From Systems Biology to Systems Ecology: A Computational Journey Through Biological Scales

Damien Eveillard

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE NANTES

HABILITATION A DIRIGER DES RECHERCHES

L'UNIVERSITE DE NANTES

COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601

Mathématiques et Sciences et Technologies de l'Information et de la Communication

Spécialité : Bioinformatique

Par

Damien EVEILLARD

From Systems Biology to Systems Ecology: a computational journey through biological scales

Habilitation présentée et soutenue à Nantes, le 13 octobre 2020
Unité de recherche : Laboratoire des Sciences Numériques de Nantes (LS2N)
Thèse N° : ##

Rapporteurs avant soutenance :

Christopher QUINCE	Associate Professor, University of Warwick
Claudine MEDIGUE	Directrice de Recherches, CNRS Genoscope
Éric RIVALS	Directeur de Recherches, CNRS LIRMM

Composition du Jury :

Président :	Prénom Nom	Fonction et établissement d'exercice	*(à préciser après la soutenance)*
Examinateurs :			
	Alexander BOCKMAYR	Professor, Freie Universität Berlin	
	Jérémie BOURDON	Professeur, Université de Nantes	
	Karoline FAUST	Assistant Professor, KU Leuven	
	Philippe VANDENKOORNHUYSE	Professeur, Université de Rennes 2	

From Systems Biology to Systems Ecology: a computational journey through biological scales

Damien Eveillard

October 13th, 2020

Avant propos

Je voudrais exprimer ma profonde gratitude aux membres du jury qui ont bien voulu évaluer mon travail. Si un travail interdisciplinaire est gratifiant, réunir un jury de grande qualité l'est encore plus.

Je voudrais remercier chaleureusement Claudine MEDIGUE, Christopher QUINCE et Eric RIVALS pour l'évaluation de ce rapport d'habilitation qui est une étape importante dans une expérience académique. J'ai conscience des contraintes que cela impliquent, et que celles ci sont amplifiées dans les conditions sanitaires que nous vivons.

Je voudrais également remercier Karoline FAUST pour avoir donné son accord à l'évaluation préalable de mon dossier et pour avoir accepté de siéger dans ce jury. Je tiens aussi à remercier Alexander BOCKMAYR, Jérémie BOURDON, et Philippe VANDENKOORNHUYSE de bien vouloir venir compléter un formidable jury interdisciplinaire.

Les travaux présentés ici ont été développés au sein de deux laboratoires, qui n'en font qu'un, le LINA puis le LS2N, dans lesquels j'ai trouvé des conditions de travail et de liberté exceptionnelles. Ces conditions sont suffisamment rares aujourd'hui dans un contexte de compétitions pour des financements, pour lesquels, il aurait été plus simple de minimiser la prise de risque. Pour cela, je voudrais remercier chaleureusement Frédéric BENHAMOU, Pierre COINTE et Claude JARD pour m'avoir laissé jouer une carte de l'interdisciplinarité et de la curiosité dans un système contraint.

Dans ce laboratoire, Jérémie BOURDON a joué un rôle pivot. Nous avons partagé plus qu'un bureau. Nombre de nos discussions (et de cafés) ont été le fondement d'un cadre de travail riche en nouvelle théories et idées plus ou moins fructueuses – mais qui sont les raisons pour lesquelles j'aime la recherche.

Ce plaisir pour les travaux de recherche n'aurait été possible sans certains chercheurs qui ont marqué mon éducation scientifique. Pour cela, je voudrais remercier dans l'ordre chronologique, Olivier BERNARD pour m'avoir dévoyé de l'océanographie traditionnelle et m'avoir ouvert les yeux sur la modélisation au sens large et sa rigueur de validation. Sans Alexander BOCKMAYR, je n'aurais jamais entrevu la beauté et la puissance des paradigmes informatiques. Mon expérience de thèse à ses côtés a fait naitre une curiosité sans fin pour la compréhension d'un monde biologique sous contraintes et de la beauté de sa seule formalisation. Je suis également plus que redevable à George JACKSON et à son ouverture d'esprit. Au-delà de son accueil dans un laboratoire qui rapproche les mondes scientifiques aussi divergents que la physique et la biologie des océans (et qui forme à l'interdisciplinarité), j'ai découvert et redécouvert à son contact tant de questions fascinantes, et beaucoup d'autres choses. J'ai également une pensée chaleureuse

pour Philippe VANDENKOORNHUYSE qui a mis le feu aux poudres de mes aspirations écologiques. Nos discussions autour de l'énergie et la diversité des écosystèmes raisonnent encore. Sans celles-ci et nos différents projets, je n'aurais pas osé faire le pas vers l'écologie dans un contexte francophone.

