non-negative weights (Zanghellini et al., 2013). Besides elementary modes, the concept of Extreme Pathways has been introduced in the literature (Schilling et al., 2000). Both concepts are closely related, as it has been shown that Extreme Pathways are a subset of Elementary modes (Klamt and Gilles, 2004).

From a biological perspective, an elementary mode can be interpreted as a generalized pathway (Schuster et al., 2000) which conveys useful interpretations. However, in practice, the number of elementary modes increases exponentially with the number of reactions. Furthermore, the complexity of enumerating elementary modes remains as an open question (Acuña et al., 2010). According to empirical observations, the running time is approximately quadratic in the size (Terzer and Stelling, 2008).

To cope with the huge number of elementary flux modes in genome-scale metabolic networks, Müller and Bockmayr (2013) proposed the concept of **flux modules**:

Definition. *A non empty set of reactions $A \subseteq \mathcal{R}$ is called a flux module w.r.t. a flux space $P \subseteq \mathbb{R}^{|\mathcal{R}|}$ (shortly A is called a P-module or just a module), if there exists a vector $d \in \mathbb{R}^{|\mathcal{M}|}$ with $\mathbf{S_A v_A} = d, \forall$ flux vectors $\mathbf{v} \in P$. The vector d is called the* right-hand side *of the P-module A. Since d operates as the interface of the P-module to the rest of the network, we refer to d also as the* interface flux *of A.*

Intuitively, modules are set of reactions that behave as one w.r.t P, given that they have a common interface to the rest of the network. By taking the set of minimal modules (minimal in the sense of modules not contained by other modules), it is shown that flux spaces can be partitioned into these minimal modules; furthermore, this decomposition is unique. Therefore, modules can be used also as a description of the flux space.

2.2.2 Optimization Based Techniques

As stated before, methods that deal with whole flux spaces can be cumbersome to explore and hard to interpret. Therefore, usually the analysis is restricted to "interesting" fluxes. In particular, we are interested into fluxes that maximizes certain "objective function". As said before, scientific rationale of optimal flux distributions is supported by natural selection: individuals being optimal present advantageous traits and therefore those individuals are prone to be selected (Varma and Palsson, 1994a).

Flux Balance Analysis (FBA) approach formalises these concepts. It relies on Linear Programing (LP) to determine a steady-state distribution of fluxes. It is important to note that the function $\mathbf{c}$ to maximize (or minimize), is usually related to biomass, but ATP production or overall sum of fluxes can be used also. The formulation is then as follows:

$$\begin{aligned} \text{maximize} \quad & z = \mathbf{c}^\intercal \mathbf{v} \\ \text{subject to} \quad & \\ & \mathbf{S_{int}} \mathbf{v} = \mathbf{0} \\ & \mathbf{S_{ext}} \mathbf{v} = \mathbf{L} \\ & l_i \leq v_i \leq u_i \quad i = 1, \ldots, n \end{aligned}$$

$\mathbf{S}$ is the stoichiometric matrix with m metabolites and n reactions (therefore, $\mathbf{S} \in \mathbb{R}^{m \times n}$). $\mathbf{v} \in \mathbb{R}^n$ is the flux vector and is determined by optimization; v_i is the i^{th} component of $\mathbf{v}$. If the value of $v_i \leq 0$ we assume that the reaction is occurring in the reverse sense, *i.e.*, from products to substrates. $\mathbf{S_{int}} \mathbf{v} = \mathbf{0}$ is the mass balance for internal metabolites and $\mathbf{S_{ext}} \mathbf{v} = \mathbf{L}$ are the exchange rates of external metabolites. l_i and u_i are the upper and lower boundaries of v_i. If a reaction r_i is irreversible (*i.e.* can only go from substrates to products), then $0 = l_i \leq v_i \leq u_i$. $\mathbf{c}$ is the objective function, a vector of n coefficients c_i for the v_i fluxes.

Similar formulations can be found in literature, somewhat equivalents (e.g, if we consider $\mathbf{v} \geq \mathbf{0}$, then we can expand $\mathbf{S}$ and attach the columns corresponding to reversible reactions with inverted sign and set all $l = 0$). It is known from Linear Programing that if an optimal value $\mathbf{z}$ exists, it is unique, but unfortunately it is not possible to guarantee the same for $\mathbf{v}$. Therefore, there are (*a priori*) many flux distributions that could yield optimal objective function values (see Raman and Chandra (2009) for a focused introduction and review on FBA).

Flux Variability Analysis (FVA) technique was developed to explore these multiple optimal flux distributions (Mahadevan and Schilling, 2003). In this approach, we are interested into fluxes that satisfy a

given an optimal value, which will give us feasible values of flux spaces. If we denote as $\mathbf{z_{obj}}$ the objective value (e.g, calculated through a previous FBA) then we can formalize the following LPs problems for each $v_i \in \mathbf{v}$:

Case 1		Case 2	
maximize	v_i	minimize	v_i
subject to		subject to	
	$\mathbf{c}^\intercal \mathbf{v} = \mathbf{z_{obj}}$		$\mathbf{c}^\intercal \mathbf{v} = \mathbf{z_{obj}}$
	$\mathbf{S_{int}} \mathbf{v} = \mathbf{0}$		$\mathbf{S_{int}} \mathbf{v} = \mathbf{0}$
	$\mathbf{S_{ext}} \mathbf{v} = \mathbf{L}$		$\mathbf{S_{ext}} \mathbf{v} = \mathbf{L}$
	$l_i \leq v_i \leq u_i, \quad i = 1, \ldots, n$		$l_i \leq v_i \leq u_i, \quad i = 1, \ldots, n$

Note that we have $2n$ LP problems to solve. As result of this procedure, we will obtain a range of values for each flux for the given objective value. This allows to explore the solution space surrounding the given set of conditions, which is more realistic to investigate concrete biological problems.

2.3 Solving Constraint Based Models

From a mathematical perspective, CBM formulations correspond to a class of problems know as "Optimization Problems". In general, a mathematical optimization problem has the following structure:

$$\begin{aligned} \text{maximize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq b_i, \quad i = 1, \ldots, l \end{aligned}$$

Where x is a vector of n components , *i.e.*, $x = (x_1, \ldots, x_n)$ and $x \in \mathbb{R}^n$. In this context, x is called *optimization* or *decision* variable of the problem. $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *objective* function and functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are the *constraint functions*. Constants $b_1, \ldots, b_m$ are the limits or *bounds* for the constraints. A vector x^* is called *optimal* or a *solution* of the problem if the value of $f_0(x)$ is the smallest among all vectors satisfying the constraints; *i.e*, for $z \in \mathbb{R}^n$ such as $f_1(z) \leq b_1, \ldots, f_m(z) \leq b_m$ we have $f_0(z) \geq f_0(x^*)$. Note also that maximize $f_0(x)$ = minimize $-f_0(x)$, so they are equivalent problems (Boyd and Vandenberghe, 2004).

When functions $f_0, \ldots, f_m$ are linear, we obtain a particular class of optimization problems called *linear programing (LP)*:

$$\begin{aligned} \text{maximize} \quad & c^\intercal x \\ \text{subject to} \quad & a_i^\intercal x \leq b_i, \quad i = 1, \ldots, m \end{aligned}$$

Where $c, a_1, \ldots, a_m \in \mathbb{R}^n$. Easily one can see that by picking and appropriate $\mathbf{c}$ and making $f_i(x) = \sum_j A_{ij} v_j = b_i, i = 1 \ldots m$, where A is a suitable matrix for representing the constraints, we can effectively represent FBA and FVA problems as LP. Furthermore, many algorithms related with Elementary Modes and Flux Modules can be implemented using LP (Acuña et al., 2009).

Current algorithms for solving LP, such as simplex and interior point methods, are implemented in several optimization packages. In general, LP have been shown to be solvable in polynomial time. In practice, however, simplex method is fast, although it has only been shown to run in polynomial time in the average case (Schrijver, 1998). For most of the practical applications, we can say that solving LPs is a mature technology, in the sense that solvers for these problems are embedded in such tools and applications (Boyd and Vandenberghe, 2004).

Besides LP, mathematical optimization comprises also two other types of problems: *Convex* and *Non-linear* optimization. A function $f_i(x) : \mathbb{R}^n \to \mathbb{R}$ it is said to be convex if it satisfies:

$$f_i(\alpha x + \beta y) \leq \alpha f_i(x) + \beta f_i(y)$$

for all $x, y \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1, \alpha \geq 0, \beta \geq 0$. If the objective and restrictions functions of a mathematical program are convex, we call the problem a convex optimization problem. For example, LPs are a particular case of convex problems. As for LPs, there is no general analytic solution for these problems, but there are efficient algorithms for solving them, such as interior point methods. However, it is not possible yet to claim that solving convex problems is a mature technology, but it is expected to become one in a few years.

Nonlinear optimization is the term used to describe a mathematical optimization problem were the objective or constrains are not linear but is not known if they are convex. Nowadays, there are not effective methods for solving them. Solving approaches usually involve some compromises, such as focusing in *local optimization*, *i.e.*, finding an optimal solution among feasible points around it, but not guaranteeing that it is the lowest possible value. Local optimization methods can be very fast and handle large-scale problems. In the other hand, *global optimization*, searches for the true global solution, but they are computationally expensive. Calculating a solution even for small problems with tens of variables can take long running times (Boyd and Vandenberghe, 2004).

3

Capacitance in *Escherichia coli*

In most of CBM applications, an implicit assumption is that the metabolic network has a fixed structure. However, it is known that organisms transfer genes from one to another, increasing their capabilities (Jain et al., 1999; Treangen and Rocha, 2011). Besides, advances in genetic engineering enabled researchers to knock-out (delete) genes as well as express new genes in organisms.

Therefore, the question of how much impact will have a gene deletion or addition in a particular system fitness is important and can be treated using CBMs framework. In contrast with other approaches, the strength of using CBM techniques relies in that no expert knowledge is needed to select appropriate transformations.

Gene deletions can be modeled by adding constrains of the type $v_j = 0$. In particular, using CBM to suggest knock-out candidate genes was analyzed by Burgard et al. (2003). In this work, authors propose a bi-level problem, where the inner problem maximizes the cellular objective using fluxes and the outer problem maximizes the biotechnological objective (*e.g.* production of a metabolite) using binary variables, *i.e.*, variables $y_i \in \{0, 1\}$ where additional restrictions imply $v_i = 0 \Leftrightarrow y_i = 0$ and $v_i \neq 0 \Leftrightarrow y_i \neq 0$.

Add a new reaction r, provided by an inserted gene for example, translates to add a new column in the stoichiometric matrix with their corresponding new flux v_r. Using this approach, we can simulate *in-silico* possible experimental outcomes. The problem of adding new reactions to the stoichiometric network has been analyzed by Pharkya et al. (2004) who introduce the OptStrain procedure. OptStrain procedure is constituted by four steps:

i) Construction of a database containing known reactions.
ii) Calculation of the higher yield set of reactions to obtain a product from a given substrate.
iii) Identification of the pathway that minimizes the the number of non native reactions in the production host
iv) Introduction of non native reactions in the host and select gene deletions to assure production.

Other approaches had been proposed in the literature: Pharkya and Maranas (2006) investigate reaction activations/inhibitions and deletions in metabolite production, in framework called OptReg; Hädicke and Klamt (2010) proposed an approach using Elementary Modes fluxes, termed CASOP; Ranganathan et al. (2010) introduced OptForce procedure, which focuses in minimal and maximal flux values to achieve an over producing network and (Yang et al., 2011) presented EMILiO approach, which seeks to improve the minimal production rate under optimal growth conditions. In general terms, OptStrain, OptReg, OptForce and EMILiO rely on bi-level formulations to optimize a design objective (production) while keeping microorganism near their natural optima, *i.e.*, maximal biomass production. For a more detailed review of methods, reader is referred to Fong (2014).

Recently, Larhlimi et al. (2012a) have described the concept of "stoichiometric capacitance" (SC) in

metabolic networks. SC corresponds to reaction that is added to the network and increases the objective function value. This increment is optimal, in the sense that a SC is chosen in such a way that objective function reaches its theoretical maxima. Furthermore, it is possible to decompose SC into known enzymatic reactions by the use of a companion LP problem and therefore effectively propose genes to be inserted in the target organism. This could be interpreted as "Which is **the best** possible gene insertion?". Because it is possible to define "the best" in terms of an objective function, this question can be formalized as an optimization problem. Mathematically, SC is computed using Mixed Integer Linear Programming (MILP), a variant of LP problems: In MILPs, some of the variables are restricted to be integers, *i.e.*, it has variables $\{x_0, \ldots, x_n\} \in \mathbb{Z}$. While complexity of MILPs is known to be NP-Hard (Schrijver, 1998), solutions methods are implemented in several platforms and available in commercial and academic packages (Achterberg et al., 2005).

As the SC effectively increases the value of objective function, added reactions are likely to be used by the augmented metabolic network, making the capacitance an interesting tool to propose genetic transformation of an organism. Most of applications are motivated from a biotechnological perspective; usually, one is interested into increase the production of a certain metabolite while keeping growth rates over certain threshold.

In this chapter we explore this possibility by adding constrains to the original SC formulation to enforce ethanol production, using *E. coli* as case study. Theoretical yields were compared with experimental results from Wargacki et al. (2012). Capacitances for amino acid production were also studied. As result, three possible transformations were obtained and tested *in silico* for ethanol and two for amino acids.

Beyond SC computation, this work emphasizes CBMs flexibility to model metabolic networks. Form a biological perspective, SC explores the space of gene insertions that are not in the original network but increases the fitness (objective function) of an organism. Therefore, SC is suited to capture in a computational efficient algorithm ecological phenomena as lateral gene transfer.

The following article is currently in preparation for journal consideration.

Using Constraint-based Modeling to Design Synthetic Biology Experiments: Application to Biofuel Production

Marko Budinich, Abdelhalim Larhlimi, Jérémie Bourdon, and Damien Eveillard

Computational Biology Group, LINA UMR 6241 CNRS, EMN, Université de Nantes, Nantes, France

Corresponding author:
Marko Budinich

Email address: marko.budinich@univ-nantes.fr

ABSTRACT

The last decade saw the rise of synthetic biology studies that promote the use of molecular techniques to modify biological systems for industrial purposes. Protocols now exist either to develop modern selection tools for biological systems of interest, or to optimize genome evolution towards those that express valuable biomolecules. However, the majority of these synthetic studies are driven by molecular facilities and mostly made from empirical expertise. Here we show that, synthetic results could be replicated by computational approaches. Indeed, when one considers all available genomic and genome-scale metabolic knowledge, producing a bioproduct by a bacteria could be seen as a proper optimization problem. For the sake of application, this study advocates for the use of optimization methods to further understand previous synthetic studies but also for promoting putative genes that must be incorporated within *E. coli* metabolic network for experimental designs to product compounds of interest such as ethanol, glutamine or alanine.

INTRODUCTION

Biotechnology is today more challenged than ever. Humankind shows great expectations in the control of Biological Systems, for instance, to increase the food production, to promote new molecules with pharmaceutical or industrial interest (*e.g.* biofuel or bioplastic). Abundant studies recently advocated that molecular techniques could overcome these challenges. Protocols now exist to develop modern selection tools for biological systems of interest, and to optimize genome evolution towards those that express valuable biomolecules (Lee et al., 2012; Enquist-Newman et al., 2014). In particular, all these techniques contribute to the settlement of Synthetic Biology as a promising engineering field. However, the majority of these studies are rather made from empirical expertises and driven by extensive molecular facilities.

Here, we show that, along previous computational studies (Pharkya et al. (2004); Pharkya and Maranas (2006); Hädicke and Klamt (2010); Ranganathan et al. (2010); Yang et al. (2011); see Fong (2014) for a review of methods) considering available genomic knowledge, producing one bioproduct could be seen as a proper optimization problem, that, once solved, emphasizes mathematically optimal combinations of genes that must be targeted in priority in further molecular experiments.

Considering microbial systems in such an optimization paradigm allows (i) not only to better understand the underneath biological mechanisms of previous successful synthetic biology studies (Lee et al., 2012; Nielsen and Keasling, 2011) (ii) but also demonstrates that previously published synthetic study gene candidates are not always those that provide theoretical optimal productions of targeted bio-products.

Microorganisms are assumed to be evolutionary optimized for converting substrates to biomass, while their metabolic networks are subject to physico-chemical, thermodynamical and environmental constraints. Based on this assumption, *Flux Balance Analysis (FBA)* has been widely used to predict the phenotype of micro-organisms.

Using FBA, one can predict the growth rate as well as the behaviors of the cell in different environmental conditions (see Raman et al. (2005); Perumal et al. (2011); Knoop et al. (2013) for biological

illustration). As living organisms are known to be redundant and robust to genetic perturbations and environmental changes, a living system can often display several optimal metabolic behaviors which all guarantee an optimal biomass production.

To assess these different optimal metabolic behaviors, *Flux Variability Analysis (FVA)* has been proposed to compute the range of possible fluxes for each reaction in the metabolic network. The main outcome of FVA is to partition reactions into different classes depending on their flux ranges. Later on, Larhlimi et al. (2012) have proposed an in-silico approach to further improve the biomass yield by adding a chemical transformation, called *stoichiometric capacitance* (SC) to the metabolic network (see Fig. 1 for a capacitance method overview an illustrative application). Such a transformation represents a theoretical bypass within the metabolic network such as its addition optimizes an objective (*i.e.*, usually increasing the overall biomass), while satisfying thermodynamical and mass-balance constraints. The later can be seen as an overall biochemical transformation that can hopefully be expressed as the sum of a set of enzymatic reactions.

In this paper, we will use the above mentioned *in-silico* approaches to replicate an established experimental result of Wargacki et al. (2012) who built a synthetic strain by introducing DNA fragment from *Vibrio splendidus* to *E. coli* in order to simultaneously degrade, uptake and metabolize alginate for the sake of the bio-ethanol production. For this purpose, we will first use FBA and FVA to simulate such a synthetic construction which was mainly relying on genetic and molecular expertises. Afterwards, we will use the stoichiometric capacitance approach to not only increase biomass production, but also include further constraints based on expert knowledge to design dedicated synthetic strain models that improve biofuel production. FBA and FVA were then applied on synthetic model to investigate the addition of selected stoichiometric capacitances on putative synthetic models. Complementary, following a last optimization, we will propose putative genes that must be incorporated within *E. coli* genome to build the corresponding synthetic strains. For the sake of computational replication, this study provides as well a framework called MeDUSA that can be used by researchers to make further simulations.

1 MATERIALS AND METHODS

1.1 In-silico optimization-based approaches

We used in this study the state-of-the-art in-silico approaches that are based on mathematical optimization. Namely, following *Flux Balance Analysis (FBA)* Varma and Palsson (1994), the optimal flux distribution displayed by a micro-organism can be obtained by solving the following Linear Program (LP) :

$$\begin{aligned} & z^* = \text{maximum } v_{\text{biomass}} \\ & \text{subject to:} \\ & \quad S\,v = 0, \\ & \quad lb \leq v \leq ub, \end{aligned} \tag{FBA}$$

where $S \in \mathbb{R}^{m \times n}$ stands for the stoichiometric matrix, lb and ub are respectively the lower and upper bounds of flux capacity, v is the flux vector and v_{biomass} denotes the growth rate. From linear programming theory, it is known that the optimal value z^* of the objective function is unique, although there may exists many flux distributions (*i.e.*, values of $\mathbf{v}$) that achieve the same optimal value z^*. This is in agreement with the fact that micro-organisms have multiple metabolic pathways that all achieve the same performance (*i.e.*, biomass production).

To describe these multiple optimal pathways, we propose to use *Flux Variability Analysis (FVA)* (Bordbar et al., 2014) which calculates all the possible fluxes of reactions within the optimal metabolic pathways. Indeed, given a metabolic reaction i, the maximum (resp., minimum) possible flux through reaction i can be obtained by solving the following LP :

$$\begin{aligned} & \text{maximum/minimum } v_i \\ & \text{subject to} \\ & \quad S\,v = 0, \\ & \quad lb \leq v \leq ub, \\ & \quad v_{\text{biomass}} \geq \alpha \cdot z^*, \end{aligned} \tag{FVA}$$

where $\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$ represents the fraction of optimum with respect to the FBA objective value that is to be considered. Using FVA, we can classify reactions into three types. Indeed, given a reaction i :

- If reaction i is not involved in any optimal pathway, this reaction is called an *excluded reaction*. In this case, reaction i can not carry a non-zero flux in any optimal pathway. These reactions are pictured in red in Fig. 2 and Fig. 4.

- If reaction i is involved in all optimal pathways, this reaction is called an *indispensable reaction*. In this case, zero does not belong to its possible range of fluxes within the optimal pathways. These reactions are pictured in green in Fig. 2 and Fig. 4.

- If reaction i is not involved in any feasible solution, *i.e.*, in any steady-state pathway, this reaction is called a *blocked reaction*. In this case, reaction i can not carry a non-zero flux in any feasible solution. These reactions are pictured in black in Fig. 4.

- Otherwise, reaction i is called an *alternative reaction* and depicted as yellow in Fig. 2 and Fig. 4.

Given the same physico-chemical constraints defining all the possible steady-state flux distributions through a metabolic network, a significant improvement of biomass production can be obtained by adding to the network a chemical transformation, called *stoichiometric capacitance* (SC), whose stoichiometry can be represented by a sparse vector r than can be obtained by solving the following MILP (Larhlimi et al., 2012):

$$
\begin{array}{ll}
\text{maximum } v_{\text{biomass}} & \\
\text{subject to} & \\
\quad S\,v + r = 0, & \\
\quad lb \leq v \leq ub, & \\
\quad Mr = 0, & (i) \\
\quad Tr \leq 0, & (ii) \\
\quad -\lambda x \leq r \leq \lambda x, & (iii) \\
\quad \sum_{i=1}^{m} x_i \leq \mu, & (iv) \\
\quad v \in \mathbb{R}^n, \;\; r \in \mathbb{R}^m, \;\; x \in \{0,1\}^m. &
\end{array}
\tag{SC}
$$

Conditions (i) and (ii) ensure that the SC is more likely to occur in nature. Indeed, the added reaction r must be (i) mass-balanced, *i.e.*, r must lie in the kernel of the mass matrix M, and (ii) thermodynamically feasible, with T denoting the vector of the standard Gibbs free energy of the metabolites. Conditions (iii) and (iv) guarantee that r involves a limited number of metabolites. Further constraints can easily be included to consider biological hypotheses.

SC does not always correspond to a single reaction but rather a combination of several enzymatic steps (*n.b.* these reactions are not always successive since bypass can be created by are not always successive). The general problem of finding a decomposition of the capacitance r is already described in Larhlimi et al. (2012); here we used a variation of the method, called stoichiometric capacitance decomposition (SCD). In particular, we search a feasible solution satisfying the following constraints:

$$
\begin{array}{l}
D\alpha = r \\
\alpha_i \leq 0; i \in \text{Irr}, \\
-Nb_i \leq \alpha_i \leq Nb_i, \\
\sum b_i \leq k, \\
b_i \in \{0,1\}.
\end{array}
\tag{SCD}
$$

where r is the capacitance, D is the stoichiometric matrix of the reaction database, b is a binary vector associated to , Irr is the set of index corresponding to irreversible reactions, N a large number and k the fixed number of reactions allowed. When solved, non-zero coefficients of α and b indicates which reactions of D correspond to a decomposition of r.

1.2 Case of Study: Using *E. coli* as an ethanol factory

This study aims at mimicking the experimental microbial platform described by Wargacki et al. (2012) using above optimization criteria. Thus, we propose to tune the metabolic model iJR904 of *E. coli* (Reed et al., 2003) by finding capacitances that increases the microbial capabilities to excrete a relevant byproduct. In this case, besides bioethanol as target, we also looked for SC for industrially relevant amino-acids. In this purpose, in addition to constraints as considered in the MILP above, we added constraints to select interesting capacitances such as those that avoid oxygen production and CO_2 consumption (resp., NH_4 consumption), in order to increase bio-ethanol (resp., amino acids) productions.

After the stoichiometric capacitance calculation (see Equation SC), these were decomposed (see Equation SCD) in known enzymatic reactions using METACYC as reaction database (Caspi et al., 2014). For the ease of the calculation, we used a $k = 20$, to limit the number of genes to consider to produce the stoichiometric capacitance in the present work (see Table 2). SC and SCD calculations were implemented using the MeDUSA framework (see below).

1.3 MeDUSA: a SAGE implementation to identify Stoichoimetric Capacitance

The computation of stoichiometric capacitance (SC) is available within an open source python based tool called MeDUSA which, given a metabolic network and an objective function, (i) computes a SC and (ii) investigates consequences of including the calculated SC in the wild-type model.

For the sake of illustration applicative tutorial, MeDUSA was applied to explore the possibility of increasing the Glutamate (Glu) production in the metabolic network of amino acid synthesis (see Fig. 1 for illustration). This case study was introduced by Schuster et al. (1999) for the sake of metabolic modeling validation. The corresponding toy network consists of 16 metabolites and 24 reactions. In addition to the stoichiometric matrix (`S`) and the lower (`lb`) and upper (`ub`) bounds on the fluxes through reactions, the capacitance calculation requires the vector (`T`) of the standard Gibbs free energy of metabolite formation, the mass matrix (`M`) which contains the molecular sum formulas of the metabolites, the reversibility of reactions (`rev`) and the index (`obj`) stating the objective function to be optimized (ex. biomass production). Vectors indicating the names of reactions (`rxnnames`) and metabolites (`metnames`) and the indices (`excindices`) of metabolites that must not occur in the capacitance can be used as well. All data, including biological and thermodynamic knowledge, that are necessary to replicate the stoichoimetric capacitance computation are available in the MeDUSA package (available for download in https://logiciels.lina.univ-nantes.fr/redmine/projects/medusa/).

We first construct an object `my_model` using the loaded data. Then, we use `my_model` to build the corresponding MILP problem in order to perform the capacitance calculation. The resulting MILP problem can be solved using the state-of-the-art MILP solvers (ex. Cplex or Gurobi in this example) by calling:

```
sage: my_model = MetabolicModel.create_model(S,M,ub,lb,T,
                                metnames,rxnnames,obj,rev,excindices,'Gurobi')   (1)
```

Next, we compute a stoichiometric capacitance by fixing an upper bound on its flux (ex. 1000) and a maximum number of metabolites to be used (ex. 4).

```
sage: (sol,cap) = my_model.capacitance(1000,4)                                   (2)
```

Finally, we use FVA to investigate the changes in the importance of metabolic reactions for performing the network objective. To achieve this, we call the function `fva` while setting fluxes through the reactions `indices` to `values`.

```
sage: (min values,max values) = my model.fva(indices,values)                     (3)
```

Based on above calculations, we obtain the following stoichiometric capacitance

$$3\,Fum + 2\,Suc \longrightarrow 4OG \tag{4}$$

whose inclusion in the metabolic model results in an increase of the Glutamate (Glu) production by 60%. To further analyze consequences of adding such a stoichiometric capacitance to a given model, MeDUSA makes use of Flux Variability Analysis (FVA) with and without adding the calculated stoichiometric capacitance. Results can be then exported in different graph formats, allowing data exchange with standard graph tools like Graphviz (Gansner and North, 2000) and Gephi (Bastian et al., 2009). For more details, please refer to the tutorial available from https://logiciels.lina.univ-nantes.fr/redmine/projects/medusa/

2 RESULTS AND DISCUSSION

2.1 Biological meanings of the stoichiometric capacitance

Fig. 1 and Fig. 2 show the stoichoimetric capacitance that optimizes the glutamate production in *E. coli* (see Section MeDUSA: a SAGE implementation to identify Stoichoimetric Capacitance) and its consequences for overall metabolic behaviors. As already shown in Larhlimi et al. (2012), the capacitance transformation replicates the previous study of Schuster et al. (1999). Fig. 2A illustrates the state of each reaction of the metabolic network. FVA allows to decipher the set of reactions into four distinct types: blocked (unable to carry a non-zero flux in any condition); excluded (unable to carry a non-zero flux in all optimal metabolic pathways); indispensable (carrying a non-zero flux in all optimal metabolic pathways) and the remaining reactions are called alternative (resp. green, red and yellow in Fig. 2). By comparing the FVA results before and after adding the capacitance to the metabolic network. Fig. 2 emphasizes as a table the changes in the type of reactions due to the capacitance inclusion. Conversely, Fig. 2B summarizes similar results by a graph that emphasizes connections between reactions (*i.e.*, nodes as circles) when they share metabolites (*i.e.*, nodes not circled). Following Fig. 2A conventions, red, green and yellow nodes are respectively blocked, obligatory and alternative reactions that remains unchanged after the stoichoimetric capacitance inclusion. Reversely, blue nodes depict reactions that change their reaction states. In particular, the inclusion of the Stoichiometric Capacitance blocks Pck, ACeEF, GltA, lcd and Acn; whereas Sdh, Fum_r and Mdh becomes indispensable. These changes summarize an overall bypass of Oxoglutarate (OG) synthesis, which is the only substrate of Glu. Indeed, a bypass of Oxoglutarate (OG) synthesis monitors re-allocations of flux from Pyruvate (Pyr), Co-enzyme A (CoA) and Isocitrate (Isocit) synthesis leads to modify Succinate (Succ) and Fumarate (Fum) synthesis pathways from Oxalacetate (OAA).

2.2 Computational investigation of synthetic strains

Previous works advocate for the use of bacterial strains as a framework for industrial bio-product synthesis (see Nogales et al. (2008); Liang et al. (2011) for illustration). Among other synthetic studies dedicated to ethanol production, Wargacki et al. (2012) designed a synthetic model (called WG in the sequel) that modifies *E.coli* (WT) to produce ethanol via alginate degradation.The corresponding genetic construction adds alginate, oligoalginate lyase and transporters within *E. coli* metabolism while knocking out *p-flB-focA*, *frdABCD*, and *ldhA* genes, which forces the strain to follow an anaerobic behavior.

Both WG and WT were simulated using constraint-based approaches and show differential behaviors. Corresponding results are summarized in Fig. 3. First, FBA confirm experiments from Wargacki et al. (2012) because they emphasize a global increase of WG ethanol production over a wide range of O_2 conditions, while the WG biomass remains smaller than WT one (resp. blue and purple in Fig. 3). Second, FVA promote further investigations of WG biological features. In particular, FVA pinpoints 11 reactions, that could not be used in WT model (*i.e.*, blocked or alternative if their flux are respectively 0 or ≤ 0), become obligatory used (*i.e.*, flux $\neq 0$) in WG model after genetic modifications (see Fig. 4 for details). These new obligatory reactions mainly belong to glycolysis, pentose phosphate and alternate carbon metabolism. Conversely several reactions from oxidative phosphorylation, pyruvate and alternate carbon metabolisms switch from excluded to alternative, even in low O_2 concentrations.

2.3 *Escherichia coli* as a microbial factory

However, the WG synthetic strain was not built upon an optimization hypothesis. So, using the stoichoimetric capacitance framework, we applied our optimization protocol to maximize the ethanol production while maintaining an overall biomass production, and this over a wide range of oxygen concentrations. In practice, this objective consists in finding, among all capacitances that increase the biomass production, those that avoid oxygen production and CO_2 consumption (*i.e.*, constraints that modify the aerobic shunt), while either maximizing ethanol production (C1) or biomass (C2) or avoiding undesired side products like H_2O_2 (C3) (see Table 1 for details). These three capacitances are quantitatively satisfying (see Fig. 3 and Table 1) when their corresponding models are computationally analyzed by FBA. All C1, C2 and C3 models produce more ethanol and biomass than WG and WT models in similar O_2 conditions. Moreover, C1, C2 and C3 models show an aerobic shunt in higher O_2 concentrations than WT: respectively, $\sim$ 29.9 mmol.gDW.h^{-1}, $\sim$ 23.7 mmol.gDW.h^{-1} and $\sim$ 24.5 mmol.gDW.h^{-1} compared to 16 for WT and WG.

When focusing on fluxes within each model, FVA shows that C1, C2 and C3 capacitances exclude most of oxidative phosphorylation reactions, while maintaining threonine and lysine reactions (see Fig. 4

for details). Transport and extracellular reactions remain unchanged after adding capacitance, which pinpoints intrinsic properties of capacitance that do not explicitly modify boundary conditions as proposed by WG in Wargacki et al. (2012). Each capacitance model shows distinct biological features. Nevertheless, similar than WG, C1, C2 and C3 models are all forced to use carbon compounds as electron acceptors, reallocating extra-energy extracted using this mechanism for growth. As result, capacitance models are able to produce ethanol, i.e., translocation of the aerobic shunt at higher oxygen availability than WT. However, contrary to WG, C1, C2 and C3 models maintain a significant core metabolic activity as depicted by acetate, formate and lactate productions. Overall, C3 appears as the best compromise while C1 maximizes ethanol production and C2 biomass only.

As a companion optimization problem, we then deciphered sets of reactions and corresponding encoding genes, that are necessary to decompose C1, C2 and C3 stoichoimetric capacitances (see Equation SCD and Table 2). Stoichoimetric capacitances appear then encoded by a combination of *E. coli* genes with genes available in other strains that necessitate the design of further genetic constructions. Note herein that a similar application succeed to identify putative genes for optimizing L-glutamine and L-alanine amino acids production by selecting stoichoimetric capacitances that enforce the use of NH_4 as substrate and amino acids as products: 5 describes corresponding FBAs while Table. 1 and Table. 2 detail their corresponding efficiencies and putative genes that could be used for genetic constructions.

The role of Computer Sciences herein goes beyond traditional expectations; such as computing and storage capacities, by promoting abilities to formalize, to automatically extract knowledge and to infer new ones. From a general perspective, considering biotechnological challenges as direct results of an optimization paradigm paves the way for designing promising molecular protocols downstream of biological modelings, which not only formally reinforces connections between synthetic and metabolic engineerings (Nielsen and Keasling, 2011), but also promotes synthetic approaches to understand microbial strategies in evolutionary context (Papp et al., 2011). Optimal solutions will represent combination of genes that are assets in synthetic biology protocol, decreasing the time cost of deciphering efficient genetic candidates from the library of Life.

ACKNOWLEDGMENTS

The authors deeply thank Pr. Alejandro Maass for seeding our collaboration. This study is funded by *Région Pays de la Loire* GRIOTE project. MB was funded by INRIA-Chile.

REFERENCES

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *ICWSM*.

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

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(Database issue):D459–71.

Enquist-Newman, M., Faust, A. M. E., Bravo, D. D., Santos, C. N. S., Raisner, R. M., Hanel, A., Sarvabhowman, P., Le, C., Regitsky, D. D., Cooper, S. R., Peereboom, L., Clark, A., Martinez, Y., Goldsmith, J., Cho, M. Y., Donohoue, P. D., Luo, L., Lamberson, B., Tamrakar, P., Kim, E. J., Villari, J. L., Gill, A., Tripathi, S. A., Karamchedu, P., Paredes, C. J., Rajgarhia, V., Kotlar, H. K., Bailey, R. B., Miller, D. J., Ohler, N. L., Swimmer, C., and Yoshikuni, Y. (2014). Efficient ethanol production from brown macroalgae sugars by a synthetic yeast platform. *Nature*, 505(7482):239–243.

Fong, S. S. (2014). Computational approaches to metabolic engineering utilizing systems biology and synthetic biology. *Computational and Structural Biotechnology*, 11(18):28–34.

Gansner, E. R. and North, S. C. (2000). An open graph visualization system and its applications to software engineering. *Software Practice and Experience*.

Hädicke, O. and Klamt, S. (2010). CASOP: A Computational Approach for Strain Optimization aiming at high Productivity. *Journal of Biotechnology*, 147(2):88–101.

Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., and Steuer, R. (2013). Flux balance analysis of cyanobacterial metabolism: the metabolic network of Synechocystis sp. PCC 6803. *PLoS computational biology*, 9(6):e1003081.

Larhlimi, A., Basler, G., Grimbs, S., Selbig, J., and Nikoloski, Z. (2012). Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics (Oxford, England)*, 28(18):i502–i508.

Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., and Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature Chemical Biology*, 8(6):536–546.

Liang, J., Luo, Y., and Zhao, H. (2011). Synthetic biology: putting synthesis into biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(1):7–20.

Nielsen, J. and Keasling, J. D. (2011). Synergies between synthetic biology and metabolic engineering. *Nature Publishing Group*, 29(8):693–695.

Nogales, J., Palsson, B. O., and Thiele, I. (2008). A genome-scale metabolic reconstruction of Pseudomonas putida KT2440: iJN746 as a cell factory. *BMC Systems Biology*, 2(1):79–20.

Papp, B., Notebaart, R. A., and Pál, C. (2011). Systems-biology approaches for predicting genomic evolution. *Nature Reviews Genetics*, 12(9):591–602.

Perumal, D., Samal, A., Sakharkar, K. R., and Sakharkar, M. K. (2011). Targeting multiple targets in Pseudomonas aeruginosaPAO1 using flux balance analysis of a reconstructed genome-scale metabolic network. *Journal of Drug Targeting*, 19(1):1–13.

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Research*, 14(11):2367–2376.

Pharkya, P. and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic engineering*, 8(1):1–13.

Raman, K., Rajagopalan, P., and Chandra, N. (2005). Flux balance analysis of mycolic acid pathway: targets for anti-tubercular drugs. *PLoS computational biology*, 1(5):e46.

Ranganathan, S., Suthers, P. F., and Maranas, C. D. (2010). OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions. *PLoS computational biology*, 6(4):e1000744–11.

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

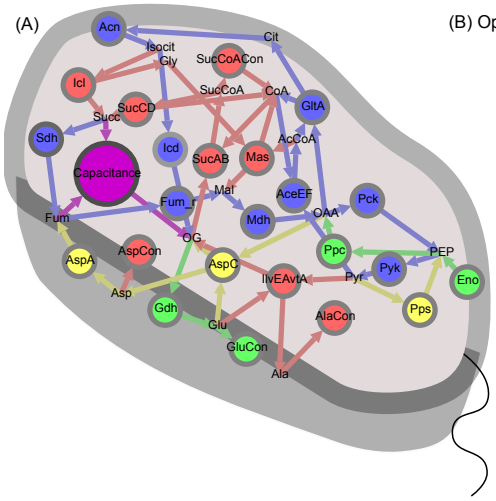
Schuster, S. S., Dandekar, T. T., and Fell, D. A. D. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in biotechnology*, 17(2):53–60.

Varma, A. and Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/technology*.

Wargacki, A. J., Leonard, E., Win, M. N., Regitsky, D. D., Santos, C. N. S., Kim, P. B., Cooper, S. R., Raisner, R. M., Herman, A., Sivitz, A. B., Lakshmanaswamy, A., Kashiyama, Y., Baker, D., and Yoshikuni, Y. (2012). An Engineered Microbial Platform for Direct Biofuel Production from Brown Macroalgae. *Science (New York, NY)*, 335(6066):308–313.

Yang, L., Cluett, W. R., and Mahadevan, R. (2011). EMILiO A fast algorithm for genome-scale strain design. *Metabolic engineering*, 13(3):272–281.

FIGURES



(B) Optimization problem

$$
\begin{aligned}
&\max \quad c^T v && c, v \in \mathbb{R}^n \\
&\text{s.t.} \quad Sv + r = 0 && S \in \mathbb{N}^{m \times n} && (1)\\
&l_b \le v \le u_b && l_b, u_b \in \mathbb{R}^n && (2)\\
&Mr = 0 && M \in \mathbb{R}^{k \times m} && (3)\\
&Tr \le 0 && T \in \mathbb{R}^m && (4)\\
&-\lambda x \le r \le \lambda x && \lambda \in \mathbb{R} && (5)\\
&\sum_{i=1}^{m} x_i \le \mu && x_i = \begin{cases} 0 & \text{if} \quad r_i = 0; \\ 1 & \text{otherwise.} \end{cases} && (6)
\end{aligned}
$$

Figure 1. Finding a stoichoimetric capacitance that increases the Glutamate (Glu) production in the amino acid synthesis in *E.coli* (A) represents the network that consists of 16 metabolites and 24 reactions. Metabolic reactions are depicted by color nodes whereas metabolites are shown by their names only. Edges depict connections between metabolites and reactions. The capacitance reaction is depicted as a magenta node that transforms Fumarate and Succinate into OG. Note herein that such synthetic reaction replicates results as obtained by Schuster et al. (1999). Following the addition of the capacitance, qualitative statuses of the reactions are then evaluated and summarized by different colors: blue nodes represent reactions that change their qualitative status when the capacitance is added, whereas red nodes are reactions that remain blocked, green nodes remain obligatory, yellow reactions remain alternative. (B) represents the formal definition of the capacitance maximization problem (MILP) while maintaining flux balance constrained (1). In addition to the stoichiometric matrix (S) and the lower (lb) and upper (ub) bounds on the fluxes through reactions (2), the capacitance calculation requires the overall mass conservation via the mass matrix (M) that contains the molecular sum formulas of the metabolites (3), while satisfying the energetic constrained of metabolic reactions via the vector (T) of the standard Gibbs free energy of metabolite formation (4). Complementary constraints limit the number of metabolites used by the capacitance (5) as well as the necessity of using metabolites that already belong to the metabolic network (6).

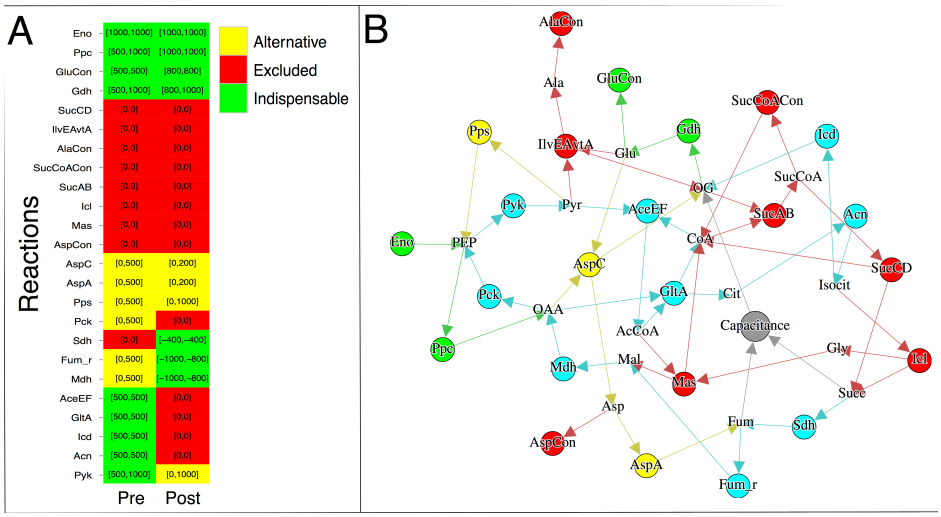


Figure 2. Visualization of the results obtained from the analysis of the metabolic model of amino acid synthesis in *E. coli* from (Schuster et al.(1999)). (A) Changes in the types of reactions (blocked,, obligatory and alternative) before (Pre) and after (Pos) the inclusion of the calculated stoichiometric capacitance (SC). The FVA flux ranges before and after adding the SC are indicated as well. (B) The metabolic network including the calculated SC. Metabolites are shown by their names and reactions are depicted as colored nodes: gray for the SC, cyan for reactions whose type changes due to the inclusion of the SC, and yellow (resp. red or green) for reactions which are alternative (resp. blocked or obligatory) before and after adding the SC.

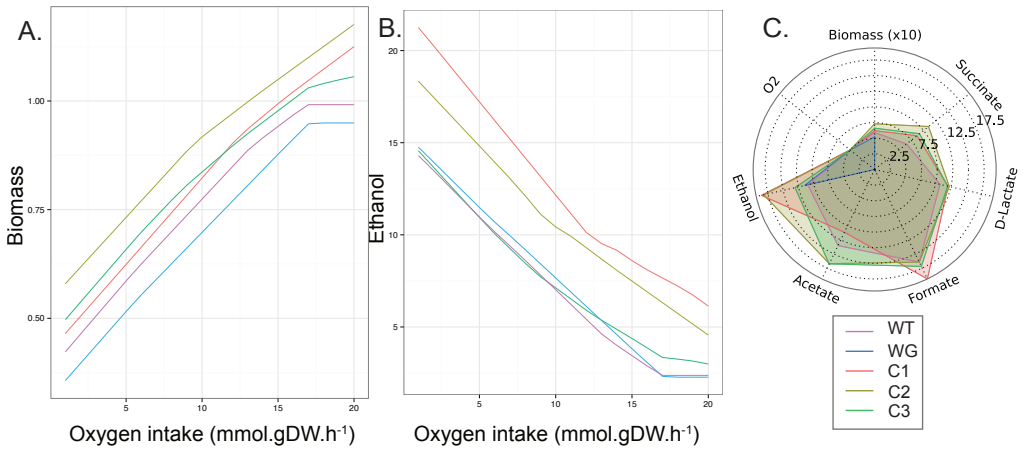


Figure 3. Comparative Flux Balance Analysis between different metabolic models. WT represents the original metabolic network; WG the metabolic model after genetic construction of WT as proposed by [6]; C1 the model after adding the capacitance transformation C1 within WT; C2 the model after adding the capacitance transformation C2 within WT; C3 the model after adding the capacitance transformation C3 within WT. Details of capacitance transformation are depicted in (Table 1). (A) represents the biomass estimation of each model over a range of O_2 conditions (mmol.gr^{-1}.DW.hr^{-1}). (B) depicts the ethanol estimation production (mmol/gr DW hr) of each model over a range of O_2 conditions (mmol/gr DW hr). In order to compare all model, (C) synthesizes biomass production, ethanol production as well as the main fermentation side product concentrations (mmol.gr^{-1}.DW.hr^{-1}) for a given oxygen concentration that maximizes ethanol production.

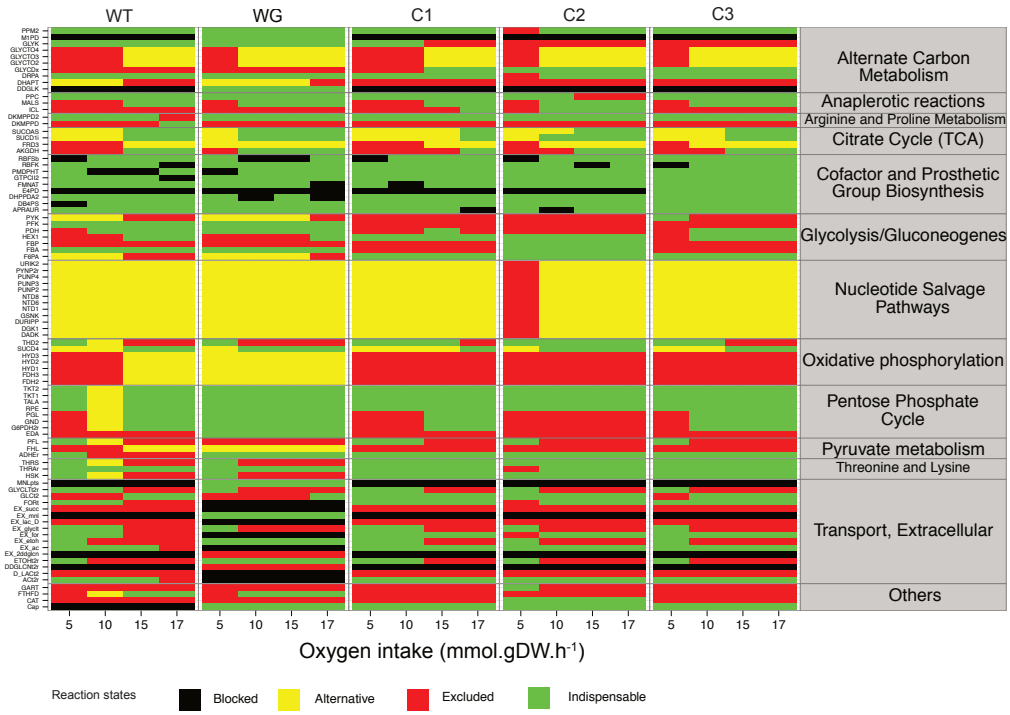


Figure 4. Flux Variability Analyses between different metabolic models. WT represents the original metabolic network; WG the metabolic model after genetic construction of WT as proposed by [6]; C1 the model after adding the capacitance transformation C1 within WT; C2 the model after adding the capacitance transformation C2 within WT; C3 the model after adding the capacitance transformation C3 within WT. Details of capacitance transformation are depicted in Table 1. Flux Variability Analysis of all models depicts the state of each metabolic reaction: red when they appear as excluded (i.e. their flux = 0), green when obligatory (*i.e.*, their flux>0) or yellow (i.e. their flux ≤ 0, black when blocked which implies that the reaction could not be used. Reactions are ranked based on their main metabolic function assignations.

TABLES

Table 1. Detailed description of stoichiometric capacitances for several biological objectives. Comparisons for ethanol as target were calculated under anoxic condition, i.e., $V_{O_2} \leq 0.5$ mmol.g^{-1}.DW.h^{-1}. For each stoichiometric capacitances, column BM represents the % of Biomass improvement, whereas colum Export depicts the % of improvement of Products maximal theoretical target export. Additional constraints in ethanol stoichiometric capacitances were imposed: No oxygen production (to keep the anoxic environment) and no CO_2 net consumption; in order to avoid carbon fixation mechanisms that are difficult to implement in practice.

	Target	**Capacitance**	**Reaction**	**BM**	**Export**
C1	ethanol	Dihydroxyacetone + Proton → Ethanol + Water	2 $C_3H_6O_3$ + 12 H^+ → 3 C_2H_5OH + 3 H_2O	6.5%	69.25%
C2	ethanol	Dihydroxyacetone + Proton → Ethanol + Hydrogen peroxide	4 $C_3H_6O_3$ + 18 H^+ → 6 C_2H_5OH + 3 H_2O_2	24.9%	65.9%
C3	ethanol	Dihydroxyacetone + D-glyceraldehyde-3-phosphate → Ethanol + Erythronate-4-phosphate	3 $C_3H_4O_{10}P_2$ + 7 $C_3H_6O_3$ → 3 C_2H_5OH + 6 $C_4H_6O_8P$	12.4%	19.6%
C4	L-glutamine	Ammonia + Pyruvate → D-galactarate + L-glutamine	6 NH_4 + 16 $C_6H_8O_8$ → 3 $C_6H_8O_8$ + 6 $C_5H_8NO_4$	3.6%	8.5%
C5	L-alanine	Ammonia + Pyruvate → L-alanine + Malate	3 NH_4 + 7 $C_3H_3O_3$ → 3 $C_3H_7NO_2$ + 3 $C_4H_4O_5$	6.2%	-0.14%

Table 2. METACYC reactions for each capacitance

	List of METACYC reactions	Associated EC numbers
C1	1.2.99.3-RXN, 1.2.99.7-RXN, 3.4.22.32-RXN, 3.4.24.62-RXN, 3.4.24.76-RXN, 3.4.25.1-RXN, 3.6.5.4-RXN, FUMARATE-REDUCTASE-NADH-RXN, GLYCDEH-RXN, NADH-DEHYDROG-A-RXN, RXN-11334, RXN-13191, RXN-13852, RXN0-6373, SORBOSE-5-DEHYDROGENASE-NADP-RXN, SORBOSE-DEHYDROGENASE-RXN, SUCC-FUM-OXRED-RXN, TRANS-RXN-131, TRANS-RXN0-474, TRANS-RXN0-546	EC-1.3.1.6, EC-3.4.24.62, EC-3.4.25.1, EC-3.4.22.32, EC-1.2.99.3, EC-1.2.99.7, EC-1.1.1.6, EC-3.4.24.76, EC-1.17.1, EC-1.1.1.123, EC-1.1.99.12, EC-1.1.99.35, EC-1.14.13.81, EC-1.6.5.3, EC-1.1.5.2
C2	1.7.2.3-RXN, 2.7.11.11-RXN, 3.1.3.16-RXN, 3.4.21.79-RXN, ALCOHOL-DEHYDROG-RXN, ASPARTATE-4-DECARBOXYLASE-RXN, ATPSYN-RXN, GLUTAREDOXIN-RXN, GLYCEROL-DEHYDROGENASE-ACCEPTOR-RXN, GLYCEROL-KIN-RXN, NADPH-PEROXIDASE-RXN, PEPDEPHOS-RXN, RXN-8667, RXN-9615, RXN0-5052, SULFITE-REDUCT-RXN, SULFITE-REDUCTASE-RXN, TRANS-RXN-131, TRANS-RXN0-546, TRANS-RXN0-547	EC-3.4.11, EC-4.1.1.12, EC-2.7.1.40, EC-3.6.3.14, EC-1.11.1.21, EC-1.11.1.16, EC-3.4.21.79, EC-1.1.99.22, EC-2.7.1.30, EC-1.11.1.2, EC-1.8.1.2, EC-1.8.99.1, EC-1.1.1.1, EC-1.7.2.3, EC-3.1.3.16
C3	1.14.19.2-RXN, 1TRANSKETO-RXN, 2TRANSKETO-RXN, 3.4.16.5-RXN, ATPSYN-RXN, ERYTH4PDEHYDROG-RXN, GAPOXNPHOSPHN-RXN, GLYCEROL-2-DEHYDROGENASE-NADP-RXN, GLYCERONE-KINASE-RXN, RIB5PISOM-RXN, RIBULP3EPIM-RXN, RXN-13567, RXN-2785, RXN0-313, TRANS-RXN-131, TRANS-RXN0-277, TRANS-RXN0-546, TRANS-RXN0-547, TRANSALDOL-RXN, TRIOSEPISOMERIZATION-RXN	EC-2.2.1.1, EC-3.6.3.14, EC-4.1.2, EC-3.4.16.5, EC-5.3.1.6, EC-5.1.3.1, EC-1.2.1.72, EC-2.7.1.29, EC-1.2.1.12, EC-1.1.1.156, EC-2.2.1.1, EC-1.6.1.2, EC-2.2.1.2, EC-5.3.1.1, EC-1.14.19.2
C4	URONATE-DEHYDROGENASE-RXN, RXN-2301, RXN-8653, PYRUVATE-OXIDASE-COA-ACETYLATING-RXN, GLUTAMINASE-ASPARAGINASE-RXN, 1.3.3.12-RXN, ACETYL-COA-HYDROLASE-RXN, RXN-8092, RXN-11152, NADH-PEROXIDASE-RXN, RXN-11383, ALCOHOL-DEHYDROG-RXN, PYRUVDEH-RXN, NADH-DEHYDROG-RXN, RXN0-5268, RXN-8220, TRANS-RXN-208, TRANS-RXN-234, TRANS-RXN0-546, TRANS-RXN0-545	EC-1.1.1.203, EC-2.6.1.55, EC-1.2.3.6, EC-1.3.3.12, EC-3.1.2.1, EC-1.2.3.1, EC-1.11.1.1, EC-1.1.1.1, EC-1.2.1, EC-1.6.5.3, EC-1.10.3.10, EC-3.1.1.80
C5	RXN-2802, ASPARTASE-RXN, ASPARTATE-4-DECARBOXYLASE-RXN, FUMHYDR-RXN, PYRNUTRANSHYDROGEN-RXN, RXN-12878, RXN-12079, RXN-12081, RXN-12082, MALSYN-RXN, SERINE-DEHYDROGENASE-RXN, SUCC-FUM-OXRED-RXN, SULFITE-REDUCT-RXN,ALANINE-DEHYDROGENASE-RXN, PYRUVDEH-RXN, NADH-DEHYDROG-RXN, SUCCINATE-DEHYDROGENASE-UBIQUINONE-RXN, RXN-974, ALANINE-AMINOTRANSFERASE-RXN, SULFITE-REDUCTASE-FERREDOXIN-RXN	EC-4.3.1.1, EC-4.1.1.12, EC-4.2.1.2, EC-1.6.1.2, EC-1.6.1.3, EC-1.6.1.1, EC-1.1.1, EC-2.3.3.9, EC-1.4.1.7, EC-1.8.1.2, EC-1.4.1.1, EC-1.2.1, EC-1.6.5.3, EC-1.3.5.1, EC-2.6.1.2, EC-1.8.7.1

SUPPLEMENTAL DATA

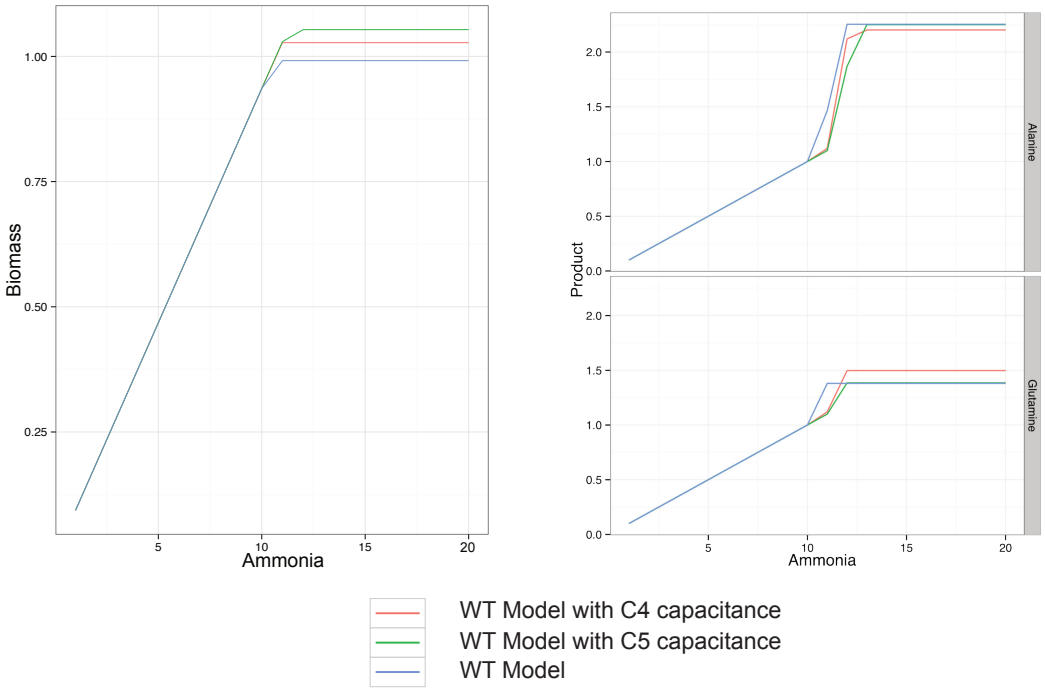


Figure 5. S1 Fig. **Comparative Flux Balance Analysis between different metabolic models.** WT represents the original metabolic network; C4 the model after adding the capacitance transformation C4 within WT, and C5 the model after adding the capacitance transformation C5 within WT. Details of capacitance transformation are depicted in (Table 1). (A) represents the biomass estimation of each model over a range of NH_4 conditions ($mmol.gr^{-1}.DW.hr^{-1}$). (B) depicts respectively the alanine and glutamine estimation production ($mmol.gr^{-1}.DW.hr^{-1}$) of each model over a range of NH_4 conditions ($mmol.gr^{-1}.DW.hr^{-1}$)

.

4

Constraint Based Modeling for testing Evolution Theories

Constraint Based Methods provide a framework able to produce predictions about biological systems, given its quantitative nature. Notion of objective functions that are maximized is strongly related with the concept of *fitness* in ecology. Indeed, by means of natural selection, traits that provide better fitness, *i.e.* are better adapted to a particular niche, are preserved in the population; maximization of growth rate is then an appropriate model to study metabolic network behavior. Therefore, it is possible to use CBMs to model evolutionary processes and test theories. In this context, as part of a larger study, FBA and FVA were used to both select interesting conditions and make predictions in a model organism study. The work presented here was developed as a collaboration with Alix Mas and Philippe Vandenkoornhuyse , from RBPE team, ECOBIO laboratory at Rennes.

4.1 Sibling Queens Theories

4.1.1 Red Queen Hypothesis: Evolution by function gain

One of strengths of evolutionary theory is its broad application in all branches of biology. In 1973, Leigh V an Valen (Van Valen, 1973), while analyzing extinction rates in several taxon, noted that for groups in a given taxon the probability of extinction is independent of the age of the group, which implies that both extinction and generation rates of groups are constant. However, probabilities varied among taxons and geological times, while correlated well with adaptive zones. This led Van Valen to propose the Red Queen Hypothesis (RQH): If some species within the group gained an advantage to increase its fitness, other species will be impacted negatively. Then, new species may replace those that were severely affected. Another way to visualize the situation is to imagine that resources (in a broad sense) are fixed for an adaptive zone. If one species gain more control of the resources, others using the same resources will forcedly decrease their use. To avoid extinction, some of the negatively impacted species will have to regain control of the resources again, reaching a new equilibrium. Later, this idea was extended to molecular evolution (Van Valen, 1974): Proteins evolve to confer gains in an organism fitness; negatively impacted organisms adapt their proteins by natural selection to counteract this adverse scenario. In this sense, RQH can explain why organisms gain or improve functions and how they evolve by antagonism.