Puisqu'une Habilitation à Diriger des Recherches marque un jalon académique pour l'encadrement doctoral, je ne peux pas oublier de mentionner Philippe BORDRON pour les développements en biologie intégrative et Marko BUDINICH pour la modélisation des communautés bactériennes. Je suis désolé si, sans le savoir, je n'ai pas été à la hauteur ; mais je voudrais vous remercier pour m'avoir fait évoluer et pour nos discussions devant un tableau blanc ou un verre.

La recherche est par ailleurs un travail d'équipe. Ce manuscrit est donc aussi le résultat de beaucoup de choses qui ont plus de sens que leurs simples accumulations. Je voudrais remercier les membres de l'équipe ComBi passés et présents et particulièrement Iréna et Guillaume pour leurs confiances à mon arrivée. Merci à Géraldine pour ta franchise et pour partager tes doutes et aspirations. Merci à Abdelhalim pour ton amitié qui m'a accompagné tout au long des travaux présentés dans ce manuscrit – sans parler de ton thé. Merci à Benoit sans qui la passerelle n'existerait pas et les probabilités ne seraient pas pareils – ou l'inverse. Merci à Sam de repousser mes connaissances des microbiotes et savoir relâcher le krachen qui sommeil en chacun de nous. Merci à Audrey de mettre mes travaux en perspectives via BIRD et plus généralement pour nos discussions plus ou moins débridées en fonction du substrat "Chouffe". Merci à Alex pour avoir co-fondé avec moi la belle science qu'est la « spotologie » à la nantaise et pour nos belles échappées. Merci aux Taranautes, parmi lesquels Lionel et Lucie, pour votre écoute bienveillante de mes délires de modélisateur.

Je remercie également mes très proches, Michel et Marielle, pour leurs soutiens depuis le tout début, et Nadège pour être Nadège Finalement, ce manuscrit est dédié à Pauline qui aura, sans le vouloir, retardé sa finalisation, mais qui me démontre tous les jours la beauté de la résilience du système vivant mais aussi de son entropie dans un contexte chaotique.

Abstract

Recent progress in metagenomics has promoted a change of paradigm to investigate microbial ecosystems. These ecosystems are today analyzed by their gene content that, in particular, allows to emphasize the microbial composition in terms of taxonomy (i.e., « who is there and who is not ») or, more recently, their putative functions. However, understanding the interactions between microbial communities and their environment well enough to be able to predict diversity based on physicochemical parameters is a fundamental pursuit of microbial ecology that still eludes us. This task requires deciphering the mechanistic rules that prevail at the molecular level. Such a task must be achieved by dedicated computational approaches or modelings, as inspired by Systems Biology. Nevertheless, the direct application of standard cellular systems biology approaches is a complicated task. Indeed, the metagenomic description of ecosystems shows a large number of variables to investigate. Furthermore, communities are (i) complex, (ii) mostly described qualitatively, and (iii) the quantitative understanding of the way communities interact with their surroundings remains incomplete. Within this research summary, we will illustrate how systems biology approaches must be adapted to overcome these points in different manners. First, we will present the application of bioinformatics protocol on metagenomics data, with a particular emphasis on network analysis. In particular, we will use environmental and metagenomic data gathered during the Tara Oceans expedition to improve understanding of a biological process such as the carbon export. Second, we will describe how to integrate heterogeneous omics knowledge via logic programming. Such integration will emphasize putative functional units at the community level. Third, we will illustrate the design and the use of quantitative modeling from this network. In particular, constraint-Based modeling will predict the microbial community structure and its behaviors based on genome-scale knowledge.