4.1.2 Black Queen Hypothesis: Evolution by function loss

Recently, RQH was echoed by a novel proposition called "Black Queen Hypothesis" or BQH, (Morris et al., 2012; Mas et al., 2016). In BQH, focus is put in how gene loss can drive adaptations of free-living organisms. Its main argument is that in the context of communities, several biological functions "leak" in the external environment, producing common goods that are available for all members of the community. In this context, organisms can show fitness gains by not producing these functions (*i.e.* loosing the associated genes), as they are provided by the media. This loss of functions should then be reflected at the genomic level.

Thanks to advances in sequencing technologies, it could be possible to check the BQH by analyzing how genes change at their sequence level in a time course experiment. Unfortunately, testing the BQH at this level is complex because it would involve separating several species before sequencing. However, if a common good is directly supplied in the culture media, it is possible to analyze a monoculture experiment that mimics the BQH.

From this point of view, it is possible to use CBMs to model experiments in a BQH context. By controlling nutrient exchange, reactions that are no longer needed to maximize the fitness can be detected. More precisely, by simulating culture conditions as nutrient availability, reactions forced to carry zero flux are predicted, providing a set of candidate genes to be affected by the BQH. Indeed, if a reaction is superfluous, it is expected to the genes involved in this function should non functional and thus could accumulate mutations without it being deleterious for the organisms (rise of selection), leading to gene decay.

4.2 Testing BQH in *Pseudomonas fluorescens*

Pseudomonas fluorescens Pf0-1 (Silby et al., 2009) was selected as model organism to test BQH theory. *P. fluorescens* is a a common Gram-negative, rod-shaped bacterium considered as a generalist species that can use several sources of carbon (such as glucose, xylose, fructose, succinate, etc.) for its growth.

P. fluorescens Pf0-1 model was obtained from Seed Database (http://theseed.org/). This model consist in 1430 reactions and 1235 metabolites, separated in 2 compartments (external and cytosol, respectively). One hundred and twenty seven of them are exchange reactions, *i.e.*, they perform the intake/export of metabolites from environment.

4.2.1 Using FBA to define an *in silico* medium

A series of FBA were used to check model response to different carbon sources, in order to determine which culture conditions could enforce observation of BQH in *P. fluorescens* Pf0-1. Analysis were run using COBRA Toolbox under MATLAB environment (Schellenberger et al., 2011).

In a first stage, 11 essential metabolites (*i.e.*, in any scenario, they are required to produce biomass) were identified, corresponding mainly to trace nutrients: Mg, Cl^-, O_2, Cu^{2+}, Co^{2+}, $SO_4{}^-$, Ca^{2+}, K, Zn^{2+}, Mn^{2+} and spermidine. Next, it was noted that adding Fe^{2+}, Fe^{3+}, $PO_4{}^-$, $NH_3{}^-$ and vitamin B12, the model was able to produce biomass using succinate as carbon source. With this, an *in silico* medium was defined. Following simulations were carried keeping constant all input metabolites except for the carbon source.

The second stage was to test which carbon sources were able to sustain growth. Model was tested using sucrose, lactate, arabinose, D-fructose, D-glucose, D-xylose, succinate and fumarate as carbon sources. From them, model was unable to produce any biomass using the first three, which were in consequence discarded for further simulations.

Next, for each carbon source, growth rate vs carbon source uptake were calculated using a series of FBA (Figure 4.1). Note that fructose and glucose have the same behavior, as well as succinate and fumarate in the range showed. Figure 4.1 shows that better culture yields will be achieved using glucose or fructose, and they will probably be preferred as carbon sources.

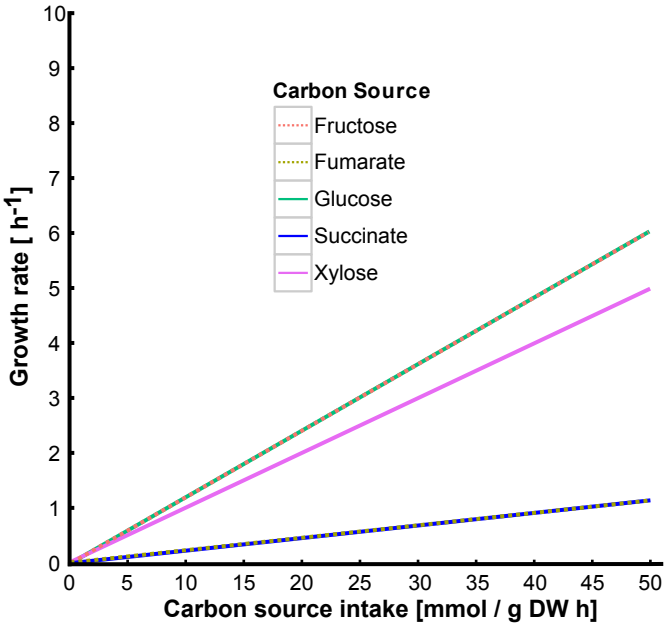


Figure 4.1 – Growth rates of *P. fluorescens* in different carbon sources

A Flux Variability Analysis (FVA) was performed, for each carbon source. Then, each reaction was classified into three Status: "Excluded" (if its flux is equal to 0), "Obligatory" if it carries flux (*i.e*, always carry a flux not equal to 0 in maximal biomass scenario) or "Alternative" (if it can carry flux in maximal biomass scenario). Count of reactions for each status is given in Table 4.1. Recalls that under a FVA, biomass functions is constrained to be close to their theoretical maxima, simulating an organism maximizing its fitness. Under that hypothesis, reactions not carrying flux are not used to increase fitness and therefore, they are candidates to be removed from the genome.

Table 4.1 – Alternative, blocked and indispensable reactions by carbon sources

Status	Fructose	Glucose	Xylose	Succinate	Fumarate
Alternative	436	436	434	433	433
Excluded	703	702	699	703	703
Obligatory	291	292	297	294	294

From these results, glucose was determined to be used as carbon source in the experimental set-up.

4.3 Experimental Set-Up and Sequence Analysis

A culture of *P. fluorescens* Pf0-1 was grown in a chemostat with glucose (1.3 g/L) as the sole source of carbon for a period of 4 weeks. Population was sampled at the beginning of the experiment (0 hours) and after 24 hours, 111 hours, 303 hours and 468 hours of culture.

Culture samples were used to extract DNA for whole genome sequencing. DNA was extracted and sequenced on Illumina HiSeq2000 platform. Next, single base mutations in *P. fluorescens* genome, called single nucleotide polymorphisms (SNP) were detected using DiscoSNP (http://colibread.inria.fr/software/discosnp/). A total of 23172 SNP were detected. Finally, these SNP were mapped into genes using NCBI *P. fluorescens* Pf0-1 genome sequence as reference (accession number NC_007492). Of the total 5678 of genes contained in the genome, 391 genes were affected by at least three SNPs.

4.4 Linking Fluxes and Genes

P. fluorecens Pf0-1 model contains associations between genes and reactions, useful to explore genome-phenotype relations. From 5678 genes present in the genome, 1130 are directly involved in the metabolic model. It is important to bear in mind that in some cases, one gene is associated to multiple reactions, and, in others, many genes are necessary for one reaction takes place. According to the BQH, in this experimental context it is expected that functions involved in unused pathways should be prone to decay, as they are useless (e.g., pathways of other sugars than glucose).

Using FVA, genes linked to metabolic reactions were classified reactions in four types: obligatory, alternative, excluded and blocked. Excluded reactions were distinguished from blocked reactions by running an FVA in a rich *in-silico* medium (*i.e.* a medium without nutrient restriction); reactions which carried zero flux in this condition were deemed as blocked. A second FVA was run setting the minimal *in silico* media with glucose and blocked reactions were subtracted from excluded reactions. With this distinction, excluded reactions correspond to reactions only blocked when glucose is used as carbon source.

Genes associated to the metabolic model were classified as obligatory, alternative, excluded or accordingly to the following rules:

— A gene associated with at least one obligatory reaction is classified as obligatory
— If a gene is not obligatory, is classified as alternative if is associated with at least one alternative reaction
— If a gene is not obligatory nor alternative, if is associated with one excluded reaction is classified as excluded
— Else, a gene belonging to a blocked reaction is classified as blocked

Of 1130 metabolic genes, 76 genes showed three or more SNP mutations (Table 4.2). A χ^2 goodness of fitness test shows (p-value = 0.03) that genes with more than three SNPs mutations follow a different proportion pattern genes in obligatory, alternative, excluded and blocked categories than those with less than three mutations.

Table 4.2 – Metabolic genes

Classification	SNP < 3	% of genes	SNP > 3	% of genes	Total
Obligatory	351	33.30%	20	26.32%	371
Alternative	419	39.75%	42	55.26%	461
Excluded	80	7.59%	2	2.63%	82
Blocked	204	19.35%	12	15.79%	216
Total	1054	100%	76	100%	1130

4.5 Conclusion and Perspectives

In this chapter it was shown how CBMs can be used to analyze problems derived from evolutionary theory. By using FBA and FVA over *P. fluorescens* CBM, simulations were able to link genes to environmental pressures.

P. fluorescens Pf0-1 model have some caveats. Only a ∼20% of genes are involved directly in the metabolic network; furthermore, blocked reactions varied from ∼50% to ∼30% of total reactions depending of the simulation conditions. This point to a deficiency in model reconstruction process, which has been already pointed as a general problem of genome scale metabolic problems (Monk et al., 2014). Nevertheless, metabolites detected as essential correspond well with Pseudomonas Minimal Medium (Kirner et al., 1996), so model correspond well with experimental conditions.

Linking metabolic genes with metabolites allowed to classify genes accordingly to the metabolic state of the reactions which involve them. As expected, genes which accumulated more than 3 SNPs mutations shown a different proportion pattern than those with less than three mutations, pointing to different selection pressures acting on both set of genes.

Interestingly, a higher than expected number of genes with more than three mutations than expected appeared in the alternative category and a lower number in the rest. Under the hypothesis that only beneficial mutated genes will be kept, results could be understood as obligatory genes maintaining their function and only fixing beneficial mutations as their function is critical. This can be interpreted in the context of the RQH, as populations are probably competing for carbon sources. As around 75% of the genes are not obligatory, they are susceptible to be affected by BQH phenomena.

Of particular interest are the two excluded genes with more than 3 SNP mutations, corresponding to a methionine transport and 2,5-dioxovalerate dehydrogenase. This last enzyme has been shown to be expressed in *Pesudomonas spp.* when exposed to different carbon sources (Adams and Rosso, 1967).

To deepen these results, two other networks should be mapped to the genes. First, no regulating genes have been linked to the metabolic activites. Secondly, as products of some reactions are substrates of others, reactions are said to be coupled (Larhlimi et al., 2012b), meaning that changes in one reaction will affect others. In practice, this mean that changes n one reaction could have an impact “downstream” in the metabolic network. Mapping this coupling network could give insights about indirect effects within metabolic network.

5

A bi-level formulation for linking Evolution and Constraint Based Modeling

In the previous chapter, it was shown how CBMs can be used to simulate a serie of possible carbon sources to select those conditions which could enable the observation of the BQH hypothesis. To this end, simulations conditions were set-up to emulate different experimental possibilities and all these possibilities were explored before to make an appropriate selection based in the internal metabolic status. This was possible in part because they were only a limited number of scenarios to explore. Is easy to see that the number of combinations increases exponentially with the number of variables to select a scenario, making this exhaustive approach inefficient to explore a high number of possibilities.

However, it is possible to reverse the biological question, by asking **which are the environmental conditions needed to produce a given metabolic status in the cell?** Motivated in part to provide conditions for BQH observation in chemostats, a method was sought to search conditions that maximized the number of excluded reactions. A two stage procedure was concived, configuring a bi-level problem.

First, it was noted that environmental conditions are given by the values of the set exchange fluxes, noted by $\mathcal{L}$. Next, with certain abuse of notation, we can formalize an environmental condition by $\mathbf{v}_L = \mathbf{E}$ where $L \in \mathcal{L}$. Given an environmental condition $\mathbf{E}$, the flux distribution which maximizes the number of active fluxes is calculated. To this, a binary variable f_i for each flux v_i is introduced. This variable is equal to one if and only if the corresponding flux is active ,*i.e.* $f_i \in \{0, 1\}$ with $f_i = 1 \Leftrightarrow v_i \neq 0$ and $f_i = 0 \Leftrightarrow v_i = 0$. Maximizing the sum of such variables gives the flux distribution where the only reactions with flux equal to zero are the excluded ones. This problem, henceforth denominated P1, constitutes the inner level of the optimization.

Next, in the upper level, the sum of active reactions is minimized, subject to the values of the external conditions, *i.e.*, $\mathbf{v}_L$. With this, an $\mathbf{E}$ environmental condition is found, which minimizes the set of maximal active reactions.

This problem is a Mixed Integer Bi Level Problem. Unfortunately, there is no general procedure to solve these type of problems, so any attempt of solving it would require a dedicated implementation of a solution procedure. For instance, progress has been made in the field of Answer Set Programming, a declarative problem solving approach, which have efficient algorithms to solve combinatorial problems (Gebser et al., 2012). However, implementing such procedure is out of the scope of the present thesis.

Nevertheless, formulation itself brings some interesting notes. First, problem P1 has been linked to the problem of finding the maximal cardinality Elementary Mode, *i.e.* the Elementary Mode including most reactions. Therefore, we can interpret the problem as finding conditions $\mathbf{E}$ to reduce the longest EM. Secondly, it is shown that bi-level formulations can capture well the nature of problems associated to two agents

making decisions in a hierarchically operation, which can offer interesting biotechnological applications. Indeed, the link between between optimization and decision problems provides tools to capture biological systems response and integrate them in engineering applications.

The above notions were formalized and presented in an article published as an opinion paper in the proceedings of 13th International Conference of Computational Methods in Systems Biology, 2015 (Budinich et al., 2015).

OPINION PAPER
Evolutionary Constraint-based Formulation Requires New Bi-level Solving Techniques

Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, Damien Eveillard

LINA, UMR 6241 CNRS, EMN, Université de Nantes,
2 rue de la Houssinière, Nantes, France

Abstract. Constraint Based Methods had been successfully used to simulate genome-scale metabolic behaviors over a range of experimental conditions. In most applications, environmental constraints are parameterized, and the use of metabolic reactions and corresponding genes is the direct consequence of the tuning of these parameters.
However, in evolutionary studies, the problem is different: one knows the relative importance of reactions and one seeks environmental conditions that could explain such a biological fitness.
This study details this modeling paradigm change and discuss a putative formalization of such a biological problem in the form of a Mixed Integer Bi-level Linear Problem (MIBLP). Unfortunately, solving a MIBLP is difficult, paving the way for the need of further constraint based method developments for understanding evolutionary processes.

Constraint Based Methods (CBMs) are considered as efficient approaches to predict phenotypic responses and explore the structure of genome-scale networks of a variety of organisms [1, 2]. For instance, they tackle effects of genetic mutations (resp. gene deletions [3, 4] and gene insertion [5]) on metabolic behaviors, whereas complementary analysis focused on gene transfers [6], gene dispensability [7] or nutrient adaptation [8]. Similarly, high-throughput sequencing allows today to compare lineages and biological studies to infer evolutionary patterns [9], paving the way to bridge evolutionary studies and CBMs.

From an evolutionary viewpoint, environment exerts or relaxes pressure in biological systems. Thus, in front of detrimental or beneficial environments, organisms adapt themselves by gaining or loosing functions [10, 11]. Those knowledge being available nowadays, it is of great interest to decipher the environmental conditions that maximize lineage evolution, pointing conditions that could lead to metabolic reaction losses [12].

When CBM is applied in evolutionary contexts, environment usually is first parameterized and its effect is then studied and interpreted *via* a range of simulations [6, 13]. Herein, instead of standard approaches, we propose to focus on selecting environmental conditions that make most reactions unable to carry fluxes (see Fig. 1a). Indeed, recent evolutionary studies hypothesize that such blocked reactions are likely to be lost as functions due to evolution [12].

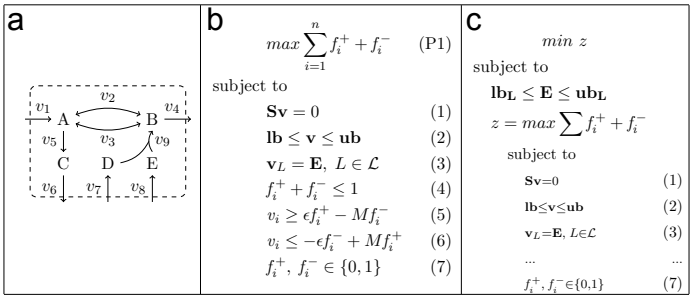


Fig. 1. Evolutionary problem formulation. Considering a putative metabolic network (a), we assume the production of metabolite B as a fitness proxy. If A is the only substrate in a particular environment, we expect that genes coding for v_7, v_8 and v_9 disappear upon evolution. b) The inner Problem (P1) identify blocked reactions, i.e., those that can not carry a non-zero flux under steady-state conditions. A variation of (P1) is used in [14, 15]. c) A mixed integer bi-level linear problem seeking for an environmental setting (i.e, defined values for environmental variables in $\mathcal{L}$, see text) $\mathbf{E}$ that maximizes the number of blocked reactions.

Formalization of the previous statements leads to an optimization problem as shown in Fig. 1b. Constraints in (1) and (2) are mass balance and boundary conditions. Equations in (3) represent environmental variables as a subset of reaction fluxes indexed by $\mathcal{L}$.

To identify blocked reactions, we introduce for each reaction i two binary variables f_i^+ and f_i^- (resp. forward and reverse flux) in (7). Constraints in (4), (5) and (6) guarantee that a reaction i is blocked if and only if $f_i^+ = f_i^- = 0$. By M (resp. ϵ), we denote a large (resp. small) number. Given an environmental setting $\mathbf{E}$, maximizing $\sum f_i^+ + f_i^-$ identifying all blocked reactions.

As a next step in our study, we propose to use the Mixed Integer Bi-level Linear Problem (MIBLP) shown in Fig. 1c in order to select an environmental setting $\mathbf{E}$ that maximizes the number of blocked reactions. The main difference with other bi-level approaches is the focus on controlling metabolic networks using only environmental variables and not genetic manipulations [16].

Unfortunately, despite several tentatives [17, 18], no general solution is available for this type of problem [19], emphasizing the need for an *ad-hoc* algorithm implementation to solve this new evolutionary problem. Furthermore, for the sake of generalization, any method that handle this type of bi-level program, will lead to theoretical and practical advances in system biology.

From an evolutionary viewpoint, we expect that solving this problem will pinpoint the environmental conditions that are responsible for the specification of lineages or microbial strains. This question is particularly vivid considering drastic environmental condition changes that are expected in a near future.

References

1. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict metabolic and associated cellular functions. Nat Rev Genet 15, 107-120 (2014).
2. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. Nat Rev Microbiol 10, 291-305 (2012).
3. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84, 647-657 (2003).
4. Tepper, N. & Shlomi, T. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 26, 536-543 (2010).
5. Larhlimi, A., Basler, G., Grimbs, S., Selbig, J. & Nikoloski, Z. Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. Bioinformatics 28, i502-i508 (2012).
6. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37, 1372-1375 (2005).
7. Papp, B., Pál, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429, 661-664 (2004).
8. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420, 186-189 (2002).
9. Koonin, Eugene V, The Logic of Chance: The Nature and Origin of Biological Evolution FT Press, (2011),
10. Van Valen, L. A new evolutionary law. Evolutionary theory 1, 1-30 (1973).
11. Van Valen, L. Molecular evolution as predicted by natural selection. J Mol Evol 3, 89-101 (1974).
12. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio 3, (2012).
13. Yang, H., Roth, C. M. & Ierapetritou, M. G. A rational design approach for amino acid supplementation in hepatocyte culture. Biotechnol Bioeng 103, 1176-1191 (2009).
14. de Figueiredo,L. F., Podhorski,A., Rubio,A., Kaleta,C., Beasley, J. E.,Schuster S., & Planes, F. J. Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics 25, 3158-3165 (2009).
15. Goldstein, Yaron A. B. & Bockmayr, Alexander. A Lattice-Theoretic Framework for Metabolic Pathway Analysis. 1?14 (2013)
16. Chowdhury, A, Zomorrodi, A.R & Maranas, CD. Bilevel optimization techniques in computational strain design Computers and Chemical Engineering, 72,363–372 (2015)
17. Saharidis, G. K. & Ierapetritou, M. G. Resolution method for mixed integer bi-level linear problems based on decomposition technique. J Glob Optim 44, 29?51 (2008).
18. Xu, P. & Wang, L. An exact algorithm for the bilevel mixed integer linear programming problem under three simplifying assumptions Computers & Operations Research. Computers and Operation Research 41, 309-318 (2014).
19. Georgios K. D. Saharidis, Antonio J. Conejo, George Kozanidis Exact Solution Methodologies for Linear and (Mixed) Integer Bilevel Programming Chapter 8 in Metaheuristics for Bi-level Optimization, Springer Berlin Heidelberg, (2013)

Modeling Microbial Communities: Accounting for Multiple Metabolic Networks

6

Compartment Definition: Effects on Quantitative Modeling

6.1 Introduction

Section I of the present thesis explored applications of CBMs to single organisms. In nature, however, organisms are often found living in communities, forming intricated exchange networks. Recent advances in genome-scale metabolic network reconstruction paved the way to the use of quantitative modelings such as FBA. However, despite the great interest of these techniques to tackle quantitative features, microbial community modeling remains unclear. Two recent reviews list four types of approaches to model metabolic networks in microbial communities: *Lumped* (or *"Soup"*), *Compartmentalization*, *Dynamic* and *Bi-Level* (Biggs et al., 2015; Perez-Garcia et al., 2016).

Briefly, the Lumped approach suggests to collecting all reactions in a single metabolic network and to apply conventional analysis such as FBA to this entity. Compartmentalization, in the other hand, suggests to treat each microorganism as a different "compartment", meaning that metabolites and reactions are considered to be spatially separated if they belong to a particular microorganism. This is implemented by labelling metabolites and reactions accordingly to their compartments. In addition, Bi-Level approach also considers compartments (among other elements) while Dynamic approach differentiates from previous approaches by focusing on time evolution of metabolic networks. Further discussion of these approaches will be carried out in Chapter 7.

Because both Lumped or Compartmentalized assumptions implies distinct experimental efforts, this study proposes an analysis of the consequences of choosing one or the other on microbial community metabolic models. The objective is to study if FBA-like methods predictions differ qualitatively and/or quantitatively if a Lumped or a Compartment approach is used to model a given community metabolic network. It is worth to notice herein that the matter was analyzed in Klitgord and Segrè (2009). In their study, authors proposed an analysis over the impact of compartmentalization in metabolic flux models by considering yeast metabolic network as an ecosystem of organelles. They concluded trough a series of FBA simulations that Lumped models over predicted the amount of biomass produced with respect to their Compartmentalized counterparts. In this chapter, we propose to extend Klitgord and Segrè result in a more automatic manner on a realistic microbial ecosystem [1].

Both Lumped and Compartmentalization approaches are compared in terms of their predictive capabilities. As application, two microbial ecosystems are analyzed: a hot spring microbial community, represented by *Synechococcus spp.*, *Chloroflexus spp.*and *Roseiflexus spp.* and Sulfur Reducing Bacteria, (Taffs et al.,

2009), and a microbial methanogenic system composed by *D. vulgaris* and *M. maripaludis* (Stolyar et al., 2007). To compare differences in both approaches, each system is described as a Lumped model or a Compartment model. Next, a technique called "Flux Modules' (Müller and Bockmayr, 2013), is applied to the four models (Lumped and Compartment version of hot spring model and Lumped and Compartment version of methanogenic system). Flux Modules compute sets of reactions which are strongly correlated (a module). In addition, this decomposition describe uniquely the flux space. By comparing which reactions compose each modules, flux spaces of Lumped and Compartment models will be compared.

6.2 Material and Methods

6.2.1 FBA, FVA and Flux Modules

Flux Balance Analysis (FBA) calculates the maximal value of biomass function (the specific growth of an organism). However, in general, multiple combinations of fluxes can lead to the same maxima. Flux Variability Analysis (FVA) improves this unique description by calculating the minimal and maximal values of each flux near the optimal value. This notion of solutions near the optimal is implemented by adding constraints over the biomass, specifying that it should attain at least a percentage (usually %95-%99) of their theoretical maxima (calculated using FBA).

Besides optimization based techniques, other approaches study CBMs by analyzing the whole solution space. Recently, promoting a systematic exploration of these solutions, The Flux Module (FM) technique (Müller and Bockmayr, 2013) analyzes how, among all solutions, some reactions are systematically correlated - emphasizing subnetworks that connect a subset of substrates and products (Kelk et al., 2012). Mathematically, these subnetworks or modules are unique and they stem from all potential quantitative solutions. Intuitively, different modules imply (potentially) different quantitative predictions, as solutions spaces are not equal.

6.2.2 Ecosystems Models

Hot Spring Mat Community

For the sake of application, one first considered the phototrophic microbial community system during day light composed *Synechococcus spp.*, abbreviated SYN, filamentous anoxygenic phototrophs related to *Chloroflexus spp.* and *Roseiflexus spp.*, abbreviated FAP, and sulfate reducing bacteria (abbreviated SRB, Taffs et al. (2009)). This community consumes CO_2 and releases O_2 by photosynthesis. As a byproduct of the rubisco activity, glycolate is produced by SYN, which will be later used as an organic substrate by FAP, along with acetate. Besides, SRB can consume organic compounds and reduce sulfate using H_2. Compartmentalized community model describes a metabolic network for each strain as well as external metabolites such as H_2, O_2, NH_3, glycogen and acetate (136 reactions, Taffs et al. (2009)). As a modeling contribution, a community biomass function was included to represent the ecosystem growth plus one extra reaction for preserving O_2 / CO_2 ratio as used by rubisco. As reported in Taffs et al. (2009), the so-called "Pool model" represents the Lumped community model (59 reactions: 48 core and 11 exchange reactions).

Methanogenic Anoxic Community

Syntrophy is a form of microbial mutualism, usually involved in organic matter degradation. In this type of interactions, transfer of metabolites between species is essential for growth. A well known case of syntrophy is the association between methanogenic archaea and hydrogen-producing microorganisms

1. This chapter was published first as pre-print in bioRxiv (http://dx.doi.org/10.1101/018010) and hal.archives-ouvertes.fr (https://hal.archives-ouvertes.fr/hal-01145858). However, in those texts, Lumped and Compartiment terms were exchanged to Single Cell Hypothesis (SCH) and Multiple Compartment Hypothesis (MCH), respectively.

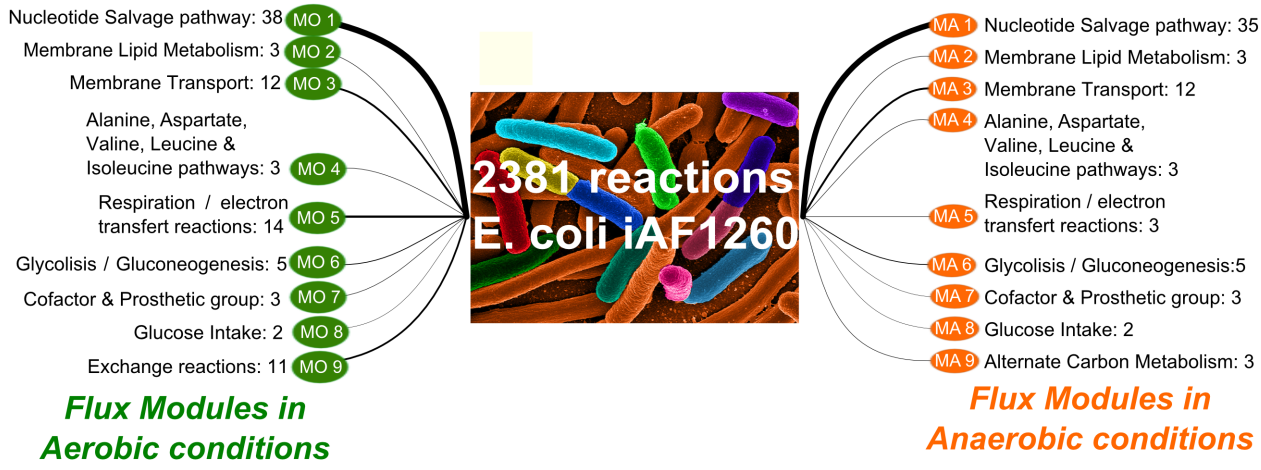


Figure 6.1 – Biological relevance of modules. Application of modules on *E.coli*, assuming availability of O_2 (aerobic, in green) and assuming no presence of O_2 (anaerobic, in orange)

such as *Desulfovibrio vulgaris* and *Methanococcus maripaludis* (Stolyar et al., 2007). Briefly, *D. vulgaris* degrades lactate producing H_2, CO_2, formate and acetate in the process. *M. maripaludis* scavenges H_2 to produce CH_4.

The metabolic model of *D. vulgaris* contains 145 reactions (Zomorrodi et al., 2014), whereas *M. maripaludis* model is composed of 97 reactions (Stolyar et al., 2007). In order to link both strains within a Compartmentalized model, we duplicated exchange reactions of H_2 in order to import/export the metabolite with either the other microorganism or environment. A similar procedure was done for formate, acetate and CO_2, which overall introduces 12 exchange reactions. Finally, the whole ecosystem biomass was design to fit biomass functions of *D. vulgaris* and *M. maripaludis* as already published, while maintaining a respective proportion of 2:1 for both strains. The Lumped model of this community consists in merging both metabolic networks and removing all replicated reactions for considering one unique representative reaction. As result, Lumped model is composed of 221 unique reactions, as a reduction of 243 reactions of the Compartmentalized model (respectively 145 and 97 reactions for *D. vulgaris* and *M. maripaludis*). Lumped model presents a disadvantage regarding the interplay between interchange fluxes; in this case, acetate and H_2 are major players of electron transfer in anaerobic systems which role is an active area of research. Besides, for these two systems, Lumped model links the fluxes of pentose phosphate system which could impact interpretations in future *in-silico* developments.

6.3 Results

6.3.1 Flux Modules are biologically relevant

In order to test the biological relevance of Flux Modules, this technique was applied over *E. coli* model iAF1260 (Feist et al., 2007) in (i) aerobic and (ii) anaerobic conditions. Figure 6.1 depicts modules obtained in *E. coli* for both aerobic (left) and anaerobic (right) growth conditions. For each condition, one extracts 9 modules that are composed of distinct numbers of reactions (line width is proportional). Each module is associated to the pathways in which the flux module reactions are involved. Modules are in accordance to biological conditions. When challenged by an oxidative stress, most of flux modules of *E. coli* are conserved, except exchange reaction, respiration & electron transfert, alternate carbon metabolism, which is in accordance to physiological knowledge. To a lesser extent, nucleotide salvage pathway is impacted by oxygen growth conditions.

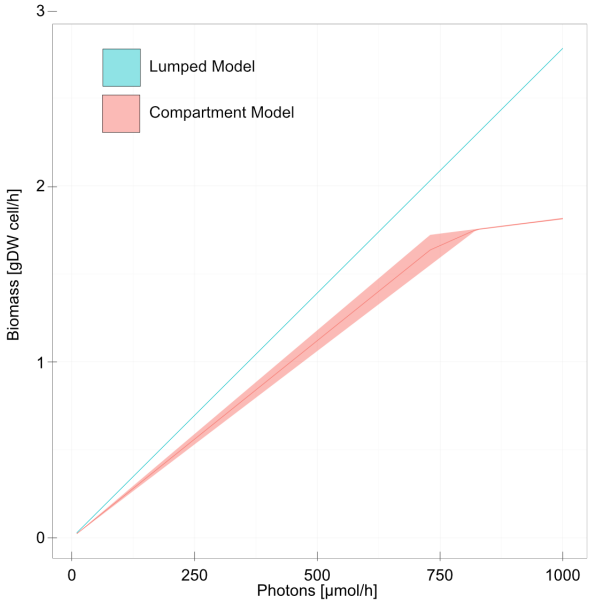


Figure 6.2 – Quantitative simulations of Lumped and Compartment models of a hot spring microbial mat system. For a fixed photon influx (abscissa axis), a FBA and a FVA (using biomass > 95% of max. biomass) where run to calculate the biomass boundaries (ordinate axis). Solid lines represent the average between minimal and maximal biomass. For the case of Lumped model, minimal and maximal values of biomass were equal. In the Compartment model, biomass varied for photon influx between 0 - 760 [μmol/h]

.

6.3.2 Metabolic modules of a three guild microbial system composed in a Hot Spring Mat

Total biomass of the microbial mat community was calculated by using FBA and FVA in both Compartment and Lumped model. (Figure 6.2). Both models show similar values of biomass production at each photon influx, in agreement with previous results and qualitatively consistent with available experiments (Stolyar et al., 2007; Zomorrodi et al., 2014). Naively, these similar predictions may lead to over-interpret that both models are identical, which do not advocate for the use of Compartmentalized that is experimentally expensive.

However, Compartment and Lumped models produce distinct modules, which clearly emphasizes fundamental differences between Compartmentalized and Lumped solutions (Figure 6.3). Lumped shows only one module (purple reactions in Figure 6.3), containing 31 reactions (52.2% of overall reactions): 20 reactions covered by Compartmentalized modules and 11 not previously highlighted. 7 reactions in Compartmentalized module do not belong to the Lumped module. Compartmentalized model SYN reactions are decoupled from other networks, confirming previous studies that highlights SYN as a primary producer for all possible microbial interactions (Taffs et al., 2009). Complementary, FAP and SRB are linked via acetate and H_2 metabolisms. As additional differences, the first glycolysis phase (R1-R2) and pentose phosphate reactions (R5-R9) are connected in Lumped model, which is not the case when each organism is considered separately. Lumped model module is independent from uptake reactions; whereas Compartmentalized model modules depict acetate processing of FAP and SRB linked to O_2, H_2 and CO_2 exchanges.

6.3.3 Metabolic modules of a methanogenic microbial system composed of *Desulfovibrio vulgaris* and *Methanococcus maripaludis*

For the sake of illustration, a similar modeling comparison was applied on a methanogenic microbial system composed of *Desulfovibrio vulgaris* and *Methanococcus maripaludis*(Figure 6.4). *D. vulgaris* uses

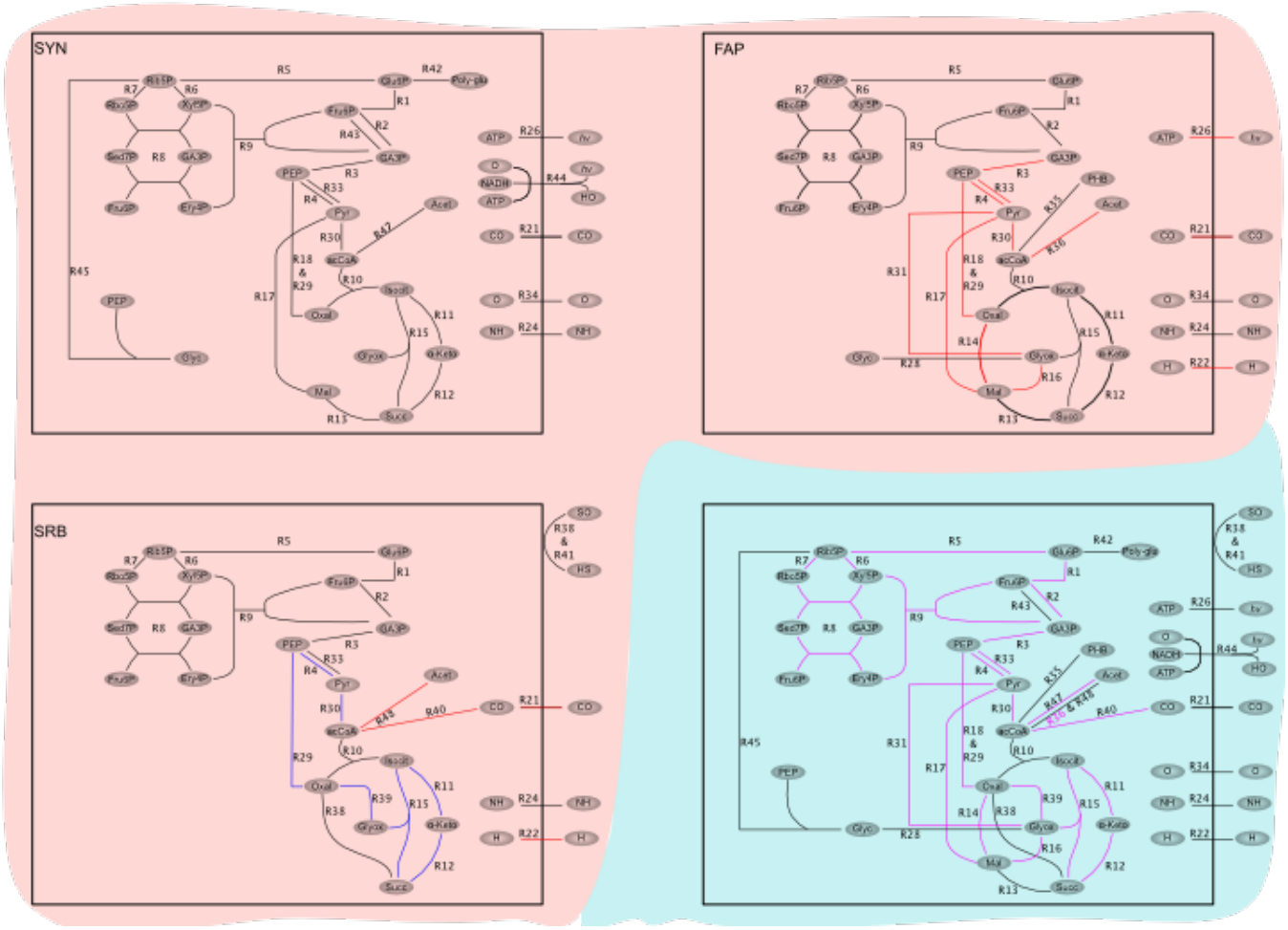


Figure 6.3 – Description of metabolic networks related to the microbial mat community system and corresponding modules illustrations. SYN, FAP and SRB depict bacterial strains of the Compartment metabolic model (highlighted in soft red background). Lumped model (highlighted in soft blue background) represents the same metabolic system with no consideration of the compartments, while conserving the naming convention of the Compartment model. For the sake of illustration exchange reactions between compartments are not shown. Compartment model reveals 2 modules (26.5% of the whole set of metabolic reactions). One module contains 28 reactions (red) that span through FAP and SRB, whereas another (blue) involves 8 reactions. Reactions of the Lumped module are depicted in purple.

lactate fermentation and sulfur reduction to gain energy, while producing gaseous hydrogen. *M. maripaludis* uses hydrogen to reduce CO_2 into methane, which avoid the accumulation of H_2 that might decrease the chemical energetic potential of *D. vulgaris*. The corresponding Compartmentalized model has 243 reactions (respectively 145 and 97 reactions for *D. vulgaris* and *M. maripaludis*, Figure 6.4A) . Lumped model is composed of 221 unique reactions, after deletion of redundant reactions. Again Compartmentalized and Lumped models modules are different. Both models show a unique module with 124 reactions (48.6% of Compartmentalized model) and 187 reactions (84.6% of Lumped model), as illustrated in Figure 6.4B and C. Reactions related to H_2 and acetate transport are not related in Lumped model, whereas they are in Compartmentalized model. Additionally, pentose phosphate cycle reactions of *D. vulgaris* and *M. maripaludis* are linked in the Lumped model module but not in the Compartmentalized model module, which might lead to erroneous biological interpretations.

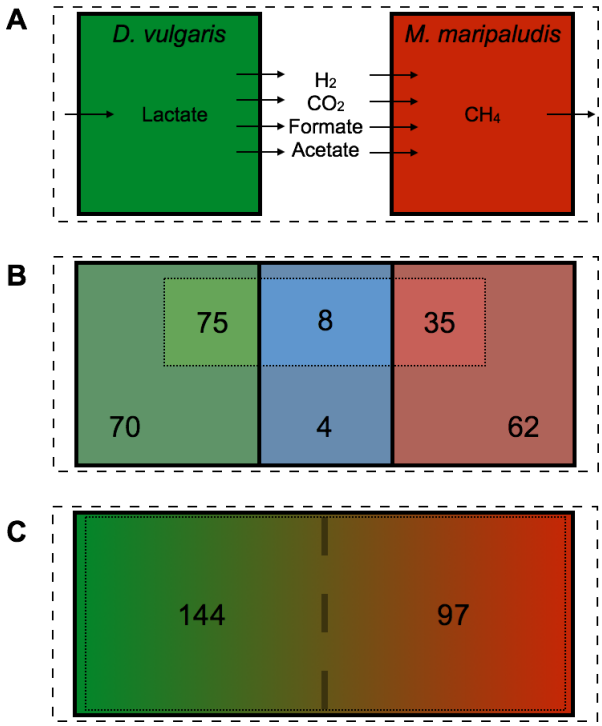


Figure 6.4 – Description of metabolic networks related in the *D. vulgaris* (green) and *M. maripaludis* (red) community model. Blue depicts a compartment were exchange reactions occurrs. A) Depiction of metabolite exchange between *D. vulgaris* and *M. maripaludis*. B) Numbers of reactions in the Compartment model; green represents *D. vulgaris* and red represents *M. maripaludis*. Reactions participating in module detected are highlighted in lighter tones. C) Number of reactions involved in the Lumped module detected.

6.4 Discussion

Despite similar quantitative values obtained by FBA and FVA, this study shows significant differences between Lumped and Compartmentalized models. However, we do not advocate for either of both modeling assumptions. Biologically, Lumped models have been widely employed to study metabolite exchanges between species (e.g., cocultures Wintermute and Silver (2010); Hanly and Henson (2010) or species within a complex environment Klitgord and Segrè (2011)), whereas Compartmentalized models have been used to describe microbial communities, where each member seeks to maximize their own biomass (Tzamali et al., 2011). Both assumptions are equivalent when one is interested by predicting overall quantitative behaviors of a microbial community, which is mostly explained by similar exchange reactions between Lumped and Compartmentalized models. A protocol driven by Lumped modeling might be mostly sufficient for overall predictions with non further functional investigations. Reversely, protocols driven by Compartmentalized models present a significant cost to decipher boundaries between species and origin of genes within a metagenome (Thiele et al., 2013), but appear as necessary to investigate fine quantitative interactions within the community.

From a methodological viewpoint, this study advocates for the use of Flux Modules to compare metabolic models. Modules represent an abstraction of all Flux Variability simulations for a given metabolic model. Indeed Flux Module technique is a natural way to resume the methodological work of Klitgord and Segrè (2009) that proposes an extensive analysis of yeast metabolic flux estimation with and without compartmentalization. Since our study pinpoints similar conclusions to Klitgord and Segrè (2009), both studies reinforces the need for further constraint-based modelings dedicated to multiple compartments simulations as motivated by Zomorrodi and Maranas (2012) and Zomorrodi et al. (2014).

7

Extensions of Constraint Based Methods: Multiple Objectives

Following results of previous chapter, modeling of microbial ecosystem would benefit from constrained approches capable to deal with multiple compartments. In the following, we will study more deeply how microbial communities can be modeled using different compartments.

Recently, two publications by Biggs et al. (2015) and Perez-Garcia et al. (2016), had reviewed the use of metabolic modeling in communities. In both works,with respect to CBMs, authors mention four approaches: Lumped or "Soup", compartmentalization, bi-level optimization and dynamic extensions.

Lumped or "Soup" approach is perhaps the most straight-forward approach. It consists into ignore the boundaries between species and put all detected reactions in a single entity, assuming a generalized biomass function which represents the whole community. Here, the focus is put in the metabolic capabilities of all organisms present. However, it has been noted that this approach changes basic properties of the network and the accuracy of flux values (see Chapter 6 and Klitgord and Segrè (2009)).

In **compartmentalization** approach, each specie is modeled as a "compartment" of the network and exchangeable metabolites are shared trough an extra compartment, common to all members, which represents the extracellular environment. Individual metabolic models are then incorporated into a *ecosystem metabolic matrix* and each metabolite is defined as entity accordingly to how many compartments it participates. For example, a metabolite M_i in a system with two species $\mathbf{A}$ and $\mathbf{B}$ will be defined three times as $M_i^{\mathbf{A}}$, $M_i^{\mathbf{B}}$ and $M_i^{\mathbf{C}}$, where subscripts $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ denote each compartment ($\mathbf{C}$ stands for the shared compartment). Additionally, exchange reactions between each specie and the shared extracellular space are included, allowing to capture interactions such as mutualism or competition (Stolyar et al., 2007; Taffs et al., 2009; Klitgord and Segrè, 2010; Khandelwal et al., 2013; Hanemaaijer et al., 2015). In these formulations, the biomass function of the system is usually modeled as a sum of the biomass of all modeled species, which is then optimized.

Dynamic extensions are designed to overcome the steady state hypothesis in CBMs, by including kinetic and differential equations which capture the dynamics of the process. Generally speaking, these methods divide the simulation time in intervals, where kinetic equations are used to estimate uptake rates. Then, these values are feed to the optimization problem, which allows estimation of fluxes to feed differential equations, in order to calculate variations in metabolite concentration and biomass. Finally, these concentration are used as starting point for the next round of simulation (Mahadevan et al., 2002). While capturing metabolic complexity as well as dynamic behavior, these approaches have two major drawbacks: (i) They are computationally demanding and (ii) They require knowledge of kinetic parameters such as maximum reaction rates and kinetic constants, which is somewhat opposed to the no-parameter advantage

of CBMs.

The **Bi-Level Optimization** approach is constituted by the OptCom framework (Zomorrodi and Maranas, 2012). In this work, authors present two closely related bi-level formulations (OptCom and Descriptive OptCom, respectively) which address the problem of modeling metabolic microbial interactions as several optimizations problems. OptCom formulation is as follows:

$$\begin{array}{ll} \text{maximize} & z = \textit{Community-level objective} \\ \text{subject to} & \end{array}$$

$$\left\{\begin{array}{ll} \text{maximize} & (\mathbf{c^{k_1}})^\intercal \mathbf{v^{k_1}} \\ \text{subject to} & \\ & \mathbf{S^{k_1}}\mathbf{v^{k_1}} = \mathbf{b^{k_1}} \\ & \mathbf{l^{k_1}} \leq \mathbf{v^{k_1}} \leq \mathbf{u^{k_1}} \\ & u_i^{k_1} = uval_i^{k_1}, \forall i \in \mathbf{I}_{up}^{k_1} \\ & e_i^{k_1} = eval_i^{k_1}, \forall i \in \mathbf{I}_{ex}^{k_1} \end{array}\right\}, \cdots, \left\{\begin{array}{ll} \text{maximize} & (\mathbf{c^{k_N}})^\intercal \mathbf{v^{k_N}} \\ \text{subject to} & \\ & \mathbf{S^{k_N}}\mathbf{v^{k_N}} = \mathbf{b^{k_N}} \\ & \mathbf{l^{k_N}} \leq \mathbf{v^{k_N}} \leq \mathbf{u^{k_N}} \\ & u_i^{k_N} = uval_i^{k_N}, \forall i \in \mathbf{I}_{up}^{k_N} \\ & e_i^{k_N} = eval_i^{k_N}, \forall i \in \mathbf{I}_{ex}^{k_N} \end{array}\right\}$$

Inter-organism flow constraints

Each $\mathbf{k_i}$ represents the organisms from the same species present in the system, also called *guilds*. $\mathbf{k_1}$ to $\mathbf{k_N}$ represent $\mathbf{N}$ different guilds present in the system. u_i and e_i are particular fluxes which make the uptake or export of shared metabolites, represented by $\mathbf{I}$. $uval_i$ and $eval_i$ are parameters for the inner problem (each guild) which are imposed for the outer problem (the ecosystem). Inter-organism constraints controls the behavior of $uval_i$ and $eval_i$ and is where the ecological relation is described. For example, in a syntrophy scenario between two communities, where the organism $\mathbf{k_1}$ produce the metabolite i which is consumed by $\mathbf{k_2}$ (and no import or export is made form outside the system) the inter-organism constraints are simply described by:

$$eval_i^{k_1} = uval_i^{k_2}$$

Solving approach for this formulation relies into the Primal-Dual theorem (Schrijver, 2011). First, for a given optimization problem, we can state a related optimization problem named *Dual*; for clarity, the original problem is termed as the *Primal* problem.

Definition. *Given the following LP problem,*

$$\begin{array}{ll} \text{minimize} & z = \mathbf{c}^\intercal \mathbf{x} \\ \text{subject to} & \\ & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \in \mathbb{R}^n \end{array}$$

We call the Dual *problem to the following derived LP problem,*

$$\begin{array}{ll} \text{maximize} & w = \mathbf{y}^\intercal \mathbf{b} \\ \text{subject to} & \\ & \mathbf{y}^\intercal \mathbf{A} = \mathbf{c}^\intercal \\ & \mathbf{y} \in \mathbb{R}_+^m \end{array}$$

Where $\mathbf{A} \in \mathbb{R}^{m \times n}$, $z \in \mathbb{R}$, $w \in \mathbb{R}$, $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$ *and* $\mathbb{R}_+$ *is the set of real numbers* ≥ 0. ***y*** *components are called* dual variables, ***b*** *is called the* right-hand side *vector and* ***c*** *the* objective function.

Corollary. *From the* Dual *definition, there is a one-to-one correspondence between constraints and variables of both problems, summarized in the following table:*

maximize	minimize
$\leq$ constraint	variable ≥ 0
$\geq$ constraint	variable ≤ 0
$=$ constraint	unconstrained variable
variable ≥ 0	$\geq$ constraint
variable ≤ 0	$\leq$ constraint
unconstrained variable	$=$ constraint
right-hand side	objective function
objective function	right-hand side

Theorem. *If* Primal *or* Dual *are feasible, then the other is feasible too. Furthermore, if* (z, w) *are their respective optimal values, then* $z = w$.

By adding dual constrains to primal problem and an equality constraint given by the Primal-Dual theorem, inner optimization problems are transformed into a set of constraints. OptCom approach uses this strategy to solve the community model, by transforming inner optimization problems into a set of constraints and keeping the outer optimization problem.

However, these transformations induce non-convexity in the problem. In general, bi-level linear programs belong to the NP-Hard class of problems (Bard, 1991). Moreover, checking the local optimality in a continuous linear bi-optimization problem is a NP-Hard problem (Vicente et al., 1994).

7.1 Multiple Objective Optimization

Motivated for the non-convexity of current formulations, an alternative mathematical framework permitting: (i) Considering multiple organisms as multiple compartments, (ii) Considering a (different) objective function for each of these microorganisms and (iii) Maximizing all these objective functions, was searched in the literature.

From a mathematical perspective, when more than one objective are required to be optimized simultaneously, the problem is referred as a **Multiple Optimization Problem** or **Multi Objective Problem** (MOP). A key concept in MOPs is the notion of **Pareto Optimality**, which can be informally defined as follows: A feasible solution to a MOP is considered to be Pareto Optimal if any of their objectives can not be improved simultaneously; *i.e.*, increasing the value of one objective will reduce the value of the others. Therefore, solution points of MOP configures a **Pareto Front**, a set of points which describe the trade-off between different objectives. Therefore, "to solve" a MOP will be considered as to find a subset of points that belong to the Pareto Front.

As organisms are thought to optimize several functions (e.g., biomass, ATP consumption, total flux; see Schuetz et al. (2012) for a detailed study), these type of problems have been considered by some authors in the context of CBMs. Vo et al. (2004) used Pareto Optimality concept to study flux distributions for human mitochondria model. In their work, authors were interested in how this organelle distributes resources when maximizing ATP production to the cell, Heme group biosynthesis and phospholipids biosynthesis objectives. To this end, they used two different approaches: weighted sum and lexicographic method. In weighted sum, an objective function is constructed as the sum of the different objectives weighted by a positive coefficient, transforming the multi objective problem into a single objective problem. In the lexicographic approach, each objective is hierarchically ordered. Then, a serie of single objective optimization problems is solved, where each objective is optimized according their order and then the result is added as a constraint to the following problem. In the case of mitochondria, both methods yielded similar results. If ATP flux is maximized, neither heme or phospholipids could be maximized. By contrast, these two last objectives can be maximized simultaneously if ATP demand is strictly less than the maxima. This is interpreted as heme and phospholipids operate relatively independently one from the other, whereas both objectives consume energy in form of ATP, reducing available ATP to the cell.

Oh et al. (2004, 2009) investigated flux distribution profiles to characterize trade offs between different products in a *in-silico E. coli* model. To this end, authors applied the noninferior set estimation (NISE) method. NISE method generate a series of points to construct a lower approximation of Pareto Front, until the distance between two consecutive aproximations is under a threshold. First, each objective is optimized, giving initial points. Then, a new point of the Pareto Front is calculated, by performing a weighted sum optimization where the weights are calculated as the coefficients of the hyperplane supporting the convex hull of the current approximation. If the distance between this new calculated point and the current Pareto Front approximation is below an specified threshold, the algorithm stops. Otherwise, this point is included in the estimation and a new point is calculated. Authors used this approach to investigate solutions sets which maximized the biomass while maximizing product outcome, allowing them to propose candidates genes for knock-out based in flux distributions within the model.

Nagrath et al. (2007, 2010) have focussed in hepatocytes. In these systems, cells usually do not go under proliferation, so growth rate can not be directly used as optimization objective by FBA techniques. Instead, these cells perform an array of metabolic functions were multiple objectives should be taken into account. In Nagrath et al. (2007), authors combine Flux Balance Analysis and Energy Balance Analysis (EBA, a variant of FBA that focuses in thermodynamic constraints) to investigate pair combinations of liver-specific objectives. Normal Constraint (NC) method is used as optimization technique. In this method, a series of evenly distributed points in the Pareto Front are generated. First, anchor points are calculated by optimizing each objective. Next, the objective space is normalized and then a predefined number of evenly distributed points along the hyperplane between anchor points (the utopia plane, UP) are calculated. Finally, a normal to the UP (NU) is determined to reduce the feasible region and allows to calculate a set of solutions by a single optimization problem. This set of solutions are used to calculate Pareto points in the original space. In Nagrath et al. (2010), authors changed the NC approach by Linear Physical Programming (LPP). Generally speaking, in LPP objectives are classified in *Soft* and *Hard* classes, depending if objectives follow a "Larger/Smaller is better" or a "Must be larger/smaller" constraint type. These preferences are feed into an Aggregate Objective Function (AOF) which is minimized. In both works, authors emphasize as result a serie of trade-offs between functions such as NADPH, ATP, ammonia and albumin production under different set of conditions, which can be used in the design of a bioartificial liver.

Pozo et al. (2012) propose an optimization method to cope with models including dynamical aspects, assuming a generalized mass action (GMA) for reactions. By using GMA, is possible to take into account changes in basal levels of enzymes. Then, the proposed formulation seeks to maximize a given product while minimizing metabolite concentration and the individual changes in enzyme activities. To solve this problem, authors used an ϵ-constraint method, which consists in to minimize one objective while setting additional constraints for the others by setting upper bound of value ϵ. By varying these upper bounds values, a set of solutions belonging to the Pareto Front are obtained. Finally, these solutions are filtered to (i) discard indistinguishable alternatives and (ii) select variables with "good" performance in all objectives. In this way, interesting solutions can be ranked and tested in the laboratory. For application, authors used the metabolic network of *Saccharomyces cerevisiae* and optimized ethanol production.

7.1.1 Formulation and Concepts in Multi Objective Optimization

In this section we presents some definitions and results from which will help us to formalize concepts to model microbial ecosystems from a mathematical perspective. In particular, our objective is to set a general framework as it was set for FBA and FVA.

One of the main difficulties in these kind of problems is that solutions are not a single values (such in the case of LP problems a in Flux Balance Analysis), but a vector, which are not always comparable. To illustrate this point, we can compare the Single Optimization Problem against a Multi Objective Problem (Ehrgott and Wiecek, 2005):

Single Objective Problem (SOP)	Multi Objective Problem (MOP)
maximize $f(x) \in \mathbb{R}$	maximize $(f_1(x), \ldots, f_p(x)) \in \mathbb{R}^p$
subject to $x \in X \subseteq \mathbb{R}^n$	subject to $x \in X \subseteq \mathbb{R}^n$

In MOP literature, $\mathbb{R}^n$ and $\mathbb{R}^p$ are often referred as *decision space* and *objective space*, respectively. Historically, MOP problem arrises from choosing values for decision variables (*i.e.*, picking a certain $x \in \mathbb{R}^n$); given such x, we will compute the values of the objective vector $f(x) = (f_1(x), \ldots, f_p(x)) \in \mathbb{R}^p$.

Set $X \subseteq \mathbb{R}^n$ is the set of possible values for the arguments x of f. X is given in the form of constraints, *i.e.*, $X := \{x \in \mathbb{R}^n : h_j(x) = 0, j = 1, \ldots, k; g_j(x) \leq 0, j = 1, \ldots, l\}$, where h_i and g_i are functions. For example, in the case of flux space $F := \{\mathbf{v} : \mathbf{Sv} = \mathbf{b}, \boldsymbol{l} \leq \mathbf{v} \leq \boldsymbol{u}\}$, is possible to identify $X = F$, $x = \mathbf{v}$ and equality constraints $h_j(\mathbf{v})$ with $h_j(\mathbf{v}) = \mathbf{S_j v} - \mathbf{b}_j$, where $\mathbf{S_j}$ is the corresponding row of matrix $\mathbf{S}$. Similarly, inequalities are described by $\boldsymbol{l} \leq \mathbf{v} \leq \boldsymbol{u}$ by using $g_i(\mathbf{v}) = \mathbf{l}_i - \mathbf{v}_i$ and $g_s(\mathbf{v}) = \mathbf{v}_s - \mathbf{u}_s$. Please note that all given constraints are described by linear functions. Also, the set of all attainable values outcomes is defined as $Y := f(X) \subset \mathbb{R}^p$.

The problem now is to define precisely the meaning of "maximize" for MOPs. In Single Objective Problems (SOP), giving that $f(x) = y, y \in \mathbb{R}$, "maximize" equals to find $y^* \in \mathbb{R}$ such as $\nexists\, y > y^*, y \in \mathbb{R}$. When $(y^*, y) \in \mathbb{R}^p$, the **Pareto** notion can be used : $y^* \geq y$ if and only if $y_k^* \geq y_k$ for $k = 1, \ldots, p$ with a strict inequality for at least some k (i.e, $y^* \neq y$). Similarly, is possible to use $y^* > y := y_k^* > y_k$ for $k = 1, \ldots, p$ to define Pareto notion. Depending of the use or not of strict inequalities,following definition is made:

Definition. *Consider the MOP. A point $x \in X$ is called:*

— **a weakly efficient solution** *if there is no $x' \in X$ such that $f(x') > f(x)$. $y = f(x)$ is called a* **weakly Pareto point**.

— **an efficient solution** *if there is no $x' \in X$ such that $f(x') \geq f(x)$. $y = f(x)$ is called a* **Pareto point**.

In other words, a weakly efficient solution will improve *at least one* component of objective functions; an efficient solution will improve the values of *all* components of objective functions. The set of efficient solutions and weakly efficient solutions are called X_E and X_wE, respectively. Also, their images are denoted by Y_E and Y_wE respectively. In the general case, $X_E \subseteq X_wE$ and $Y_E \subseteq Y_wE$.

Besides efficient and weakly efficient solution, three other useful concepts are defined in MOPs:

Definition. *Consider the MOP. We define the following points:*

— *Ideal point $y^I = (y_1^I, \ldots, y_p^I)$ where $y_k^I :=$ maximize$\{f_k(x); x \in X\}$*

— *Utopia point $y^U = (y_1^U, \ldots, y_p^U)$ where $y_k^U := y_k^I + \epsilon_k$, ϵ_k is a small positive number*

— *Nadir point $y^N = (y_1^N, \ldots, y_p^N)$ where $y_k^N :=$ minimize$\{f_k(x); x$ is a efficient solution$\}$*

Solution Sets in MOPs

Approaches to generate solutions sets for MOPs are divided in two types: scalarization methods and nonscalarization methods, depending if they base their strategy into converting MOP into a SOP (or a series of SOP) or another MOP, respectively. Scalarization methods accomplish this by transforming the set of objective functions in one objective by using an explicit function, while nonscalarization methods uses other means. Additionally, it is worth noting that usually when dealing with CBMs set X is formed by linear restrictions, as showed in 7.1.1 . If objective functions are also linear (such as biomass functions), then the MOP belongs to the category of Multiple Objective Linear Programs (MOLPs). As illustration, three commonly used approaches are presented.

— **Weighted Sum** : This method gives relatives weights to each objective function $f_k(x)$ and minimize their sum:

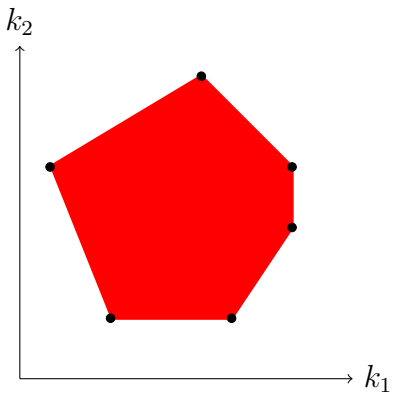


Figure 7.1 – Example of a 2D Convex Polyhedron. Black dots mark polyhedron vertices

$$\begin{aligned} &\text{maximize } \sum_{k}^{p} \lambda_k f_k(x) \\ &\text{subject to } x \in X \end{aligned}$$

Where $\lambda \in \mathbb{R}^p_{\geqq} := \{\lambda_i \in \mathbb{R} : \lambda_i \geq 0; i = 1, \ldots, p\}$. For MOLPs, x^* is a optimal solution of the Weighted Sum problem for some $\lambda \in \mathbb{R}^p_{>}$ if and only if x^* is an efficient solution.

— **ϵ-constraint approach** : This approach retains the k-th objective function as scalar objective and the others are used to generate new constraints:

$$\begin{aligned} &\text{maximize } f_k(x) \\ &\text{subject to } f_i(x) \leq \epsilon_i, \quad i = 1, \ldots, p; i \neq k \\ &x \in X \end{aligned}$$

Let $\epsilon_{-k} = (\epsilon_1, \ldots, \epsilon_{k-1}, \epsilon_{k+1}, \ldots, \epsilon_p)$ and $\Psi := \{\epsilon \in \mathbb{R}^p : \text{Problem } k\text{-th is feasible for } \epsilon_k\}$. If for some $k \in \{1, \ldots, p\}$ exist ϵ_{k-1} such that x^* is an solution of Problem k-th, then x^* it is a weak efficient solution of the original MOP. If x^* is unique, then x^* is a efficient solution.

— **Objective Space Methods**: Multi Objective Linear Programs (MOLPs) are stated as follows:

$$\begin{aligned} &\text{minimize } Cx \\ &\text{subject to } Ax = b \\ &\qquad x_i \geq 0 \quad i = 1, \ldots, n \end{aligned}$$

Where C is a $p \times n$ objective function matrix, A is a $l \times n$ restriction matrix and b $\in \mathbb{R}^l$. Both decision space $X := \{x \in \mathbb{R}^n : Ax = b, x \geqq 0\}$ and objective space $Y := \{y \in \mathbb{R}^p : y = Cx, x \in X\}$ [1], are *polyhedrons* (see figure 7.1). It is known that optimal values of Single Objective Linear Problems are located in the vertices of X polytope. Methods to find them are usually described in optimization literature.

For MOPs, usually dimension of Y is smaller than dimension of X, ie. $p << n$; therefore, some methods have been proposed (Benson, 1998; Ehrgott et al., 2010; Hamel et al., 2013) to exploit this property. In the specific case of metabolic modeling using compartments, p will depend on the number of species (*i.e.* compartments) in the community.

1. We define for $x \in \mathbb{R}^n$, $x \geqq 0$ as $x_i \geq 0$, for $i = 1, \ldots, n$

8

A Multi-Objective Constraint Based Approach for Modeling Microbial Ecosystems at Genome Scale

CBMs represent microorganisms as a set of constraints imposed by their metabolic network (represented by their stoichiometric matrix) and explore the solution space of these constraint using a mathematical representation of cellular objectives (represented by optimization of the objective function). In this context, a CBM to represent a microbial ecosystem needs to fulfill two requirements: (i) Characterization of metabolic networks of the agents involved in the ecosystem (including interactions between them) and (ii) Appropriate mathematical description of their cellular objectives.

In Chapter 6 , quantitative and qualitative differences between *Lumped* and *Compartmentalization* approaches for modeling communities were analyzed using two community models systems. Using modules framework, it was concluded that solution space of both modeling approaches are different in general. Consequently, modeling approaches designed to represent microbial communities should use compartments to obtain accurate predictions of the system.

Considering different organisms in the ecosystem has the added challenge of representing accurately each compartment biological objective in the modeling approach. Seminal studies (Stolyar et al., 2007; Taffs et al., 2009) used a total biomass objective function, represented as the sum of individual biomass rates. However, in this approach, optimization will always favor maximization of the objective with higher value in the objective function. It is possible to circumvent this difficulty by weighting the coefficients of each biomass function, but then the problem is the appropriate choice of such weights.

The question about handling multiple objectives was treated in literature, as a framework to understand conflicting objectives in single cells. The "multi-objective" concept has been used early in CBMs context and have been used to analyze compartments in single cells models (Vo et al., 2004) or different biological functions (Nagrath et al., 2007, 2010). All these work deal with multiple objectives, but being based in a single organism, they does not take into account multiple compartments (see Chapter 7 for a detailed revision).

According to two recent reviews on modeling of microbial communities using CBMs (Biggs et al., 2015; Perez-Garcia et al., 2016), OptCom framework, a bi-level approach proposed by Zomorrodi and Maranas (2012), is capable of capturing several objectives in a compartmentalized model. OptCom frameworks relays in a bi-level formulation with an inner and outer problem. In the inner problem, each compartment represents a microorganism which maximizes their own objective function. In the outer level, interactions between microorganisms such as competition and cooperation can be described by constraints over shared

metabolites, linking the inner and outer problem. Finally, an ecosystem objective function is maximized in the outer problem. Besides OptCom, authors also propose a dynamic extension, called d-OptCom (Zomorrodi et al., 2014). Further details about details of OptCom solution procedure are given in Chapter 7

In this chapter, we propose to model microbial ecosystems as follow. First, to construct the set of restrictions, a systematic way to use individual genome scale models to construct a system stoichiometric matrix is implemented by considering an additional exchange compartment for shared metabolites. (Khandelwal et al., 2013).

Next, it has been shown that biological systems operates under Pareto-Optimal conditions (Schuetz et al., 2012); therefore, is expected that this property will be maintained through different organizational levels. Under this hypothesis, we propose to use the system stoichiometric matrix to set up an MultiObjective FBA (MO-FBA), which maximizes all cellular objectives simultaneously. Solving the MO-FBA corresponds to find the Pareto Front of the system, *i.e.* the set of non-dominated points in the space defined by the cellular objectives functions. Additionally, we propose a MultiObjective FVA (MO-FVA) to explore the range of fluxes in the Pareto Front.

The proposed approach differs conceptually and practically from OptCom framework. First, the multi-objective approach prescinds from a ecosystem objective function and do not requiere prior knowledge of microbial interactions except for the set of shared metabolites. Furthermore, although OptCom considers multiple compartments and seeks to optimize an ecosystem objective while maximizing each microorganism objective, from a mathematical point of view, bi-level formulations are not equivalent to multi-objective ones (Talbi, 2013).

As result, a new method for obtaining a MO-FBA is proposed, which does not rely in assumptions over an ecosystem function. In addition, it delivers as solution a geometrical description of the Pareto Front. This enable the study of fluxes which characterizes different regions of the Pareto Front. In particular, is possible to define a multi-objective extension of the FVA, named MO-FVA, by including previously calculated optima as a restriction set.

Because MO-FBA does not relies in the assumption of an ecosystem objective function, is possible to study different propositions of such objectives. For instance, along with flux values of MO-FVA, is possible to explore criteria suggested by thermodynamical principles and compare them against the total biomass hypothesis. Results suggest that modeled ecosystem makes a compromise between both, conforming a new multi-objective proposition.

The following article was published in PLoS ONE journal.

PLOS ONE

RESEARCH ARTICLE

A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems

Marko Budinich*, Jérémie Bourdon, Abdelhalim Larhlimi, Damien Eveillard

Computational Biology group, LINA UMR 6241 CNRS, EMN, Université de Nantes, Nantes, France

* marko.budinich@univ-nantes.fr

OPEN ACCESS

Citation: Budinich M, Bourdon J, Larhlimi A, Eveillard D (2017) A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. PLoS ONE 12(2): e0171744. doi:10.1371/journal.pone.0171744

Editor: Tamir Tuller, Tel Aviv University, ISRAEL

Received: July 28, 2016

Accepted: January 25, 2017

Published: February 10, 2017

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

Funding: MB is supported by CNRS and Region Pays de la Loire funding (GRIOTE project, http://griote.univ-nantes.fr/). This study is supported by ANR (IMPEKAB, ANR-15-CE02-001-03). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Abstract

Interplay within microbial communities impacts ecosystems on several scales, and elucidation of the consequent effects is a difficult task in ecology. In particular, the integration of genome-scale data within quantitative models of microbial ecosystems remains elusive. This study advocates the use of constraint-based modeling to build predictive models from recent high-resolution -omics datasets. Following recent studies that have demonstrated the accuracy of constraint-based models (CBMs) for simulating single-strain metabolic networks, we sought to study microbial ecosystems as a combination of single-strain metabolic networks that exchange nutrients. This study presents two multi-objective extensions of CBMs for modeling communities: multi-objective flux balance analysis (MO-FBA) and multi-objective flux variability analysis (MO-FVA). Both methods were applied to a hot spring mat model ecosystem. As a result, multiple trade-offs between nutrients and growth rates, as well as thermodynamically favorable relative abundances at community level, were emphasized. We expect this approach to be used for integrating genomic information in microbial ecosystems. Following models will provide insights about behaviors (including diversity) that take place at the ecosystem scale.

Introduction

Microbial organisms comprise approximately 50% of the Earth's biomass [1, 2] and their interplay drives most biogeochemical cycles [3, 4]. The study of microbial interactions, which occur at the molecular scale, remains crucial to the elucidation of larger-scale processes [5]. Several models have attempted to simulate the quantitative impact of molecular-scale processes at an ecosystem level. Among others, trait-based approaches have gained attention as a precise way to understand and predict the quantitative behaviors of microbial communities [6, 7]. However, such models remain difficult to apply to most communities without the additional expertise required for deciphering particular traits and performing extensive experiments to design accurate parameters [8]; such expertise is often unavailable for the study of natural communities.

In the last decade, great advances have been made in the development of high-throughput techniques that enable the study of the metagenomics, meta-transcriptomics, and meta-metabolomics of natural communities. Such techniques provide 'omics-scale information for organisms, from which it is possible to identify specific molecules (*e.g.,* DNA, mRNA, metabolites) present in a particular microbial ecosystem. Such studies of microbial ecosystems have facilitated drastic changes in approaches utilized for characterizing microbial communities [9, 10], thus leading to the emergence of the field of microbial systems ecology. Further, advances in bioinformatics and computational techniques have enabled the development of next-generation sequencing technologies for the qualitative analysis of microbial environments by emphasizing *who is there and who is not* [11] and allowing the study of the co-existence of microbial strains under different environmental conditions (see [12] for illustration). However, among the most significant challenges in modeling microbial communities remains the ability to quantitatively predict microbial community composition and functions under specific environmental conditions.

We propose to overcome this challenge by using recent systems biology approaches for the prediction of quantitative behaviors of single organisms based on genome-scale data [13, 14]. This study presents a natural extension of such approaches via their application to the modeling of microbial ecosystems and the elucidation of their quantitative features [15, 16].

Genome-scale descriptions, in this context, are provided by metabolic networks. A metabolic network summarizes the set of biochemical reactions encoded by the genome of a given organism. Two reactions are linked within a metabolic network if the substrate of one reaction is the product of the other. Such genome-scale descriptions of organisms are currently applied in systems biology for the purpose of investigating physiology [17]. In particular, for an increasing number of species, current bioinformatics protocols build genome-scale metabolic networks from genome-scale transcriptomic or metabolomic data [18].

Quantitative analyses utilize such metabolic networks as inputs for constraint-based models (CBMs) in order to infer physiological features based on a genome-scale description [17]. As a central assumption, constraint-based modeling considers the constraints defined by the set of reactions as linked within a metabolic network at steady state, and assume the corresponding model to behave optimally to achieve a given objective [13, 14]. The use of constraint-based modeling for microbial ecosystems, which involves the generation of a framework to perform data integration as well as mathematical descriptions useful for numerical simulations, seems promising [16, 19].

Several attempts have been made to model the metabolic network of microbial communities. Rodrígez *et al.* [20] proposed to use a "supra-organism" assumption, which considers reactions of all members of the community as a single entity. While such an approximation was used in recent studies (see Biggs *et al.* [21] and Perez-Garcia *et al.* [22] for a review), Kiltgord and Segré [23] previously showed that fluxes from a compartmentalized network and its de-compartmentalized counterpart (*i.e.*, supra-organism approach) are significantly different in their predicted FBA and FVA values. Furthermore, they show that fluxes using both assumptions are often not correlated. Such a distinction between both modeling results, along with the indisputable presence of compartments within ecosystems, clearly advocates for the use of compartments in the modeling. Considering so, several modelings have been proposed. However, while they all assume to consider distinct compartment for each microbial strain involved, they differ in their use of choosing the objective function. Stolyar et al. [24] first proposed a compartmentalized flux balance approach for modeling a mutualistic co-culture that requires an "ecosystem function". Such a function is usually a weighted sum of each compartment objective. Nevertheless, the relative weight of each strain objective function remains

herein at the discretion of an empirical expertise that is mostly out of reach for complex or uncharacterized microbial ecosystems.

To overcome such a weakness, more elaborated modeling approaches have been proposed. Zomorrodi and collaborators [25, 26] modeled each organism in a microbial community as a single CBM with its own objective function, nested within a global ecosystem model, thereby enabling the maximization of an ecosystem objective function. This approach still require to design an ecosystem objective function but proposes a multi-level optimization that considers both microbial strain and ecosystem objectives. Meanwhile, Khandelwal and collaborators [27] (followed by [28]) advocates for the use of the "balanced growth" concept, according to which all microorganisms grow at the same rate. Accordingly, this approach considers several compartment with no ecosystem objective function per se but rather introduces community fractions into the formulation, adding new degrees of freedom to the general optimization problem. Worth noticing, such a modeling assumption is justified for microbial communities for which biomass production is monitored and constrained in chemostat, but not necessary for open systems as observed in nature.

In this study, we propose a complementary model, to investigate the general case of microbial ecosystems. Based on Pareto optimality [29], we aim at describing all the feasible solutions considering metabolic constraints from each strain with no design of ecosystem function. Consistent with previous works, the present study considers the community as a compartmentalized system in which each organism (*i.e.*, a compartment) has (i) its own objective to optimize and (ii) shares metabolites through the environment. Contrary to above methods, our approach is based on multi-objective optimization, which allows us to consider the objective function of each organism simultaneously.

Specifically, following previous works, we implemented a multi-objective flux balance analysis method [30], henceforth known as MO-FBA, for microbial communities, which is based on an exact resolution algorithm. Additionally, we introduced a complementary multi-objective flux variability analysis (MO-FVA) method. These analyses emphasize putative metabolic behaviors that are optimal at the community level, while considering metabolic constraints for each strain. Finally, we performed complementary thermodynamics analysis [31], which enabled us to pinpoint (i) favored ecosystem responses to environmental parameters and (ii) the corresponding diversity.

For the sake of MO-FBA and MO-FVA illustration, this study models a microbial ecosystem comprising three distinct phenotypes: a primary producer, *Synecococcus spp.* (SYN), filamentous anoxygenic producers (FAP), namely *Chloroflexus spp.* and *Roseiflexus spp.*; and sulfate-reducing bacteria (SRB, composed by *Thermodesulfovibrio spp.*-like activity, [32]), as described in [33]. Results emphasize trade-offs between distinct bacterial growth rates based not only on environmental conditions and genome-scale descriptions of each strain, but also thermodynamical quantitative predictions that are consistent with experimental knowledge.

Material and methods

Metabolic networks as constraint-based models

The genomic data for a particular microorganism describes a set of genes, allowing the identification of enzymes and related reactions. Reactions produce metabolites that are used as substrates in subsequent reactions; such interplay constitutes a *"metabolic network"* whose size may vary from few tens to several hundreds of reactions [14]. Metabolic networks are modeled (Fig 1A) in order to study the physiology of the relevant microorganism. In particular, metabolic models are used to infer reaction rates, also known as fluxes, without using kinetic parameters. For this purpose, a metabolic model is formally described by its stoichiometric

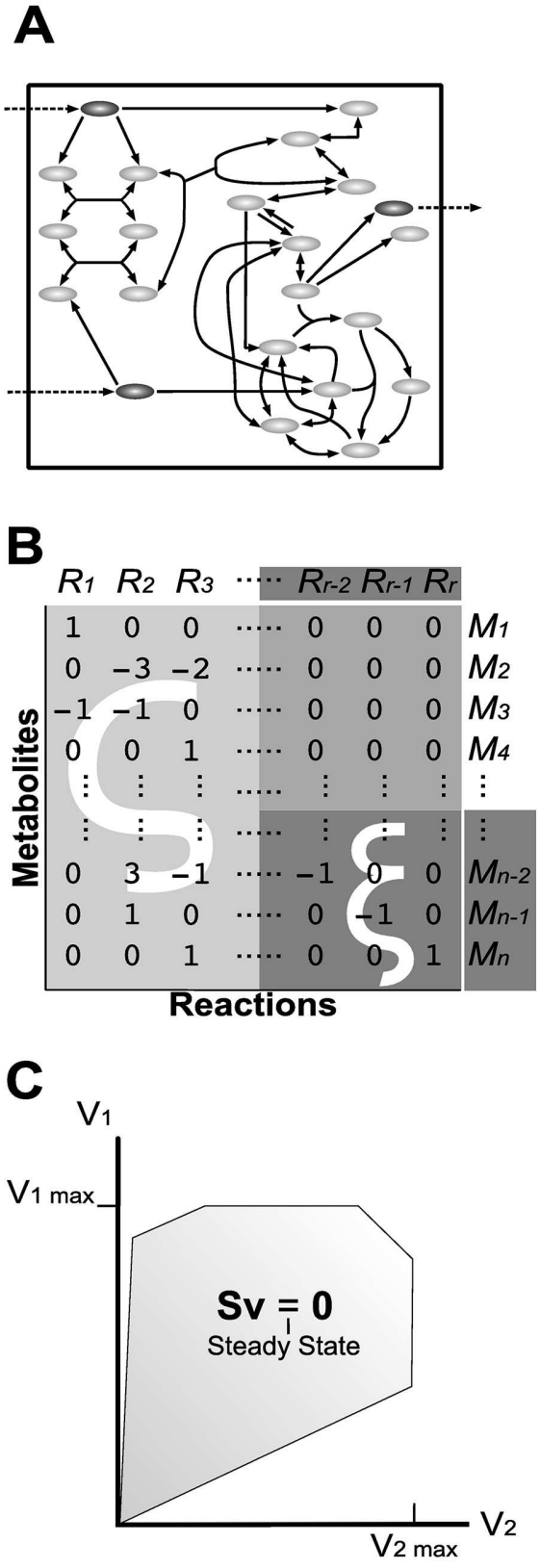

A
B
R1 R2 R3 ····· Rr-2 Rr-1 Rr
1 0 0 ····· 0 0 0 M1
0 −3 −2 ····· 0 0 0 M2
−1 −1 0 ····· 0 0 0 M3
0 0 1 ····· 0 0 0 M4
0 3 −1 ····· −1 0 0 Mn-2
0 1 0 ····· 0 −1 0 Mn-1
0 0 1 ····· 0 0 1 Mn
S
ξ
Metabolites
Reactions
C
V1
V1 max
Sv = 0
Steady State
V2
V2 max

Fig 1. Construction of a Constraint Based Model (CBM). (A) **Metabolic Network** is represented as a chart of metabolites (ellipses) trough chemical reactions (arrows); borders represent the system boundary. (B) depicts the **Stoichiometric Matrix**, in which reactions are presented as columns and metabolites as rows. Each coefficient $\mathbf{S_{ij}}$ of the matrix corresponds to the stoichiometric coefficient of metabolite M_i in reaction R_j, with reactants as negative and products positive. Exchange reactions and exchange metabolites are placed in the right and inferior section of the matrix, respectively. Therefore, submatrix ς is in the left and highlighted in light gray while submatrix ξ is highlighted in dark grey (see text). Normal gray depicts a matrix with only zeros. (C) **Flux space**, also known as "solution space", is defined by the set of restrictions of the CBM (mass balance in steady state, bounded reaction rates, etc.) and contains all possible values of **v**.

doi:10.1371/journal.pone.0171744.g001

matrix **S** (Fig 1B), where the rows correspond to the metabolites and the columns correspond to the reactions considered in the metabolic network. At steady-state conditions, the rate of formation of internal metabolites is equal to the rate of their consumption. This is expressed by the flux balance equation $\mathbf{Sv} = 0$, where $\mathbf{v} = (v_1, \ldots, v_r)$ stands for the flux vector, *i.e.*, v_j is the flux of reaction R_j for all $j = 1, \ldots, r$.

Under steady-state conditions, the continuous supply of metabolites from the media is facilitated by exchange reactions at a constant rate (dark gray eclipses and dashed lines in Fig 1A and highlighted dark gray block in Fig 1B). This matter exchange with the media allows the metabolic network to be in a non-equilibrium steady state (NESS). If metabolite exchange were not possible, then for each reaction the only possible state would be the chemical equilibrium, with all net fluxes equal to zero [31]. In the following, ς and ξ represent, respectively, internal reaction and exchange reaction submatrices (light gray and dark gray blocks in Fig 1B, respectively). Occasionally, exchange rates may be experimentally measured and incorporated into the model as equations of the form $v_i = b$ for reaction *i*. In addition, maximal and minimal flux values may be expressed as *lower* and *upper* bounds constraints, by equations of the form $l_i \leq v_i \leq u_i$, resulting in a model described as a set of constraints. Such models are termed CBMs. CBMs usually comprise more reactions than metabolites; therefore, these models are undetermined in that when a solution **v** exists, it is not unique. All feasible solutions define a *"flux space"* (Fig 1C) that may be further analyzed through several state-of-the-art approaches. For a detailed review of these methods, the reader may wish to refer to [13] and [14].

Flux balance analysis. Flux balance analysis (FBA) is one of the most widely used approaches for the identification of points of interest in the flux space [14]. Using this method, an objective function (for example, biomass production) is stated and its maximal value within the flux space is determined. In addition to the flux balance constraints, FBA utilizes flux capacity constraints that limit the fluxes of reactions. An optimal flux vector may be obtained by solving the following linear program (LP):

$$\begin{aligned} \underset{\mathbf{v} \in \mathbb{R}^n}{\text{maximize}} \quad & z = \mathbf{c}^{\mathsf{T}}\mathbf{v} \\ \text{subject to} \quad & \\ & \mathbf{Sv} = \mathbf{0} \\ & l_i \leq v_i \leq u_i \qquad i = 1, \ldots, n, \end{aligned}$$

where $\mathbf{c}^{\mathsf{T}}\mathbf{v}$ is a linear combination of fluxes that represents the objective function (*i.e.*, biomass production or growth rate). From linear programming theory, it is known that the optimal value z^* of objective function is unique; however, multiple flux distributions (*i.e.*, values of **v**) that achieve the same optimal value z^* may exist.

Flux variability analysis. The set of all optimal flux distributions, *i.e.,* those with an optimal objective value of z^*, may be investigated by using Flux Variability Analysis (FVA) to

determine the flux range of each reaction in the metabolic network [14]. Formally, FVA solves the two following LPs for each reaction R_j:

$$\begin{aligned} \underset{v_j \in \mathbb{R}}{\text{maximize / minimize}} \quad & v_j \\ \text{subject to} \quad & \\ & \mathbf{c}^{\mathrm{T}}\mathbf{v} \geq \alpha \cdot z^* \\ & \mathbf{S}\mathbf{v} = \mathbf{0} \\ & l_i \leq v_i \leq u_i, \qquad i = 1, \ldots, n \end{aligned}$$

where $\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$ represents the fraction of the optimum value with respect to the FBA objective value to be considered. FVA allows the user to infer specific properties of the fluxes involved. For example, *essential* reactions have strictly positive or negative fluxes, whereas *blocked* reactions are constrained to have a flux value equal to zero.

Both FBA and FVA are today state-of-the-art tools to explore CBMs [13]. From a computational viewpoint, several algorithms are available to solve these optimization-based approaches (see section Solving Linear Optimization Problems).

Thermodynamic constraints metabolic networks. FBA and FVA utilize constraints derived from mass conservation laws; however, it is possible to exploit thermodynamic laws to derive constraints in order to obtain further insights into the behavior of a metabolic system [31, 34, 35]. In biochemical systems, each metabolite has an associated chemical potential μ_i (expressed in J.mol^{-1}), which quantifies the potential to perform chemical work. Chemical potentials depend on metabolite concentration according to $\mu_i = \mu_i^0 + \mathrm{RT}\ ln(x_i/x_i^0)$, where x_i is the molar concentration, x_i^0 is the standard reference concentration (1 M) and μ^0 is the standard chemical potential (dependent on temperature, pressure, and ionic strength); these are usually tabulated [36, 37]. For a reaction j, the stoichiometric sum of the chemical potentials of the metabolites involved is equal to the Gibbs energy of the reaction, *i.e.*, $\Delta_r G_j = \sum_i^n \mathbf{S}_{ij}\ \mu_i$ where $\Delta_r G_j \leq 0$ for a spontaneous reaction. In the following, we note the Gibbs energy of reaction as a difference of potentials, *i.e.*, $\Delta\mu_j \doteq \Delta_r G_j$.

Under NESS conditions, the entropy balance implies that $\boldsymbol{\Delta\mu}^{\mathrm{T}}\ \mathbf{v}_\varsigma = \boldsymbol{\mu}^{\mathrm{T}}\ \mathbf{v}_\xi$, where $\mathbf{v}_\varsigma$ represents the internal portion of fluxes, $\mathbf{v}_\xi$ boundary fluxes, and $\boldsymbol{\Delta\mu}$ and $\boldsymbol{\mu}$ are vectors of components $\Delta\mu_j$ and μ_i, respectively. The term $\boldsymbol{\mu}^{\mathrm{T}}\mathbf{v}_\xi$ represents the *chemical motive force* or *cmf* of the network, which accounts for energy related to boundary fluxes [31]. This equation may be interpreted as internal fluxes being driven by the consumption of external chemical potential.

The integration of such equations into general CBMs is not straightforward, as in most of applications, concentrations x_i are not known; therefore, these must be introduced as variables. As a result of non-linear expressions, CBM formulations using these constraints are generally more complex to solve [38–40].

Solving linear optimization problems. In general, optimization problems are aim at determining $f(\mathbf{v})$ where $\mathbf{v}$ is usually required to satisfy constraints. Linear optimization problems (LPs) are a particular kind of optimization problem where both objective function and constraints may be expressed as linear functions of variables, *i.e.*, $max\ f = \mathbf{c}^{\mathrm{T}}\mathbf{v}$, $\mathbf{A}\mathbf{v} = \mathbf{b}$; where $\mathbf{v}$ is a vector of variables, $\mathbf{c}$ is a row vector of n coefficients, $\mathbf{A}$ is a matrix of n columns and m rows, and $\mathbf{b}$ a column vector of m values. The solution space of LP problems are polyhedrons that are characterized by their extreme points.

The first algorithm to solve a LP, which was proposed in 1947 by Dantizg [41], was based on the fact that if the objective function has an optimum value in the feasible region, then it reaches this value in at least one of the extreme points. The algorithm begins its search in one

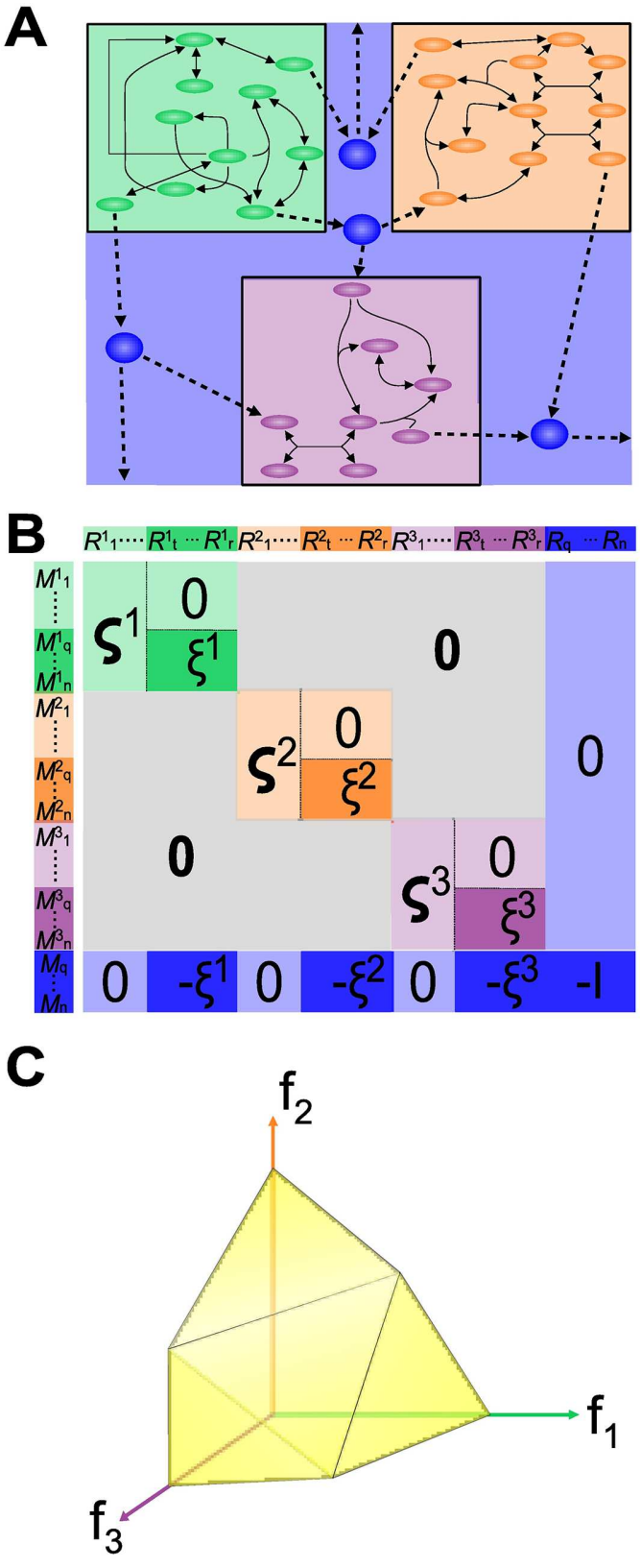

A
B
C
f2
f1
f3

Fig 2. Illustration of microbial ecosystem CBM. For the sake of illustration, an ecosystem may be considered to comprise three microbial strains. (A) According to the metabolic model, each microorganism is considered a separate compartment, depicted here in green, orange, and purple. Metabolic networks are linked via an additional compartment, termed the "pool" (blue)), which sums up all external metabolites exchanged between organisms and the environment. (B) depicts the Stoichiometric Matrix $\mathbf{S}^{\sigma}$, where each compartment is colored accordingly, with their corresponding ς and ξ submatrices. (C) Pareto front. When performing an FBA for multiple organisms, a set of points known as the Pareto front (in yellow) is obtained. Objective functions $\mathbf{f_1}$, $\mathbf{f_2}$ and $\mathbf{f_3}$ define the "objective space".

doi:10.1371/journal.pone.0171744.g002

vertex of the feasible region and then starts visiting adjoint vertexes until the objective function value cannot be improved. Currently, several solvers such as GUROBI [42] or GLPK are capable of solving LPs and other types of single objective problems (SOPs) efficiently.

From single microorganisms to microbial ecosystems

In order to model a microbial community, each strain is considered a single compartment [19, 25, 27] that shares metabolites with other strains (see Fig 2A). As the stoichiometric matrix of a single organism, the structure of the ecosystem is described by a stoichiometric matrix $\mathbf{S}^{\sigma}$, which is formed by the stoichiometric matrices of each single organism. Accordingly, for a community of k microorganisms, k metabolic models must be considered and represented by their corresponding stoichiometric matrices: $\mathbf{S}^{l}$, $l = 1, \ldots, k$.

As shown in Fig 2B, matrices $\mathbf{S}^{1}$ to $\mathbf{S}^{k}$ are used to construct a diagonal block matrix. Each block is linked to a *pool compartment*, that mirrors exchange fluxes between each organism and the environment ($-\xi^{l}$, for $l = 1, \ldots, k$ in Fig 2B). A set of exchange reactions R_q to R_n for metabolites M_q to M_n between the Pool and the external environment, is additionally set (bottom right in Fig 2B). Finally, as for single organisms, a steady state hypothesis restricts the solution set by adding a constraint $\mathbf{S}^{\sigma}\,\mathbf{v} = 0$. Together with flux bound constraints l_i and u_i, these constraints describe a solution flux space, as depicted in Fig 1C.

Multi objective flux balance analysis of a microbial ecosystem. Each compartment above corresponds to an organism with a specific objective function $\mathbf{c}_k$. Accordingly, the following multi-objective optimization problem, for analyzing flux balance conditions (MO-FBA), may be defined:

$$\underset{\mathbf{v} \in \mathbb{R}^{\bar{n}}}{\text{maximize}} \quad \begin{pmatrix} f_1 \\ \ldots \\ f_k \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1^{\mathrm{T}}\mathbf{v} \\ \ldots \\ \mathbf{c}_k^{\mathrm{T}}\mathbf{v} \end{pmatrix}$$

$$\text{subject to} \quad \mathbf{S}^{\sigma}\mathbf{v} = \mathbf{0}$$

$$l_i \leq v_i \leq u_i \quad i = 1, \ldots, \bar{n}$$

where $(f_1, \ldots, f_k)^{\mathrm{T}}$ are the objective functions of the k organisms and $\bar{n}$ is the total number of reactions (*i.e.*, the sum of reactions of each organism and exchange reactions from the pool compartment). The general class of MO-FBA problems is referred to as the *multi objective problems* (MOP) [29, 43]. Contrary to single objective problems, solution of MOPs is a set of vectors instead of a single value, producing a Pareto front (see section Solving Multi Objective Optimization Problems), defined in the objective space (Fig 2C). In our present formulation, all constraints and objective functions are linear, thereby resulting in a particular type of MOP known as the multi-objective linear problem (MOLP).

Interpretation of MO-FBA can be done in terms of growth rates and resources used to produce such growth. Indeed, if one of the members of the ecosystem decreases its growth rate, more resources are available for other members. According to their particular physiologies, they can use these new available resources to increase their own biomass. A guideline containing three ideal cases for two guilds is provided in S1 File.

Flux variability analysis of a microbial ecosystem. Given a particular point $\mathbf{f}^*$ of the Pareto Front, the multiple optimal flux solutions that achieve the optimal objective values, as given by the Pareto optima $\mathbf{f}^*$, must be explored. To this end, we propose the use of the multi-objective FVA (MO-FVA) for multiple organisms, which may be considered a straightforward extension of FVA (see Flux Variability Analysis). Indeed, given a reaction R_j with $j = 1, \ldots, \bar{n}$, the range of the flux v_j may be determined by solving the following LPs:

$$\begin{aligned} &\underset{v_j \in \mathbb{R}}{\text{maximize / minimize}} \quad v_j \\ &\text{subject to} \\ &\qquad \mathbf{C}^{\intercal}\mathbf{v} \geq \alpha \cdot \mathbf{f}^* \\ &\qquad \mathbf{S}^{\sigma}\mathbf{v} = \mathbf{0} \\ &\qquad l_i \leq v_i \leq u_i \quad i = 1, \ldots, \bar{n} \end{aligned}$$

where $\mathbf{C}$ is the matrix such as the column j corresponds to objective function $\mathbf{c}_j$, *i.e*, $\mathbf{C}$ is column defined as $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_k]$. $\alpha \in \mathbb{R}, 0 \leq \alpha \leq 1$ is the fraction of the optima considered.

Thermodynamics analysis in the context of a microbial ecosystem. Biological systems are hypothesized to favor thermodynamic states where entropy production is maximal [44, 45]. To take into account this hypothesis, given a particular point $\mathbf{f}^*$ of the front, we propose the following: First, a MO-FVA must be applied to determine R_j for each reaction, with $j = 1, \ldots, \bar{n}$ and the range $[a_j, b_j]$ of the flux v_j near the Pareto optima $\mathbf{f}^*$. Next, the following optimization problem must be considered:

$$\begin{aligned} &\underset{i \in \xi}{\text{maximize}} \qquad cmf = \sum \mu_i v_i \\ &\text{subject to} \\ &\qquad\qquad a_i \leq v_i \leq b_i, \qquad\qquad i \in \xi, \\ &\qquad \mu_i^0 - \mathrm{dg}_i \leq \mu_i \leq \mu_i^0 + \mathrm{dg}_i, \end{aligned}$$

where ξ is the set of exchange reactions and $\mathrm{dg}_i = RTln(x_i/x_i^0)$. As *cmf* is non-linear, optimization algorithms based on heuristics must be used in order to obtain a numerical solution to this problem (see Computational Procedures).

Solving multi objective optimization problems. In 1906, Vilfredo Pareto in his *Manuale di Economia Politica*, stated that, while (economic) optima have not been achieved, it is possible to increase the objective of an agent (*i.e.*, welfare) without decreasing that of another [46]. In the following, a formal definition of Pareto optima and efficient solutions is given [43] and approaches to solutions are discussed.

Let $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^n$ and $\boldsymbol{\mathcal{Y}} \subseteq \mathbb{R}^p$ represent the flux space and objective space, respectively, where $\boldsymbol{\mathcal{X}}$ is defined by the set of restrictions and $\boldsymbol{\mathcal{Y}} := \{\mathbf{y} \mid \mathbf{y} = \mathbf{f}(\mathbf{x}), \mathbf{x} \in \boldsymbol{\mathcal{X}}\}$, with $\mathbf{f} = (f_1(\mathbf{x}), \ldots, f_p(\mathbf{x}))^{\intercal}$ denoting the objective functions. If both $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{Y}}$ are constructed using linear restrictions and linear objective functions, the MOP represents a MOLP.

A point $\mathbf{y} \in \boldsymbol{\mathcal{Y}}$ is a **Pareto optimum** if there is no $\mathbf{y}^* \in \boldsymbol{\mathcal{Y}}$ such as $y_j^* \geq y_j, j = 1, \ldots, p$ and $\mathbf{y} \neq \mathbf{y}^*$. Similarly, $\mathbf{y}^{\mathbf{w}}$ is a **weak Pareto optimum** point if there is no $\mathbf{y}^*$ such as $y_j^* > y_j^w, j = 1, \ldots, p$. A point $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ is an **efficient** solution if there is not a $\mathbf{x}^* \in \boldsymbol{\mathcal{X}}$ such that

$\mathbf{f}(\mathbf{x}^*) \geq \mathbf{f}(\mathbf{x})$. A $\mathbf{x}^w \in \boldsymbol{\mathcal{X}}$ is a **weak efficient** solution if there is no $\mathbf{x}^* \in \boldsymbol{\mathcal{X}}$ such as $\mathbf{f}(\mathbf{x}^*) > \mathbf{f}(\mathbf{x}^w)$. Therefore, a (weak) Pareto optimum is the image of a (weak) efficient solution. Note that all efficient solutions are also weakly efficient solutions but no vice-versa. The collection of Pareto optimal points is termed **Pareto Front**.

Approaches for solving MOPs have been reviewed, for example, by [43] and [47]. Traditional approaches makes use of "scalarization techniques", that involve the transformation of the MOP into a SOP by using a real-valued scalar function of the objective functions. Solution approaches using scalarization techniques aim to find the set of (weak) efficient solutions $\mathbf{x}^* \in \boldsymbol{\mathcal{X}}$.

The most well known approach is the "weighted sum approach", wherein the weighted sum of the objective functions is optimized, *i.e.*, $\max \sum \lambda_k f_k(\mathbf{x})$, where $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ and $\boldsymbol{\lambda} \in \mathbb{R}^p$ is a given weight vector with components $\lambda_k \geq 0$ and at least one $\lambda_k > 0$. If $\mathbf{x}^*$ is a solution of this SOP then $\mathbf{x}^*$ is an efficient solution of the MOP. Furthermore, if the MOP is convex, the inverse is also true.

Another commonly used approach is the "ϵ-constraint method", where only one objective function is retained as the objective and the remaining objective functions are used to introduce new constraints. Then, the j-th ϵ-constraint problem is as follows: $\max f_j(\mathbf{x})$, subject to $f_i(\mathbf{x}) \geq \epsilon_i$, $i \neq j$ and $\mathbf{x} \in \boldsymbol{\mathcal{X}}$. If $\mathbf{x}^*$ is a solution of this SOP, then $\mathbf{x}^*$ is a weak efficient solution of the MOP.

Not all approaches rely on scalarization: for MOLPs, a set of algorithms describing the shape of the image of efficient points, $\boldsymbol{\mathcal{Y}}_E := \{\mathbf{C}\mathbf{x} \mid \mathbf{x} \text{ is efficient}\}$, referred to as "outer approximation" or "Benson type" algorithms, have been described [48–51]. Generally speaking, these type of algorithms calculate $\boldsymbol{\mathcal{Y}}$ and identify their vertices, which correspond to Pareto optimal points; additionally, despite their names, these algorithms provide exact solutions. BENSOLVE [52], a solver based on these approaches, computes a set of directions and points describing the image of the efficient points.

Existing CBM approaches for communities. The various approaches to studying microbial communities have been recently reviewed by Biggs *et al.* [21] and Perez-Garcia *et al.* [22]. Among the methods reviewed, OptCom most closely resembles the approach presented here, in that each member of the community is considered to maximize its own biomass. OptCom is based on bi-level optimization, where an "outer" maximization problem represents the whole community and each member of the community is represented by a "inner" optimization problem. Inner optimization problems are solved using the primal-dual theorem, which transforms the whole bi-level formulation into a non-convex single-objective form [25]. A second approach that combines compartments and FBA, known as community flux balanced analysis, advocates the application of a *"balanced growth"* hypothesis, wherein each compartment grows at the same rate. Furthermore, this approach considers the biomass fraction of each member of the community. In general, the approach is non-linear, although it may be made linear by fixing biomass fractions and solving the corresponding FBA. Then, optimal solutions for various combinations of biomass fractions may be explored [27]. For illustration purposes, the application of our approach to the analysis of a microbial ecosystem is discussed below.

Case study: Hot spring mat

In order to illustrate the application of the present approach, we modeled the microbial ecosystem of hot spring microbial mats [33]. Briefly, this ecosystem is composed of three *guilds*, representing three commonly found phenotypes: *Synechococcus spp.* (SYN), *Chloroflexus spp.* and *Roseiflexus spp.* (FAP) sulfate-reducing bacteria (SRB). SYN is a primary producer that fixes

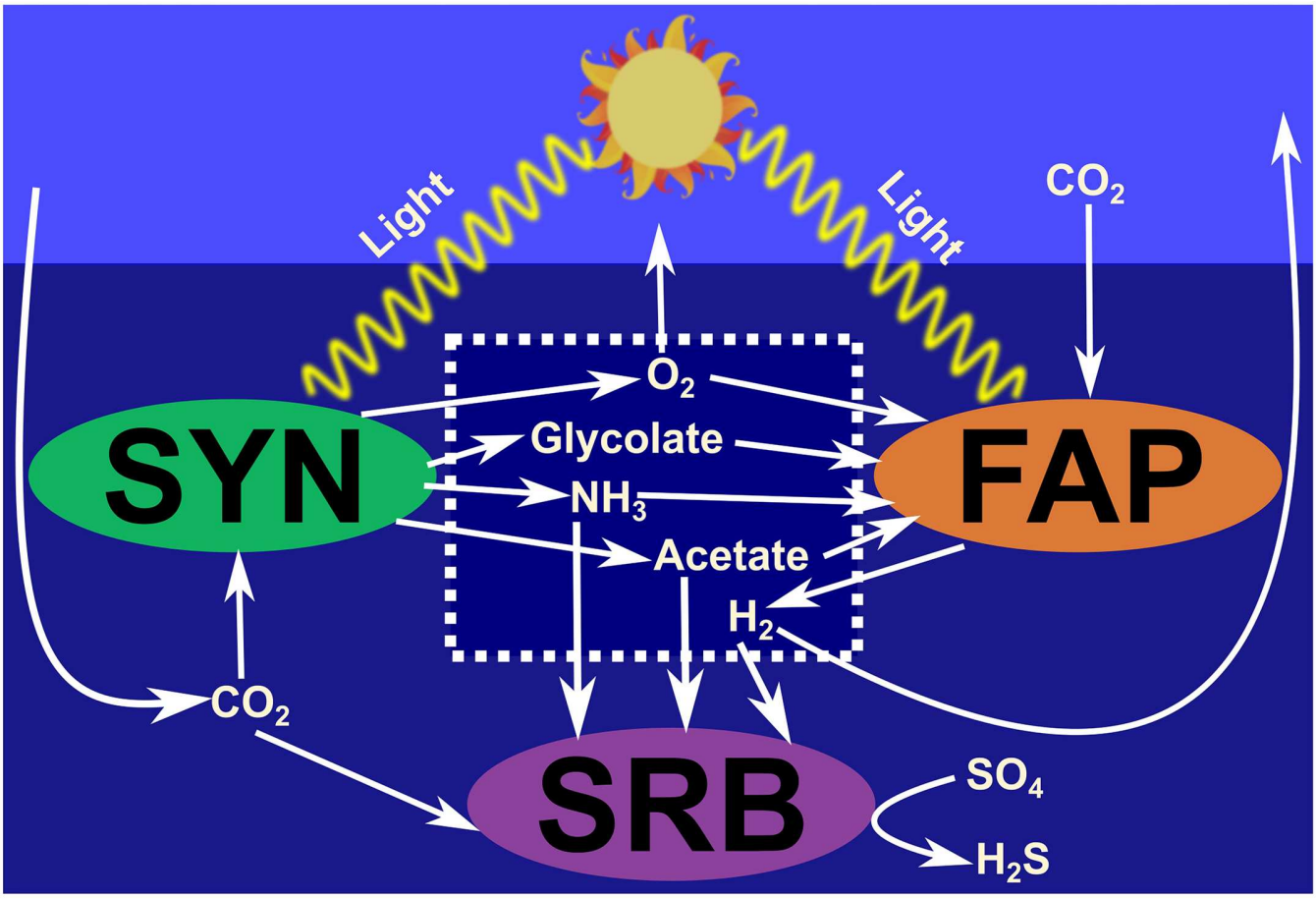


Fig 3. Day Model of the Hot Spring Mat Community. The model comprises three guilds of microorganisms of the SYN, FAP, and SRB phenotypes. Organics acids produced by SYN may be utilized by FAP and SRB. FAP is capable of fixing carbon by anoxygenic photosynthesis. Under anoxygenic fermentation conditions, FAP is additionally capable of producing hydrogen, which, in turn, may be used by SRB.

doi:10.1371/journal.pone.0171744.g003

carbon and nitrogen for further utilization by other strains. The use of these guilds allows simplification of the ecological diversity while capturing essential metabolite-exchange relationships. Under light conditions, the major fate of nutrients involves assimilation into cells [53]; therefore, most of the overall system growth occurs during the daytime. As growth rates are related to biovolumes, predictions may be compared with relative abundance data. Therefore, we will focus on the daytime model as described in [33] (Fig 3), assuming a simplified nighttime behavior, as described below.

Using the available compartment model of this system, as described in [33], we performed a manual curation (*i.e.*, balancing equations and including intermediate reactions) using METACYC [54]. Model equivalent reactions in [33] are provided in S2 File. Nitrogen fixation has been shown to take place at night and in the early morning [55, 56]; therefore, a nitrate assimilation mechanism for SYN was included and considered as functional. Finally, biomass coefficients of each guild were scaled to match 1 (h^{-1}) as maximal growth rate [57].

Glycolate is produced by the use of O_2 instead of CO_2 by the Rubisco enzyme; the flux ratio between the use of O_2 and CO_2 varies between 0.03 and 0.07. This restriction was included

linearly in the model by fixing a ratio of 0.03 between SYN reactions RXN-961 and RIBULOSE-BISPHOSPHATE-CARBOXYLASE-RXN during all calculations, under the hypothesis that the system is in anaerobic state.

Excess photosynthate producing during the day is stored as polyglucose (PG) by SYN. PG is fermented at night, producing several organic acids that accumulate in the media and are integrated as biomass mostly under light conditions [53, 58]. In order to capture this behavior in the daytime model, PG was not allowed to accumulate; therefore, the excess photosynthesis activity is redirected through acetate production. Accordingly, in our model, acetate is interpreted as equivalent to several forms of reduced carbon.

For each of the exchanged metabolites, standard Gibbs energies for biological conditions were obtained from [37], using calculations from [36]. Values used are found in S2 File. For the pseudo-compound *hv* (representing photons), a standard chemical potential was estimated based on glucose synthesis from CO_2: $6CO_2+6H_2O \xrightarrow{48\ hv} C_6H_{12}O_6$. The assumption that this reaction approaches equilibrium at standard biological conditions (*i.e.*, $\Delta\mu = 0$) implies that $\mu_{hv} = 68.6$ kJ.mol^{-1} (S2 File). The metabolite concentration was allowed to vary between 10^3 and 10^{-3} M, and therefore chemical potential equals $\mu_i = \mu_i^0 \pm$ dg, where dg = $RTln(10^3) \approx 20$ (kJ.mol^{-1}) for T = 75°Celsius. For water and *hv*, concentrations were considered as fixed at 1 M, implying $dg_{H_2O} = dg_{hv} = 0$.

Computational procedures

For each guild, a metabolic model was built in MATLAB and an ecosystem stoichiometric matrix $\mathbf{S}^\sigma$ was constructed, as described above. MO-FBA was carried out using BENSOLVE [52]. In order to analyze nitrogen and carbon fluxes through MO-FBA results, a MO-FVA was performed using GUROBI [42] through Python interface over a mesh of 5 151 equally distributed points in the Pareto surface at 90% fraction of optimum. Then, we subdivided the Pareto surface into 225 similar regions; for each of these regions, we calculated their maximum (as well as their minimum) as the average of MO-FVA maxima of mesh points contained (this procedure was repeated for the minima). Thermodynamics calculations were performed over the same mesh as the MO-FVA using a truncated Newton conjugate algorithm [59] contained in scipy optimization module. Heatmaps and surface illustrations were generated using matplotlib [60] with *ad-hoc* scripts.

From methods discussed in Biggs *et al.* [21] and Perez-Garcia *et al.* [22], OptCom [25] was chosen for comparison, as this method resembles the approach applied to the present work. We applied OptCom and Descriptive OptCom to the hot spring mat model, as follows: first, 11 points were calculated using OptCom, as described by [25], each with a different upper boundary value for SYN biomass; these values ranged from 1.0 to 0.0 with a step of 0.1 (i.e. 1.0, 0.9, 0.8, . . ., 0.0). Second, Descriptive OptCom was applied three times using SYN to FAP ratios of 1.5, 2.5, and 3.5, respectively. All programs were written in GAMS language and solved using BARON [61] through the NEOS Server [62–64].

All scripts are available in https://gitlab.univ-nantes.fr/mbudinich/MultiObjective-FBA-FVA

Results

Biomass distribution as relative microbial strain abundance

SYN, SRB, and FAP growth rates are represented in a 3-dimensional space, in each axis, respectively, in Fig 4A. MO-FBA solutions are described as a Pareto front, representing a surface with five extreme points of biomass growth: (1, 0, 0), (0, 1, 0), (0, 0, 1); the points

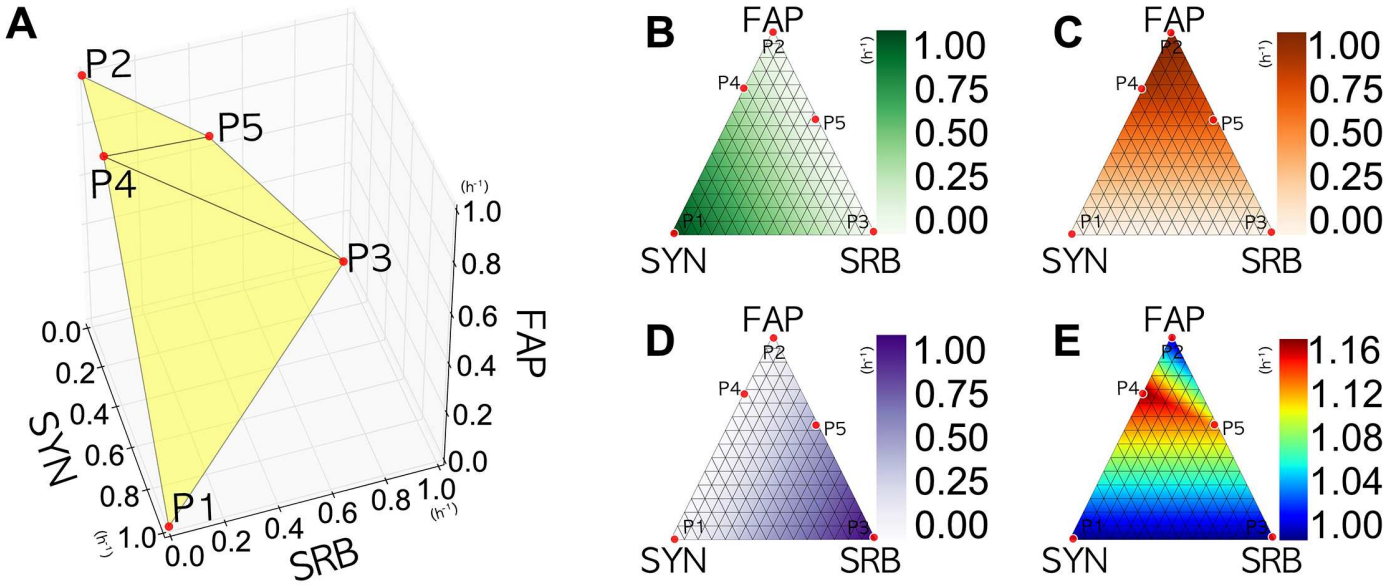


Fig 4. 3D and 2D Projections of Pareto Front. (A) shows a 3D Pareto front, in yellow, describing the maximal growth rates of SYN, FAP, and SRB (in terms of units per hour, h^{-1}), when considered as a system. It is evident that a decrease in the growth rate of one organism results in an increase in that of the other two, but not necessarily in equal proportions (see S1 Video for an animated view). The sum of the growth rates of all the guilds in P4 and P5 was 1.16 (h^{-1}) and 1.11 (h^{-1}), respectively. In (B), (C), (D), and (E), the Pareto front was projected onto the triangular surface formed by P1, P2, and P3. (B), (C), and (D) shows the respective growth rates for SYN, FAP, and SRB, respectively. (E) shows the sum of the three growth rates, which represent the total biomass of the ecosystem.

doi:10.1371/journal.pone.0171744.g004

corresponding to the maximal growth rates of each guild, and points (0.27, 0.00, 0.89) and (0.00, 0.46, 0.65). In the following, these points are designated P1, P2, P3, P4, and P5, respectively. For clarity, this Pareto front is then projected in a two-dimensional space. Therefore, over a triangular surface defined by P1, P2, and P3, heatmaps were produced using the values for the growth rate of SYN, FAP, SRB, as well as their sum, to depict the overall microbial abundance (Fig 4B–4E, respectively). Each vertex of the triangle represents the maximal growth rate of a guild, while its opposing side represents a zero growth rate for that guild.

The results show that when each guild grows at its maximal rate, no biomass is produced by the other guilds. The sum of the growth rates is always minimal in vertices (blue areas in Fig 4E). As the growth rates may be directly related to biovolumes [33], red to yellow areas in Fig 4E represent regions where most of the total biomass of the ecosystem is present. Notably, these regions correspond to guilds growing at sub-optimally rates.

Nitrogen and carbon fluxes between microbial guilds

Multi-objective FVA was performed in the P4 and P5 regions to explore NH_3 import and export fluxes between guilds (Fig 5A, upper and lower panel, respectively). Notably, the growth rate of each strain was found to be related to the use of ammonia; the SYN guild re-oxidized ferredoxins, which were reduced in the photosynthetic reactions, via nitrate assimilation reactions, thereby promoting permanent ammonia production. When growing sub-optimally, NH_3 that is not used to build biomass is excreted. This point is emphasized in Fig 5A, where both maximal and minimal reaction rates are strictly positive for SYN, resulting in an export to the pool.

Nitrogen uptake by FAP and SRB occurs solely from ammonia that is available in the pool compartment; therefore, these strains compete for its intake. When SRB is not growing

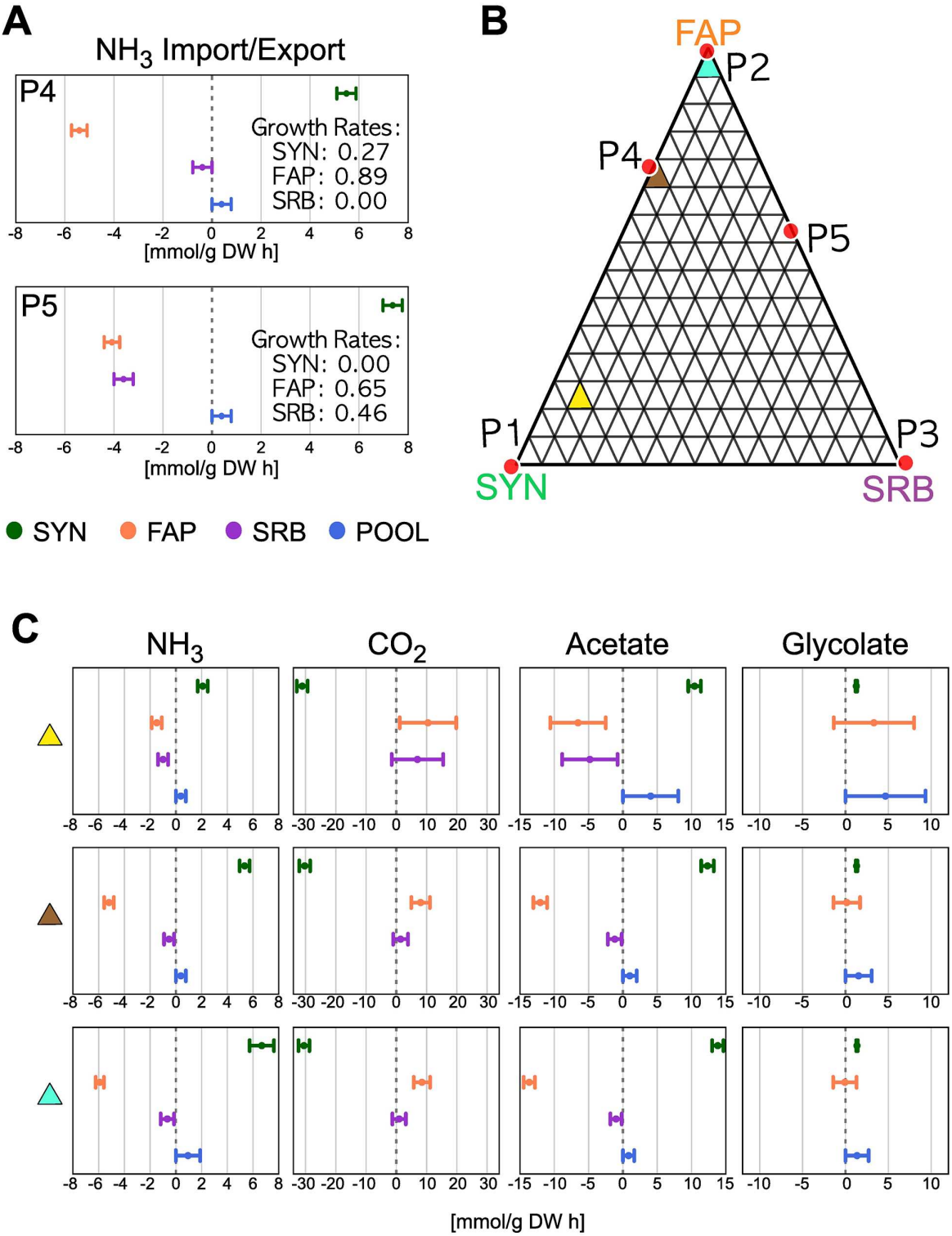


Fig 5. Multi Objective FVA. (A) shows NH_3 maximal and minimal fluxes for SYN, FAP, SRB, and pool compartments (green, yellow, purple, and blue respectively) for extreme points P4 and P5. The export of NH_3 by SYN is correlated with a drop in their growth rate; similarly, increases in NH_3 intake are correlated with increases in the growth rates of FAP and SRB. (B) Three sections selected for the illustration of MO-FVA; (C) Mean values of the minimal and maximal fluxes over selected sections of NH_3, CO_2, acetate, and glycolate (columns) for each section (rows).

doi:10.1371/journal.pone.0171744.g005

(superior panel in Fig 5A), excess of NH_3 is taken up mainly by FAP (both minima and maxima are negative, implying an intake from the pool). Small amounts that are not taken up by FAP may be either taken up by SRB (maximal rate value is null and minimal rate negative, which depicts a possible import) or excreted to the external environment (pool maximal rate value is positive and minimal rate value is null, which depicts a possible export to the media). When SRB is growing (inferior panel of Fig 5A), the uptake rate of ammonia by SRB and FAP is similar, with no export to the external media.

In order to analyze the relationships between the growth rate of each strain and nitrogen- or carbon-related fluxes, we performed a MO-FVA as described in Computational Procedures, focusing on exchange reactions. For the purpose of illustration, we highlighted three sections from 225 calculated, as shown in Fig 5B. These regions were chosen to depict the theoretical interplay between SYN and FAP when the growth rate of SRB is low [65]. Flux variability of exchange fluxes for these regions is shown in Fig 5C (see S1 Fig for an alternative representation and S2 to S5 Figs for a complete MO-FVA for ammonia, acetate, carbon dioxide and glycolate fluxes).

For NH_3 exchange reactions, high growth rates of SYN are related to lower levels of ammonia export, which represents a limiting factor for FAP and SRB growth rates. This results in the two strains competing for its use (S2 Fig). Fig 5C shows that most of the ammonia produced by SYN is captured by FAP, while a small proportion is taken up by SRB. Ammonia that is not captured is released into the pool.

SYN consumes approximately the same amount of CO_2 under all relative abundance conditions (see second column in Fig 5C and S4 Fig), indicating that carbon compounds are involved in reactions that serve functions other than biomass synthesis. Acetate intake by FAP is less restrained at low growth rates of SYN than at high growth rates (see Fig 5C and S3 Fig).

The present results additionally emphasize that FAP and SRB produce relatively small amounts of CO_2 at low growth rates. However, when the growth rate of FAP increases, the maximal excretion of CO_2 reduces, whereas its minimal excretion increases; these data indicate the theoretical efficiency of carbon management, as experimentally reported by [53]. Glycolate metabolism by FAP appears to be reversible as its minimal flux is negative (*i.e.*, intake) while its maximal flux is positive (*i.e.*, excretion), implying that intake or excretion by FAP is related to the relative abundance of other strains (see Fig 5C and S5 Fig for details).

Chemical potentials drive community growth rates

As discussed previously, the direct integration of thermodynamic constraints into MO-FBA and MO-FVA formulations is complex. Instead, we used the thermodynamic optimization problem stated in as a post-treatment analysis. Considering fluxes as computed by MO-FVA in 5 151 points of Pareto front (as a result of which growth rates are also determined), we estimated the corresponding maximal *cmf* for each point (Fig 6A).

Results show that higher *cmf* is associated with SYN growing at its optimal rate. Lower *cmf* rates are related to a higher growth rate of SRB, whereas the impact of the growth rate of FAP on the value of *cmf* appears to be lower than that of SRB.

Given that all surface showed positive values, all regions are feasible from a thermodynamic viewpoint. Under the hypothesis that a biological system prefers configurations in which entropy production is maximal, it is expected that an ecosystem would favor growth rates with higher *cmf* (redder areas in Fig 6A), predicting higher SYN growth rates. This prediction is consistent with *in vivo* field measurements of SYN: FAP relative abundance ratios in the range of 1.5 and 3.5, with a low presence of SRB [33, 65], as shown in Fig 6B.

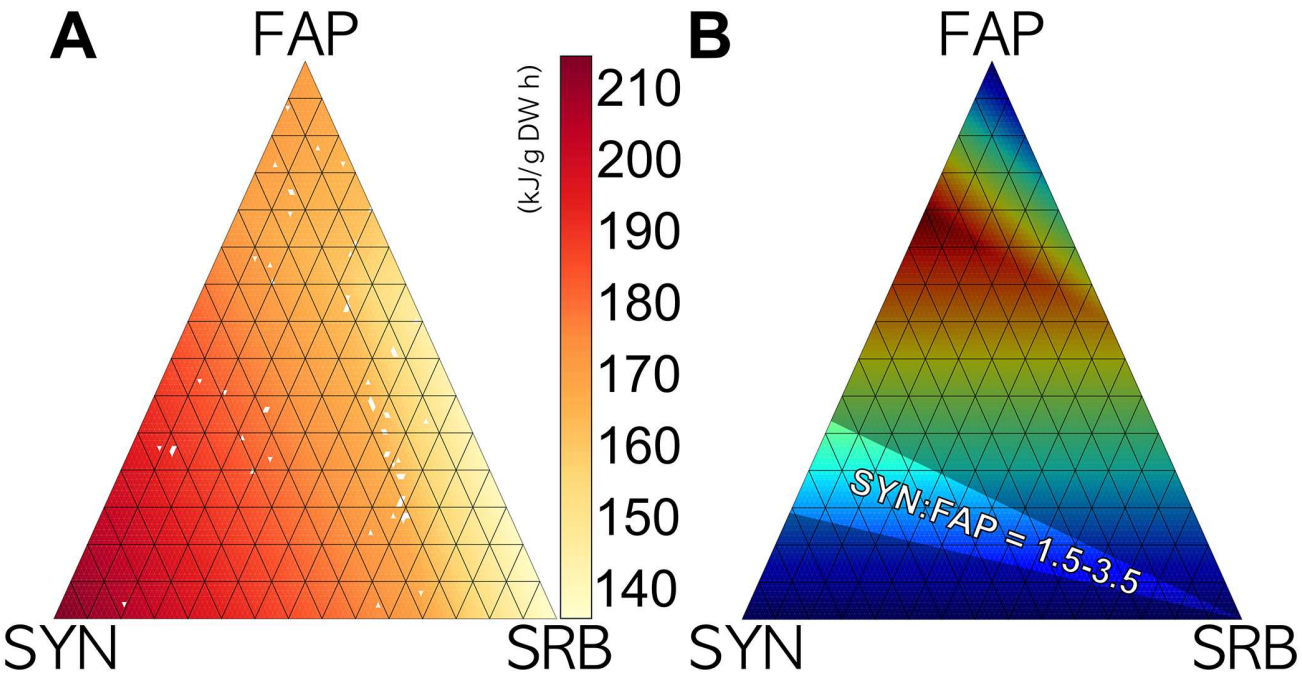


Fig 6. Thermodynamics in the Pareto front. (A) Description of the chemical motive force (kJ.gr^{-1}.DW^{-1}.h^{-1}) for each point of the Pareto front; red regions indicate thermodynamically favored growth rates, while the points where the solver does not reach the optimal criteria are shown in white. The obtained surface appears smooth, without sudden changes in neighboring values. (B) Description of the overall community biomass distribution based on the growth rate of each strain, with a particular emphasis on regions supported by experimental measurements showing a SYN: FAP ratio of between 1.5 and 3.5.

doi:10.1371/journal.pone.0171744.g006

Comparison with previous approaches

We compared growth rates and flux predictions of MO-FBA and MO-FVA with those obtained by a comparable approach (OptCom [25]), as described in Computational Procedures. Predictions obtained were mapped as points in the Pareto front (S6 Fig). Values of growth rates, as well as their corresponding flux values for NH_3, acetate, glycogen, and CO_2, are described in S2 File. As expected, all points calculated using the OptCom approach were included in the Pareto front calculated by MO-FBA (S6 Fig). Furthermore, all flux predictions for NH_3, acetate, glycogen, and CO_2 fall into the range predicted by MO-FVA. Without constraining SYN biomass (point O1), OptCom does not reach the maximal biomass optimum. However, when SYN biomass is increasingly constrained (points O2 to O11), the total biomass increases. This suggests the existence of local optima in the OptCom general formulation for this model.

The composition of a community that function in a constant environment can be also assessed using the approaches proposed in [27] and [28]. Here, we focus on modeling the composition of a community in a changing medium where the considered organisms could grow not necessarily with the same growth rate

Discussion

As reported in previous studies, in particular [25], we extended state-of-the-art systems biology constraint-based approaches to the modeling of microbial ecosystems, by considering a multi-objective optimization framework. Within the ecosystem, each microorganism, with its own

objective function, represents a building block that interacts with others via the exchange metabolites. Furthermore, the genomic knowledge of each microorganism is integrated as a set of metabolic constraints. The main advantage is represented by the capture of trade-offs on objectives and metabolite exchange between members of the ecosystem. While previous works report topological analyses that focus on pathways that promote cross-feeding between strains (see [66, 67] for example), this study quantifies fluxes through these pathways as well as their effect in objective functions, thereby representing a major step towards automatically producing trait-based models. Through the application of MO-FBA, we emphasize a full description of the Pareto front that captures trade-offs in the optimal values of the objective function of each microorganism. Additionally, we introduced MO-FVA as a tool for the analysis of exchange fluxes between members of the community. These fluxes help to characterize the optimal behavior of microorganisms, providing insights into the theoretical relative abundances (*i.e.*, a proxy for microbial diversity) and corresponding nutrients usage, that are based on *omics* descriptions.

Unlike previous works that consider multiple objectives, our approach does not rely either on assumptions about ecosystem behaviors, such as maximization of the total ecosystem biomass, ([25, 26]) nor on the balanced growth ([27, 28]) of microbial strains involved. Instead, we propose to describe all optimal solutions in the sense of Pareto in the objective space. This approach provides several advantages: firstly, it includes any solution for a system objective function expressed as a weighted sum of each compartment objective function (see [43] and section Solving Multi Objective Optimization Problems). Therefore, it comprises all solutions proposed by OptCom as system objectives for microbial communities [25]. Secondly, no additional complementary restrictions are required to focus on given solutions, i.e., imposing an equal growth rate for all members, as proposed by Kandelwal et al. [27]. This restriction remains valid for controlled microbial ecosystems. Third, the set of constraints remains linear, which allows a description of the Pareto front for realistic ecosystems. In [25] and [26], formulations are, in general, non-convex; in [27], the stated general optimization problem is non-linear. However, in order to solve MOLPs, a series of LPs must be solved for which exact algorithms are fast, thereby reducing computational complexity. Note herein that the last two points are mandatory to model natural ecosystems that are by definition composed of a large number of microbial strains and mostly unconstrained.

For illustration purposes, we applied MO-FBA to the daytime part of the diurnal cycle of the microbial hot spring mat system [33]. As most biomass fixation occurs during the day phase [53], we assumed that daytime growth rates dominate overall ecosystem rates. Results show that the maximal total biomass growth rate is achieved when each guild grows at a rate below its theoretical maximum, which may, based on genomic knowledge, be interpreted as an altruistic behavior. Mechanistically, when guilds make resources available to others, they lower their objective value by a certain proportion, based on metabolic pathways used to synthesize those resources and their biomass function. Conversely, the use of new available resources increases the value of the objective functions of the other guilds. Therefore, the growth rate of the global maximal ecosystem, which was designated P4 in our case study, should correspond to the optimal resource allocation scenario from the ecosystem viewpoint. P4 also corresponds to the optimal solution to maximal ecosystem biomass [25].

MO-FVA results show that nitrogen flux is correlated to growth rates, and that the three guilds compete for their usage. In contrast, CO_2 consumption and glycolyte and acetate production by SYN do not seem to be correlated with its growth rate, indicating that these processes are not carbon-limited. Reduced carbon, represented by acetate, appears as being the main carbon flux in the system for FAP and SRB, and becomes a limiting nutrient for FAP at

high growth rates. This result is consistent with those of [53] and [58], in which a high proportion of reduced carbon was shown to be assimilated by FAP.

By coupling MO-FVA results with chemical potentials, we were able to analyze thermodynamic constraints and study favored conditions of the Pareto front by comparing their respective maxima *cmf*. We observed that the SYN: FAP ratio, predicted using this criteria, is closer to the 1.5 to 3.5 value observed in field measurements. Thermodynamic considerations underline relative strain growth rates, or microbial diversities, that are more favorable from an energetic viewpoint, which indicates that an ecosystem behaves according to two different objectives: maximal biomass production and maximization of *cmf*, corroborating previous systems biology studies that advocate the use of distinct concurrent objectives to predict *Escherichia coli* metabolic behaviors [68]. In both cases, observations were possible by general investigation of the Pareto front.

Nevertheless, further refinement of the thermodynamic calculations is warranted. In particular, the calculation of *cmf* does not consider biomass concentration; this may be overcome by considering community fractions as proposed in [27] and [28]. Furthermore, in the current model, biomass generation does not affect the overall ecosystem entropy; however, on an intuitive basis, a larger amount of biomass should increase an entropy term, in terms of Gibbs energy, as a result of mass dispersion [69], thereby affecting *cmf* evaluation. These considerations are out of the scope of the present work; however, they but raise interesting perspectives.

Despite the above limitations, we consider the present form of the modeling approach as fruitful guidance to gain qualitative as well as quantitative data for the metabolic interplay between various species in an ecosystem. This method paves the way for improved contextualization of other -omics datasets in microbial ecology by providing a mechanistic description of species co-occurrence *via* analysis of their metabolic interactions.

Supporting information

S1 File. Guidelines for interpreting MO-FBA results.
(PDF)

S2 File. Metabolic Model of Hot Spring Community. A Stoichiometric Matrix of each guild used, along with thermodynamic data considered.
(XLSX)

S1 Video. Animated 3D version of Pareto front.
(MP4)

S1 Fig. Alternative MO-FVA illustration of Fig 5. The convention used is the same for S2–S5 Figs.
(EPS)

S2 Fig. MO-FVA for NH_3 exchange fluxes between SYN, FAP, and SRB.
(PNG)

S3 Fig. MO-FVA for acetate exchange fluxes between SYN, FAP, and SRB.
(PNG)

S4 Fig. MO-FVA for CO_2 exchange fluxes between SYN, FAP, and SRB.
(PNG)

S5 Fig. MO-FVA for glycolate exchange fluxes between SYN, FAP, and SRB.
(PNG)

S6 Fig. OptCom and Descriptive OptCom results mapped in the Pareto front.
(PNG)

Acknowledgments

MB is supported by CNRS and Region Pays de la Loire funding (GRIOTE project). This study was supported by ANR (IMPEKAB, ANR-15-CE02-001-03). We would like to thank Editage (www.editage.com) for English language editing. The authors would like also to thank the anonymous reviewer for his valuable input and comments regarding the manuscript.

Author Contributions

Conceptualization: DE JB MB.

Data curation: MB.

Formal analysis: MB.

Funding acquisition: DE JB.

Software: MB.

Supervision: DE JB AL.

Visualization: MB.

Writing – original draft: DE JB AL MB.

Writing – review & editing: DE JB AL MB.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95(12)6578–6583. doi: 10.1073/pnas.95.12.6578 PMID: 9618454
2. Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. Global distribution of microbial abundance and biomass in subseafloor sediment. Proceedings of the National Academy of Sciences. 2012; 109 (40):16213–16216. doi: 10.1073/pnas.1203849109
3. Lin BL, Sakoda A, Shibasaki R, Goto N, Suzuki M. Modelling a global biogeochemical nitrogen cycle in terrestrial ecosystems. Ecological Modelling. 2000; 135(1):89–110. doi: 10.1016/S0304-3800(00)00372-0
4. Rullkötter J. Organic Matter: The Driving Force for Early Diagenesis. In: Marine Geochemistry. Berlin/Heidelberg: Springer Berlin Heidelberg; 2006. p. 125–168.
5. Jessup CM, Kassen R, Forde SE, Kerr B, Buckling A, Rainey PB, et al. Big questions, small worlds: microbial model systems in ecology. Trends in Ecology & Evolution. 2004; 19(4):189–197. doi: 10.1016/j.tree.2004.01.008 PMID: 16701253
6. McGill BJ, Enquist BJ, Weiher E, Westoby M. Rebuilding community ecology from functional traits. Trends in Ecology & Evolution. 2006; 21(4):178–185. doi: 10.1016/j.tree.2006.02.002 PMID: 16701083
7. Krause S, Le Roux X, Niklaus PA, Van Bodegom PM, Lennon JT, Bertilsson S, et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Frontiers in Microbiology. 2014; 5(364):251. doi: 10.3389/fmicb.2014.00251 PMID: 24904563
8. Litchman E, Klausmeier CA. Trait-Based Community Ecology of Phytoplankton. Annual Review of Ecology, Evolution, and Systematics. 2008; 39(1):615–639. doi: 10.1146/annurev.ecolsys.39.110707.173549
9. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. Molecular Systems Biology. 2013; 9(666):1–15 doi: 10.1038/msb.2013.22 PMID: 23670539
10. Waldor MK, Tyson G, Borenstein E, Ochman H, Moeller A, Finlay BB, et al. Where Next for Microbiome Research? PLoS Biology. 2015; 13(1):e1002050. doi: 10.1371/journal.pbio.1002050 PMID: 25602283

11. Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. Nature Reviews. Microbiology. 2008; 6(9):693–699. PMID: 18587409
12. Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. Nature Reviews Microbiology. 2015; 13(3):133–146. doi: 10.1038/nrmicro3417 PMID: 25659323
13. 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. doi: 10.1038/nrmicro2737 PMID: 22367118
14. Bordbar A, Monk JM, King ZA, Palsson BO. Constraint-based models predict metabolic and associated cellular functions. Nature Reviews. Genetics. 2014; 15(2):107–120. PMID: 24430943
15. Klitgord N, Segrè D. Ecosystems biology of microbial metabolism. Current opinion in biotechnology. 2011; 22(4):541–546. doi: 10.1016/j.copbio.2011.04.018 PMID: 21592777
16. Zengler K, Palsson BO. A road map for the development of community systems (CoSy) biology. Nature Reviews: Microbiology. 2012; 10(5):366–372. PMID: 22450377
17. Kim TY, Sohn SB, Bin Kim Y, Kim WJ, Lee SY. Recent advances in reconstruction and applications of genome-scale metabolic models. Current Opinion in Biotechnology. 2012; 23(4):617–623. doi: 10.1016/j.copbio.2011.10.007 PMID: 22054827
18. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols. 2010; 5(1):93–121. doi: 10.1038/nprot.2009.203 PMID: 20057383
19. Hanemaaijer M, Röling WFM, Olivier BG, Khandelwal RA, Teusink B, Bruggeman FJ. Systems modeling approaches for microbial community studies: from metagenomics to inference of the community structure. Frontiers in Microbiology. 2015; 6(213):1–12. doi: 10.3389/fmicb.2015.00213 PMID: 25852671
20. Rodríguez J, Kleerebezem R, Lema JM, van Loosdrecht MCM. Modeling product formation in anaerobic mixed culture fermentations. Biotechnology and Bioengineering. (2006), 93(3), 592–606. doi: 10.1002/bit.20765
21. Biggs MB, Medlock GL, Kolling GL, Papin JA. Metabolic network modeling of microbial communities. WIRES Systems Biology and Medicine. 2015; 7:317–334 doi: 10.1002/wsbm.1308 PMID: 26109480
22. Perez-Garcia O, Lear G, Singhal N. Metabolic Network Modeling of Microbial Interactions in Natural and Engineered Environmental Systems. Frontiers in Microbiology. 2016; 7(673), 1–30. doi: 10.3389/fmicb.2016.00673 PMID: 27242701
23. Klitgord N, Segrè D The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. Genome Informatics (2009), 22, 41–55. PMID: 20238418
24. Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA. Metabolic modeling of a mutualistic microbial community. Molecular Systems Biology (2007), 3, 92:1–14.
25. Zomorrodi AR, Maranas CD. OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Computational Biology. 2012; 8(2):e1002363. doi: 10.1371/journal.pcbi.1002363 PMID: 22319433
26. Zomorrodi AR, Islam MM, Maranas CD. d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. ACS Synthetic Biology. 2014; 3(4):247–257. doi: 10.1021/sb4001307 PMID: 24742179
27. Khandelwal RA, Olivier BG, Röling WF, Teusink B, Bruggeman FJ. Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. PLoS ONE. 2013; 8(5):e64567. doi: 10.1371/journal.pone.0064567 PMID: 23741341
28. Koch S, Benndorf D, Fronk K, Reichl U, Klamt S. Predicting compositions of microbial communities from stoichiometric models with applications for the biogas process. Biotechnology for Biofuels. 2016; 9(1):1–16. doi: 10.1186/s13068-016-0429-x PMID: 26807149
29. Ehrgott M. Multicriteria Optimization. Berlin, Germany: Springer Science & Business Media; 2005.
30. Vo TD, Greenberg HJ, Palsson BO. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. Journal of Biological Chemistry. 2004; 279(38):39532–39540. doi: 10.1074/jbc.M403782200 PMID: 15205464
31. Kschischo M. A gentle introduction to the thermodynamics of biochemical stoichiometric networks in steady state. The European Physical Journal Special Topics. 2010; 187(1):255–274. doi: 10.1140/epjst/e2010-01290-3
32. Dillon JG, Fishbain S, Miller SR, Bebout BM, Habicht KS, Webb SM, et al. (2007). High rates of sulfate reduction in a low-sulfate hot spring microbial mat are driven by a low level of diversity of sulfate-respiring microorganisms. Applied and Environmental Microbiology. 2007; 73(16), 5218–5226. doi: 10.1128/AEM.00357-07 PMID: 17575000

33. Taffs R, Aston JE, Brileya K, Jay Z, Klatt CG, McGlynn S, et al. In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. BMC Systems Biology. 2009; 3(1):114. doi: 10.1186/1752-0509-3-114 PMID: 20003240

34. Beard DA, Liang Sd, Qian H. Energy balance for analysis of complex metabolic networks. Biophysical Journal. 2002; 83(1):79–86. doi: 10.1016/S0006-3495(02)75150-3 PMID: 12080101

35. Qian H, Beard DA, Liang Sd. Stoichiometric network theory for nonequilibrium biochemical systems. European Journal of Biochemistry / FEBS. 2003; 270(3):415–421. doi: 10.1046/j.1432-1033.2003.03357.x PMID: 12542691

36. Alberty RA. Appendix 2: Tables of Transformed Thermodynamic Properties. Applications of Mathmatica. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006.

37. Flamholz A, Noor E, Bar-Even A, Milo R. eQuilibrator–the biochemical thermodynamics calculator. Nucleic Acids Research. 2012; 40(Database issue):D770–D775. doi: 10.1093/nar/gkr874 PMID: 22064852

38. Hoppe A, Hoffmann S, Holzhütter HG. Including metabolite concentrations into flux balance analysis: thermodynamic realizability as a constraint on flux distributions in metabolic networks. BMC Systems Biology. 2007; 1(1):1–23 doi: 10.1186/1752-0509-1-23

39. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-Based Metabolic Flux Analysis. Biophysical Journal. 2007; 92(5):1792–1805. doi: 10.1529/biophysj.106.093138 PMID: 17172310

40. Fleming RMT, Thiele I, Provan G, Nasheuer HP. Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. Journal of Theoretical Biology. 2010; 264(3):683–692. doi: 10.1016/j.jtbi.2010.02.044 PMID: 20230840

41. Dantzig GB. Reminiscences About the Origins of Linear Programming. In: Mathematical Programming The State of the Art. Berlin, Germany: Springer; 1983. p. 78–86. Available from:

42. Gurobi Optimization I. Gurobi Optimizer Reference Manual; 2015. Available from: http://www.gurobi.com

43. Ehrgott M, Wiecek MM. Mutiobjective Programming. In: Multiple Criteria Decision Analysis: State of the Art Surveys. New York: Springer-Verlag; 2005. p. 667–708.

44. Aoki I. Entropy and exergy in the development of living systems: a case study of lake-ecosystems. Journal of the Physical Society of Japan. 1998; 67(6):2132–2139. doi: 10.1143/JPSJ.67.2132

45. Martyushev LM, Seleznev VD. Maximum entropy production principle in physics, chemistry and biology. Physics Reports. 2006; 426(1):1–45. doi: 10.1016/j.physrep.2005.12.001

46. Stadler W. A survey of multicriteria optimization or the vector maximum problem, part I: 1776–1960. Journal of Optimization Theory and Applications. 1979; 29(1):1–52. doi: 10.1007/BF00932634

47. Marler RT, Arora JS. Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization. 2004; 26(6):369–395. doi: 10.1007/s00158-003-0368-6

48. Benson HP. An Outer Approximation Algorithm for Generating AllEfficient Extreme Points in the Outcome Set of a Multiple ObjectiveLinear Programming Problem. Journal of Global Optimization. 1998; 13(1):1–24.

49. Ehrgott M, Shao L, Schöbel A. An approximation algorithm for convex multi-objective programming problems. Journal of Global Optimization. 2010; 50(3):397–416. doi: 10.1007/s10898-010-9588-7

50. Ehrgott M, Löhne A, Shao L. A dual variant of Benson's "outer approximation algorithm" for multiple objective linear programming. Journal of Global Optimization. 2012; 52(4):757–778. doi: 10.1007/s10898-011-9709-y

51. Hamel AH, Löhne A, Rudloff B. Benson type algorithms for linear vector optimization and applications. Journal of Global Optimization. 2013; 59(4):811–836. doi: 10.1007/s10898-013-0098-2

52. Löhne A, Weißing B. BENSOLVE—VLP Solver, version 2.0.2; 2015. Available from: http://www.bensolve.org

53. Anderson KL, Tayne TA, Ward DM. Formation and Fate of Fermentation Products in Hot Spring Cyanobacterial Mats. Applied and Environmental Microbiology. 1987; 53(10):2343–2352. PMID: 16347455

54. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Research. 2014; 42(Database issue):D459–471. doi: 10.1093/nar/gkt1103 PMID: 24225315

55. Steunou AS, Bhaya D, Bateson MM, Melendrez MC, Ward DM, Brecht E, et al. In situ analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(7):2398–2403. doi: 10.1073/pnas.0507513103 PMID: 16467157

56. Steunou AS, Jensen SI, Brecht E, Becraft ED, Bateson MM, Kilian O, et al. Regulation of nif gene expression and the energetics of N2 fixation over the diel cycle in a hot spring microbial mat. The ISME journal. 2008; 2(4):364–378. doi: 10.1038/ismej.2007.117 PMID: 18323780

57. Oberhardt MA, Chavali AK, Papin JA. Flux Balance Analysis: Interrogating Genome-Scale Metabolic Networks. In: Systems Biology. Totowa, NJ: Humana Press; 2009. p. 61–80.

58. Kim YM, Nowack S, Olsen MT, Becraft ED, Wood JM, Thiel V, et al. Diel metabolomics analysis of a hot spring chlorophototrophic microbial mat leads to new hypotheses of community member metabolisms. Frontiers in Microbiology. 2015; 6(209):1–14. doi: 10.3389/fmicb.2015.00209 PMID: 25941514

59. Nash SG. A survey of truncated-Newton methods. Journal of Computational and Applied Mathematics. 2000; 124(1-2):45–59. doi: 10.1016/S0377-0427(00)00426-X

60. Hunter JD. Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 2007; 9 (3):90–95. doi: 10.1109/MCSE.2007.55

61. Tawarmalani M, Sahinidis NV. A polyhedral branch-and-cut approach to global optimization. Mathematical Programming, 2005; 103(2), 225–249 doi: 10.1007/s10107-005-0581-8

62. Czyzyk J, Mesnier MP, Moré JJ. The NEOS Server. IEEE Journal on Computational Science and Engineering. 1998; 5(3), 68–75. doi: 10.1109/99.714603

63. Dolan E. The NEOS Server 4.0 Administrative Guide. Technical Memorandum ANL/MCS-TM-250, Mathematics and Computer Science Division, Argonne National Laboratory. 2001.

64. Gropp W, Moré JJ. Optimization Environments and the NEOS Server. In: Approximation Theory and Optimization, Buhmann MD and Iserles A, eds., Cambridge University Press; 1997, p167–182.

65. Klatt CG, Liu Z, Ludwig M, Kuhl M, Jensen SI, Bryant DA, et al. Temporal metatranscriptomic patterning in phototrophic Chloroflexi inhabiting a microbial mat in a geothermal spring. The ISME Journal. 2013; 7 (9):1775–1789. doi: 10.1038/ismej.2013.52 PMID: 23575369

66. Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(38):14482–14487. doi: 10.1073/pnas.0806162105 PMID: 18787117

67. Bordron P, Latorre M, Cortés MP, González M, Thiele S, Siegel A, et al. Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. MicrobiologyOpen. 2016; 5(1):106–117. doi: 10.1002/mbo3.315 PMID: 26677108

68. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional Optimality of Microbial Metabolism. Science (New York, NY). 2012; 336(6081):601–604. doi: 10.1126/science.1216882

69. England JL. Statistical physics of self-replication. The Journal of Chemical Physics. 2013; 139 (12):121923. doi: 10.1063/1.4818538 PMID: 24089735

9

Conclusions

As microbial ecology sampling efforts are intensifying, corresponding analysis has become a central scientific challenge, mainly because of the quantity and the heterogeneity of data. Complementary disciplines such as Systems Biology had been successful for integrating biological entities into organization such as networks. In this context, a call for new quantitative methods bridging both of these fields has been promoted (Widder et al., 2016; Martins Conde et al., 2016).

The present thesis represents an effort in this direction. In a general view, its main contribution is to demonstrate and extends the use of Constraint Based Models (CBM) in several problems related to microbial ecology. The use of CBM as modeling framework has several advantages. For instance, from a biological viewpoint, it allows a direct integration of genomic knowledge into the modeling. Furthermore, CBM incorporates directly the specific growth rate of organisms (either as a parameter or as a prediction goal), feature that is central in classic microbial ecology models (Poggiale et al., 2014). By interpreting specific growth rate as a measure of fitness, a new ecological dimension is added to models. From a computer science perspective, solving CBM reverts a solid theoretical foundations along with efficient solving algorithms, enabling simulations to be implemented with relative ease.

The first part of the thesis focused into showing how CBM links genome to phenotype features by using a recently developed CBM, called Stoichiometric Capacitance (SC) (Larhlimi et al., 2012a). Specifically, we estimated *in-silico* gene insertions such as: i) they increase specific growth rate and ii) they produce a metabolite of interest (here ethanol) as byproduct.

SC formulation can be interpreted from a biological viewpoint as a network reconfiguration that will increases an organism fitness by acquiring genes. In a way, it searches the space of (known) enzymatic reactions for the best possible adaptation. In addition, by studying the range of fluxes to achieve this maximal fitness, we were able to determine reactions that need to carry these fluxes, meaning that they are "obligatory" to achieve a fitness optima. Beyond practical applications such as ethanol production, SC shows that CBM can be used to study complex phenomena such as gene transfer and phenotypic adaptation.

CBM also offers a way to asses effect of evolutionary processes in cellular metabolism, demonstrating how emerging properties of metabolic networks can be studied using mathematical tools. When applied to *P. fluorescens* metabolic model, Flux Balance Analysis (FBA) and Flux Variability Analysis (FVA) were used to infer the relevance of certain fluxes to achieve optimal fitness. Using these results, genes associated with fluxes were classified as blocked, excluded, alternative or indispensable.

In this context, Red Queens Hypothesis (RQH) predicts that beneficial mutations should be fixed, affecting obligatory and alternative genes. In contrast, Black Queen Hypothesis (BQH) suggest that blocked and excluded genes could loose their function. Experimental observations show that alternative genes with 3 or more SNPs are present in higher proportions than expected by chance; also, an enzyme involved in al-

ternative carbon processing which is excluded in *P. fluorescens* is affected by mutations under experimental conditions.

Previous example assume environmental conditions as model parameters. However, it is also possible to reverse the problem by assessing excluded fluxes for a given set of conditions. In this purpose, we could attempt to calculate which environmental conditions must be satisfied to maximize excluded reactions. Formalization of this problem leds to a bi-level mixed integer program, that could not be solved by a standard method. Proposing a solution schema remains out of the scope of the present work; however, the formulation itself points to interesting interpretations of the environment-metabolism interplay. This highlights how formalization of these systems can help to find analogies in other fields of Computer Science as well as motivate new research in those fields.

In nature, however, microorganisms are rarely found living independently. Most of times they are found to be living in communities. Development of quantitative methods to model microbial ecosystems using CBMs is the main axis of the second part of present thesis. Moving from analyzing networks of interconnected cellular components to analyzing "networks of networks", *i.e.* networks of interconnected cells, such as multicellular organisms or microbial ecosystems, seems a natural extension of the field itself.

Such quantitative modelings raise questions about the mathematical properties of interconnected metabolic networks and how their modelings should be tackled. By using recent advances in the field (Müller and Bockmayr, 2013), it was shown that taking into account each microorganism as a single entity is relevant qualitatively and, to a lesser extent, quantitatively, confirming previously results obtained using a different approach (Klitgord and Segrè, 2009). Therefore, a CBM aiming to cope with microbial ecosystems should consider compartments *per se* for the sake of model investigation.

Conversely, analysis of the different approaches for model communities shown, in one hand, that current methods have limitations by either considering a single compartment or using an aggregate objective function (AOF) to represent an ecosystem objective. Conceptually, using an AOF implies weighting each of the objectives, which could lead to computational artifacts. Current approaches that consider a system objective as well as each entities objectives are based in bi-level optimization, which add a layer of mathematical complexity. In the other hand, literature review shown that multiple objectives in biological systems are well represented by the Pareto Front (Schuetz et al., 2012). Thus, results motivated a revision of the CBM mathematical framework towards a multi-objective optimization.

As a result, a dedicated CBM method was proposed, emphasizing a geometrical description of the Pareto Front and solved using a Benson Outer Optimization Algorithm. For the sake of the application, this method was applied to model a community ecosystem, comprising three distinct phenotypes: a primary producer, *Synecococcus spp.* (SYN), filamentous anoxygenic producers (FAP), namely *Chloroflexus spp.* and *Roseiflexus spp.*; and sulfate-reducing bacteria (SRB, composed by *Thermodesulfovibrio spp.*-like activity, (Dillon et al., 2007)), as described in Taffs et al. (2009). The description of the corresponding Pareto Front enabled further investigations of other principles involved in the community via microbial interactions such as entropy generation, which seems to play an important role to drive the biomass distribution within the ecosystem.

Present work advocates for the use of CBM as a modeling technique for microbial ecology, showing that CBM represents well metabolic aspects and capture emerging properties of community. However, despite their wide application, CBM still present some challenges to overcome. For instance, time dependent or non metabolic phenomena (such as regulation) are usually difficult to model using CBM, although some works proposed solutions to these limitations (Covert et al., 2001; Zomorrodi et al., 2014). Development in this direction would be beneficial to improve presented methods.

Nevertheless, despite above limitations, CBM has been proven to be flexible enough for tackling different kind problems and could be a cornerstone framework to understand microbial ecosystems. As Biology become more and more quantitative, this thesis shows that CBMs is an appropriate framework to model microbial ecology from genes to communities and can be tailored to different problems. CBMs have the potential to go from an specialized Systems Biology approach to a standard analysis toolkit for recent Biology and Biotechnology progresses.

For instance, CBM are able to capture interactions between the metabolism and its environment and vice-versa, as shown in Section I. Determining obligated and excluded reactions is an interesting perspective, for both practical and theoretical reasons. In practice, they represent potential targets for genetic modifications, either due to chance or design. In theory, these reactions are interesting from an evolutionary perspective, Indeed, one could explore if such critical functions are: (i) either product of convergent evolution or (ii) due to horizontal transfer in a particular niche, or redundant (*e.g.* organized in gene tandems) in a single specie. Likewise, one could explore why certain functions remain present if they do not have a direct impact over the microbial fitness, revealing, perhaps, an incomplete understanding of their role in the system.

From an application viewpoint, as CBMs are mathematically based in optimization, they can in turn be related to decision problems. Therefore, CBM framework can be extended beyond simulation to tackle design problems. For example, in the case of RQH-BQH, FBA and FVA were applied in a simulation context to observe the properties of the metabolic network: Given media composition, obligatory, alternative, excluded and blocked reactions are emphasized. Another application of CBMs to design problems is found in the Capacitance application. In this work, we calculated capacitances for *E. coli* to obtain products of interest (ethanol and amino acids, respectively). Furthermore, capacitances can be decomposed into known enzymatic reactions, which can be used as guide to effective gene insertions. Such modifications have the added value that as they improve the fitness, they should be fixed in the populations, helping in the maintaining such function in time.

Considering the description of a community optimum as a Pareto Front is of particular interest. With a description a community optimum as a region in the solution space, Pareto optimality can be included as restriction in more sophisticated CBMs. This could lead to new ways to control microbial fractions in heterogeneous populations, a key factor to improve the quality control in complex biotechnological processes. Furthermore, an interesting possibility is given by a redefinition of Stoichiometric Capacitance in terms of communities, by calculating which "community functions" (*e.g.*, reactions carried by a specific bacterial group) must be introduced in the system to produce a certain effect, such as waste cleaning or control population of nocive organisms. Such applications could led to the emergence of a "synthetic ecology" field, mimicking contributions of CBM to synthetic biology (Barrett et al., 2006; Burk and Van Dien, 2016).

In general, mathematical formalization of biological systems offers new perspectives to understand biological systems properties. In particular, using these types of formalisms to study microbial ecosystems promotes links between physical principles such as entropy production and self-organization of biological systems, pointing to new and exciting research lines. For example, it could be possible to model environments (either at laboratory or ecological scale) as systems which chemical energy potentials which move from one state to another where microorganisms plays the role of "entropy dissipators", by reducing potentials of media and/or mass dispersion by replication. However, as the state of the system is not necessary in chemical equilibrium, non-equilibrium thermodynamics should be used. Advances has been made recently in non-equilibrium thermodynamics, as well as statical mechanics of self replication (Glansdorff and Prigogine, 1964; England, 2013; Kondepudi and Prigogine, 2014) which could be used to derive new constraints and objective functions in CBM, as well as other modeling approaches. Hopefully, such developments would lead to new insights in underlying principles governing biological systems.

This work is expected to contribute to the development of CBMs in ecology, extending from a "gene-phenotype" link to a "genes-environment" one. Large sampling projects such as TARA Oceans, gut microbiomes and others are generating detailed maps of microbes in natural environments as well as gene catalogs of functions present in each niche (Bork et al., 2015; Chaffron et al., 2010; Mandal et al., 2015; Magne et al., 2016). We expect that the present framework and its future extensions could serve as an articulating paradigm in microbial ecology. Furthermore, we think that proposed approaches will serve to detailed exploration of those massive datasets, leading to new global understanding of effects of microbial ecology from human health to biogeochemical cycles.

List of Tables

List of Figures

Bibliography

Achterberg, T., Koch, T., and Martin, A. (2005). Branching rules revisited. *Operations Research Letters*, 33(1):42–54.

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M.-F., and Stougie, L. (2009). Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60.

Acuña, V., Marchetti-Spaccamela, A., Sagot, M.-F., and Stougie, L. (2010). A note on the complexity of finding and enumerating elementary modes. *Biosystems*, 99(3):210–214.

Adams, E. and Rosso, G. (1967). Alpha-ketoglutaric semialdehyde dehydrogenase of Pseudomonas. Properties of the purified enzyme induced by hydroxyproline and of the glucarate-induced and constitutive enzymes. *Journal of Biological Chemistry*, 242(8):1802–1814.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Bard, J. F. (1991). Some properties of the bilevel programming problem. *Journal of Optimization Theory and Applications*, 68(2):371–378.

Barrett, C. L., Kim, T. Y., Kim, H. U., Palsson, B. O., and Lee, S. Y. (2006). Systems biology as a foundation for genome-scale synthetic biology. *Current opinion in biotechnology*, 17(5):488–492.

Benson, H. P. (1998). An Outer Approximation Algorithm for Generating AllEfficient Extreme Points in the Outcome Set of a Multiple ObjectiveLinear Programming Problem. *Journal of Global Optimization*, 13(1).

Biggs, M. B., Medlock, G. L., Kolling, G. L., and Papin, J. A. (2015). Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334.

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

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science (New York, NY)*, 348(6237):873.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge, UK ; New York : Cambridge University Press.

Budinich, M., Bourdon, J., Larhlimi, A., and Eveillard, D. (2015). OPINION PAPER Evolutionary Constraint-Based Formulation Requires New Bi-level Solving Techniques. In *Computational Methods in Systems Biology*, pages 279–281. Springer International Publishing, Cham.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657.

Burk, M. J. and Van Dien, S. (2016). Biotechnology for Chemical Production: Challenges and Opportunities. *Trends in biotechnology*, 34(3):187–190.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(Database issue):D459–71.

Chaffron, S., Rehrauer, H., Pernthaler, J., and von Mering, C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959.

Covert, M. W., Schilling, C. H., and Palsson, B. (2001). Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213(1):73 – 88.

Dillon, J. G., Fishbain, S., Miller, S. R., Bebout, B. M., Habicht, K. S., Webb, S. M., and Stahl, D. A. (2007). High rates of sulfate reduction in a low-sulfate hot spring microbial mat are driven by a low level of diversity of sulfate-respiring microorganisms. *Applied and Environmental Microbiology*, 73(16):5218–5226.

Droop, M. R. (1968). Vitamin B*12* and Marine Ecology. IV. The Kinetics of Uptake, Growth and Inhibition in *Monochrysis lutheri*. *Journal of the Marine Biological Association of the United Kingdom*, 48(3):689–733.

Edwards, J. S. and Palsson, B. O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5528–5533.

Ehrgott, M., Shao, L., and Schöbel, A. (2010). An approximation algorithm for convex multi-objective programming problems. *Journal of Global Optimization*, 50(3):397–416.

Ehrgott, M. and Wiecek, M. M. (2005). Mutiobjective Programming. In *Multiple Criteria Decision Analysis: State of the Art Surveys*, pages 667–708. Springer-Verlag, New York.

England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12):121923.

Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nature Publishing Group*, 10(8):538–550.

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

Fong, S. S. (2014). Computational approaches to metabolic engineering utilizing systems biology and synthetic biology. *Computational and Structural Biotechnology*, 11(18):28–34.

Gause, G. F. (1934). *The struggle for existence*. Williams and Wilkins.

Gebser, M., Kaminski, R., Kaufmann, B., and Schaub, T. (2012). *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers.

Glansdorff, P. and Prigogine, I. (1964). On a general evolution criterion in macroscopic physics. *Physica*.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., Robbins, M. M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R. A., Vabø, R., Visser, U., and DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115–126.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., Darzi, Y., Audic, S., Berline, L., Brum, J. R., Coelho, L. P., Espinoza, J. C. I., Malviya, S., Sunagawa, S., Dimier, C., Kandels-Lewis, S., Picheral, M., Poulain, J., Searson, S., Tara Oceans Consortium Coordinators, Stemmann, L., Not, F., Hingamp, P., Speich, S., Follows, M., Karp-Boss, L., Boss, E., Ogata, H., Pesant, S., Weissenbach, J., Wincker, P., Acinas, S. G., Bork, P., de Vargas, C., Iudicone, D., Sullivan, M. B., Raes, J., Karsenti, E., Bowler, C., and Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470.

Hädicke, O. and Klamt, S. (2010). CASOP: A Computational Approach for Strain Optimization aiming at high Productivity. *Journal of Biotechnology*, 147(2):88–101.

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236.

Hagen, J. B. (2011). The origin and early reception of sequence databases. In Hamacher, M., Eisenacher, M., and Stephan, C., editors, *Data Mining in Proteomics: From Standards to Applications*, pages 61–77. Humana Press, Totowa, NJ.

Hamel, A. H., Löhne, A., and Rudloff, B. (2013). Benson type algorithms for linear vector optimization and applications. *Journal of Global Optimization*, 59(4):811–836.

Hanemaaijer, M., RÃ ling, W. F. M., Olivier, B. G., Khandelwal, R. A., Teusink, B., and Bruggeman, F. J. (2015). Systems modeling approaches for microbial community studies: from metagenomics to inference of the community structure. *Frontiers in microbiology*, 6.

Hanly, T. J. and Henson, M. A. (2010). Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. *Biotechnology and bioengineering*, 108(2):376–385.

Huerta, M., Haseltine, F., Liu, Y., Downing, G., and Seto, B. (2000). NIH Working Definition of Bioinformatics and Computational Biology.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7):3801–3806.

Johnson, C. W. (2006). What are emergent properties and how do they affect the engineering of complex systems? *Reliability Engineering & System Safety*, 91(12):1475–1481.

Joyce, A. R. and Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3):198–210.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., and D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences*, 109(40):16213–16216.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42(D1):D199–D205.

Kelk, S. M., Olivier, B. G., Stougie, L., and Bruggeman, F. J. (2012). Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Scientific reports*, 2:580.

Khandelwal, R. A., Olivier, B. G., Röling, W. F., Teusink, B., and Bruggeman, F. J. (2013). Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE*, 8(5):e64567.

Kim, T. Y., Sohn, S. B., Bin Kim, Y., Kim, W. J., and Lee, S. Y. (2012). Recent advances in reconstruction and applications of genome-scale metabolic models. *Current opinion in biotechnology*, 23(4):617–623.

Kirner, S., Krauss, S., Sury, G., Lam, S. T., Ligon, J. M., and van Pée, K.-H. (1996). The non-haem chloroperoxidase from pseudomonas fluorescens and its relationship to pyrrolnitrin biosynthesis. *Microbiology*, 142(8):2129–2135.

Kitano, H. (2002a). Computational systems biology. *Nature*, 420(6912):206–210.

Kitano, H. (2002b). Systems biology: a brief overview. *Science (New York, NY)*, 295(5560):1662–1664.

Klamt, S. and Gilles, E. D. (2004). Minimal cut sets in biochemical reaction networks. *Bioinformatics (Oxford, England)*, 20(2):226–234.

Klamt, S. and Stelling, J. (2003). Two approaches for metabolic pathway analysis? *Trends in biotechnology*, 21(2):64–69.

Klitgord, N. and Segrè, D. (2009). The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Informatics*, 22:41–55.

Klitgord, N. and Segrè, D. (2010). Environments that Induce Synthetic Microbial Ecosystems. *PLoS computational biology*, 6(11):e1001002.

Klitgord, N. and Segrè, D. (2011). Ecosystems biology of microbial metabolism. *Current opinion in biotechnology*, 22(4):541–546.

Kondepudi, D. and Prigogine, I. (2014). *Modern thermodynamics: from heat engines to dissipative structures*. John Wiley & Sons.

Krause, S., Le Roux, X., Niklaus, P. A., Van Bodegom, P. M., Lennon, J. T., Bertilsson, S., Grossart, H.-P., Philippot, L., and Bodelier, P. L. E. (2014). Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in microbiology*, 5(364):251.

Larhlimi, A., Basler, G., Grimbs, S., Selbig, J., and Nikoloski, Z. (2012a). Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics (Oxford, England)*, 28(18):i502–i508.

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

Lin, B. L., Sakoda, A., Shibasaki, R., Goto, N., and Suzuki, M. (2000). Modelling a global biogeochemical nitrogen cycle in terrestrial ecosystems. *Ecological Modelling*, 135(1):89–110.

Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science (New York, NY)*, 227(4693):1435–1441.

Lotka, A. J. (1925). Elements of Physical Biology.

Magne, F., ORyan, M. L., Vidal, R., and Farfan, M. (2016). The human gut microbiome of latin america populations: a landscape to be discovered. *Current Opinion in Infectious Diseases*, 29(5):528–537.

Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophysical Journal*, 83(3):1331–1340.

Mahadevan, R. and Schilling, C. H. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276.

Mandal, R. S., Saha, S., and Das, S. (2015). Metagenomic Surveys of Gut Microbiota. *Genomics, Proteomics & Bioinformatics*, 13(3):148–158.

Martins Conde, P. d. R., Sauter, T., and Pfau, T. (2016). Constraint Based Modeling Going Multicellular. *Frontiers in Molecular Biosciences*, 3(4):158–111.

Mas, A., Jamshidi, S., Lagadeuc, Y., Eveillard, D., and Vandenkoornhuyse, P. (2016). Beyond the Black Queen Hypothesis. *Nature Publishing Group*, pages 1–7.

Monk, J., Nogales, J., and Palsson, B. O. (2014). Optimizing genome-scale network reconstructions. *Nature Publishing Group*, 32(5):447–452.

Monod, J. (1949). The growth of bacterial cultures. *Annual Reviews in Microbiology*.

Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*, 3(2):–.

Müller, A. C. and Bockmayr, A. (2013). Flux modules in metabolic networks. *Journal of Mathematical Biology*.

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., and Yarmush, M. L. (2007). Integrated Energy and Flux Balance Based Multiobjective Framework for Large-Scale Metabolic Networks. *Annals of Biomedical Engineering*, 35(6):863–885.

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A., and Yarmush, M. L. (2010). Metabolic Engineering. *Metabolic engineering*, 12(5):429–445.

Oh, Y.-G., Lee, D.-Y., Lee, S. Y., and Park, S. (2009). Multiobjective flux balancing using the NISE method for metabolic network analysis. *Biotechnology progress*, 25(4):999–1008.

Oh, Y.-G., Lee, D.-Y., Yuri, H., Lee, S. Y., and Park, S. (2004). Multi-product trade-off analysis of e. coli by multiobjective flux balance analysis. In Barbosa-Póvoa, A. and Matos, H., editors, *European Symposium on Computer-Aided Process Engineering-14, 37th European Symposium of the Working Party on Computer-Aided Process Engineering*, volume 18 of *Computer Aided Chemical Engineering*, pages 1099 – 1104. Elsevier.

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., and Palsson, B. O. (2011). A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Molecular Systems Biology*, 7:1–9.

Perez-Garcia, O., Lear, G., and Singhal, N. (2016). Metabolic Network Modeling of Microbial Interactions in Natural and Engineered Environmental Systems. *Frontiers in microbiology*, 7(474):93–30.

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Research*, 14(11):2367–2376.

Pharkya, P. and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metabolic engineering*, 8(1):1–13.

Picioreanu, C., van Loosdrecht, M. C., and Heijnen, J. J. (1998). Mathematical modeling of biofilm structure with a hybrid differential-discrete cellular automaton approach. *Biotechnology and bioengineering*, 58(1):101–116.

Picioreanu, C., Xavier, J. B., and van Loosdrecht, M. C. M. (2004). Advances in mathematical modeling of biofilm structure. *Biofilms*, 1(4):337–349.

Poggiale, J.-C., Dantigny, P., de Wit, R., and Steinberg, C. (2014). Modeling in Microbial Ecology. In *Environmental Microbiology: Fundamentals and Applications*, pages 847–882. Springer Netherlands, Dordrecht.

Pozo, C., Guillén-Gosálbez, G., Sorribas, A., and Jiménez, L. (2012). Identifying the Preferred Subset of Enzymatic Profiles in Nonlinear Kinetic Metabolic Models via Multiobjective Global Optimization and Pareto Filters. *PLoS ONE*, 7(9):e43487–11.

Raes, J., Letunic, I., Yamada, T., Jensen, L. J., and Bork, P. (2011). Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*, 7:1–9.

Raman, K. and Chandra, N. (2009). Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics*, 10(4):435–449.

Ranganathan, S., Suthers, P. F., and Maranas, C. D. (2010). OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions. *PLoS computational biology*, 6(4):e1000744–11.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2014). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437.

Rullkötter, J. (2006). Organic Matter: The Driving Force for Early Diagenesis. In *Marine Geochemistry*, pages 125–168. Springer Berlin Heidelberg, Berlin/Heidelberg.

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

Schilling, C. H., Letscher, D., and Palsson, B. O. (2000). Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *Journal of theoretical biology*, 203(3):229–248.

Schrijver, A. (1998). *Theory of Linear and Integer Programming*. John Wiley & Sons.

Schrijver, A. (2011). *Combinatorial Optimization*. Springer-Verlag Berlin Heidelberg.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional Optimality of Microbial Metabolism. *Science (New York, NY)*, 336(6081):601–604.

Schuster, S., Fell, D. A., and Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(3):326–332.

Schuster, S. and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 02(02):165–182.

Silby, M. W., Cerdeño-Tárraga, A. M., Vernikos, G. S., Giddens, S. R., Jackson, R. W., Preston, G. M., Zhang, X.-X., Moon, C. D., Gehrig, S. M., Godfrey, S. A., Knight, C. G., Malone, J. G., Robinson, Z., Spiers, A. J., Harris, S., Challis, G. L., Yaxley, A. M., Harris, D., Seeger, K., Murphy, L., Rutter, S., Squares, R., Quail, M. A., Saunders, E., Mavromatis, K., Brettin, T. S., Bentley, S. D., Hothersall, J., Stephens, E., Thomas, C. M., Parkhill, J., Levy, S. B., Rainey, P. B., and Thomson, N. R. (2009). Genomic and genetic analyses of diversity and plant interactions of pseudomonas fluorescens. *Genome Biology*, 10(5):1–16.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193.

Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, 3:92.

Taffs, R., Aston, J. E., Brileya, K., Jay, Z., Klatt, C. G., McGlynn, S., Mallette, N., Montross, S., Gerlach, R., Inskeep, W. P., Ward, D. M., and Carlson, R. P. (2009). In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Systems Biology*, 3(1):114.

Talbi, E.-G. (2013). A taxonomy of metaheuristics for bi-level optimization. In Talbi, E.-G., editor, *Metaheuristics for Bi-level Optimization*, chapter 1, pages 1–34. Springer, Heidelberg.

Terzer, M. and Stelling, J. (2008). Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics (Oxford, England)*, 24(19):2229–2235.

Thiele, I., Heinken, A., and Fleming, R. M. (2013). A systems biology approach to studying the role of microbes in human health. *Current opinion in biotechnology*, 24(1):4–12.

Thiele, I. and Palsson, B. O. (2010). A protocol for generating a high-qualitygenome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121.

Treangen, T. J. and Rocha, E. P. C. (2011). Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genetics*, 7(1):e1001284–12.

Tzamali, E., Poirazi, P., Tollis, I. G., and Reczko, M. (2011). A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities. *BMC Systems Biology*, 5(1):167.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary theory*, 1:1–30.

Van Valen, L. (1974). Molecular evolution as predicted by natural selection. *Journal of molecular evolution*, 3(2):89–101.

Varma, A. and Palsson, B. O. (1994a). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/technology*.

Varma, A. and Palsson, B. O. (1994b). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Applied and Environmental Microbiology*, 60(10):3724–3731.

Vicente, L., Savard, G., and Júdice, J. (1994). Descent approaches for quadratic bilevel programming. *Journal of Optimization Theory and Applications*, 81(2):379–399.

Vieira-Silva, S., Falony, G., Darzi, Y., Lima-Mendez, G., Yunta, R. G., Okuda, S., Vandeputte, D., Valles-Colomer, M., Hildebrand, F., Chaffron, S., and Raes, J. (2016). Species-function relationships shape ecological properties of the human gut microbiome. *Nature Microbiology*, 1(8):1 –8.

Vo, T. D., Greenberg, H. J., and Palsson, B. O. (2004). Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *Journal of Biological Chemistry*, 279(38):39532–39540.

Volterra, V. (1926). Fluctuations in the Abundance of a Species considered Mathematically. *Nature*, 118(2972):558–560.

Wargacki, A. J., Leonard, E., Win, M. N., Regitsky, D. D., Santos, C. N. S., Kim, P. B., Cooper, S. R., Raisner, R. M., Herman, A., Sivitz, A. B., Lakshmanaswamy, A., Kashiyama, Y., Baker, D., and Yoshikuni, Y. (2012). An Engineered Microbial Platform for Direct Biofuel Production from Brown Macroalgae. *Science (New York, NY)*, 335(6066):308–313.

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583.

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., Cordero, O. X., Brown, S. P., Momeni, B., Shou, W., Kettle, H., Flint, H. J., Haas, A. F., Laroche, B., Kreft, J.-U., Rainey, P. B., Freilich, S., Schuster, S., Milferstedt, K., van der Meer, J. R., Groβkopf, T., Huisman, J., Free, A., Picioreanu, C., Quince, C., Klapper, I., Labarthe, S., Smets, B. F., Wang, H., Isaac Newton Institute Fellows, and Soyer, O. S. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME journal*, 10(11):2557–2568.

Wintermute, E. H. and Silver, P. A. (2010). Emergent cooperation in microbial metabolism. *Molecular Systems Biology*, 6:1–7.

Yang, L., Cluett, W. R., and Mahadevan, R. (2011). EMILiO A fast algorithm for genome-scale strain design. *Metabolic engineering*, 13(3):272–281.

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

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):Article17.

Zomorrodi, A. R., Islam, M. M., and Maranas, C. D. (2014). d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, 3(4):247–257.

Zomorrodi, A. R. and Maranas, C. D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology*, 8(2):e1002363.

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Marko BUDINICH ABARCA

Modélisation des Réseaux Métaboliques en interaction avec l'Environnement

Modeling Metabolic Networks and their Environment Interaction

Résumé

Les réseaux métaboliques permettent à l'utilisateur la construction de modèles détaillés en utilisant des jeux de données dites "omiques" de haute résolution. En particulier, les modèles par contraintes (CBM, en anglais) sont utilisés pour obtenir des prédictions quantitatives à partir de modèles métaboliques. Pendent les 20 dernières années, CBMs ont été appliqués avec succès à un large éventail de problèmes dans plusieurs aspects de la physiologie microbienne.

L'objectif principal de la présente thèse est d'utiliser les CBM comme une technique de modélisation dans le contexte de l'écologie microbienne. En particulier, les effets du réseau métabolique sur l'environnement et les effets des variables environnementales sur la physiologie sont explorés à l'aide de mesures de confiance.

La première section est dédiée à l'application des CBM à des réseaux métaboliques isolées. D'abord, une nouvelle application de CBM est utilisée pour étudier l'insertion de gènes tels qu'ils sont optimales pour maximiser le taux de croissance. Ensuite, les effets des conditions environnementales dans une chemostat chez le réseau métabolique, sont évalués par des approches CBM classiques et contrastés avec des observations expérimentales. Enfin, un nouveau CBM est développé pour déterminer les conditions environnementales telles qu'elles favorisent le la perte des gènes.

La deuxième section comporte sur les interactions entre plusieurs réseaux métaboliques différents. L'utilisation de compartiments pour représenter différents microorganismes est d'abord justifiée. Ensuite, une révision des approches existantes dans la littérature est réalisée. Après cette révision, un nouveau CBM basé dans l'optimisation MultiObjective pour l'écosystème microbien est développé.

On s'attend à que l'ensemble des travaux développés dans la thèse pourrait servir à rapprocher les champs de l'écologie microbienne et la modélisation par contraintes.

Mots clés

Réseaux Métaboliques, Modélisation par Contraintes, Ecologie Microbienne, Optimisation MultiObjective.

Abstract

Metabolic networks allows to the user the construction of detailed models using high resolution 'omics datasets. In particular, Constrained Based Models (CBMs) are used to obtain quantitative predictions from metabolic models. CBMs have been successfully applied to a wide range of problems for the last 20 years to several aspects of microbial physiology.

Main objective of present thesis is to use CBMs as a modeling technique in Microbial Ecology context. In particular, both metabolic network effects over the environment and effects of environmental variables over physiology are explored using CBMs.

In the first section, applications of CBMs to single metabolic networks are explored. First a novel application of CBMs is used to study gene insertion such they are optimal to maximize the growth rate. Next, effects of environmental conditions in a chemostat culture in metabolic network are assed by classical CBMs approaches and contrasted with experimental observations. Finally, a new CBMs is developed to determinate environmental conditions such as they favor gene loose.

Second section deals with interactions between multiple metabolic networks. The use of compartments to represent different microorganisms is first justified. Next, a revision of existent approaches in the literature is carried. After this revision, a new CBM based in MultiObjective Optimization for microbial ecosystem is developed.

Set of works developed in present thesis is expected to help filling the gap between Microbial Ecology and Constraint Based Modeling.

Key Words

Metabolic Networks, Constraint Based Methods, Microbial Ecology, MultiObjective Optimization.