Contents

Chapter 1

Introduction

For centuries, Biology was mostly a descriptive science. The first biological studies principally focused on describing living compounds. Such a description relied on acute observations to identify specific features for each organism. Naturalists then used these features to identify, to name, and eventually to classify living compounds based on their similarities. The generalization of this approach driven by observations seeks for a general architecture to describe Life, such as general phylogenetic trees or modular descriptions of organic elements. It is worth noticing that these great descriptive works were associated with intense archiving efforts that remain active until today by taking the form of up-to-date databases as promoted by the natural history museums. From this description of Life, fascinating studies attempt to link different organic forms. For instance, one aims at explaining the fundamental changes from one biological component behavior into another (i.e., the physiology that explains the change of state of a given living form). Also, one seeks to understand the differences between two distant living forms, promoting the rise of several evolution theories that still elude us, among which the famous one proposed by Charles Darwin [41]. Both questions propose to infer dynamical properties between observations. These properties presume the existence of mechanisms or general laws inspired by Chemistry or Physics that, as a biologist, one aims to discover.

To reach this goal, Biology increases its explanatory power by improving observation techniques. Along with this technical evolution, it is worth noticing that Biology also changes of central paradigm by empowering other Sciences. Indeed, Biology started focusing on the organism description, in which case organisms and their structure where the unit of Life (i.e., comparative anatomy). Following improvements in optics, Physics promoted the general use of microscopes to investigate organic components, which led to the identification of cells. This general iden-

tification advocated for the cell as a central paradigm of Life (i.e., embryogenesis, immunology, a new classification of organisms based on organelles) until the mid of the twentieth century. At this date, scientists deciphered the molecular bricks that compose a cell. Among other seminal works, one could mention the work of the RNA tie club consortium that built upon the discovery of DeoxyriboNucleic Acid (DNA) to break the code of proteins. Since then, one generally considers DNA (and assimilated) as the central paradigm of Life that promptly conducted to the rise of the concept of modern genes when Jacques Monod, François Jacob, and Andre Lwoff mixed the molecular concept with the biological inheritance theory, as proposed by Gregor Mendel 70 years earlier [150].

Considering this evolution of paradigm is of significant interest for this work. It is worth noticing that the paradigm of Life is getting more conceptual over time. Rapidly, the concept of model emerges in Biology, first for the sake of generalization, and more recently for the sake of investigation. In particular, a gene, the current state-of-the-art biological model, is not observable per se, but its abstraction/formalization is getting more precise following the accumulation of experimental knowledge. Conversely, the design of new experiments aims at refining existing models. All these points concur in considering Biology as a science of models. This statement is particularly crucial for the following if one acknowledges that not a sole, but several abstractions are necessary to cover the full diversity of organic forms. It is therefore not surprising that, beyond the interest of understanding the origins and mechanisms of Life, Biology inspired several formal sciences driven by different modeling abstractions.

1.1 Biology & Formal Sciences

The formalization of Life is an arduous task. An extensive review of the literature is beyond the objective of this summary. For the sake of humility, the sequel will describe only a few critical studies that motivate or inspire this manuscript and associated works. The chronology of these studies remains partial and depicts a personal interpretation that would require further investigations to give justice to the scientific field and its implications. Physicists pioneered the introduction of formal concepts to tackle biological questions. For instance, Erwin Schrödinger proposed a classical physicist's approach by emphasizing *"how can the events in space and time which take place within the spatial boundary of a living organism be accounted for by physics and chemistry"* [180]. This seminal work introduced several concepts that motivate the application of statistical physics to tackle biological questions. In particular, his study analyzed biological objects/abstractions that one barely described at the time, such as chromosomes. The work of Erwin

Schrödinger was crucial to introduce the concept of self-organization and the use of entropy to explain biological features. Fascinatingly, Sir Alan Turing proposed, almost ten years later, another and complementary analysis of biological emerging properties by focusing, this time, on discrete abstractions and the impact of rules to simulate feedback loops [203]. This study established the concept of automaton for biological systems, and from a personal viewpoint, computational biology in general. Among others, these formal concepts inspired several biological theories. One of the most famous was maybe the concept of operons proposed by Jacques Monod, François Jacob, and Andre Lwoff. Already mentioned above, it is worth noticing that this mathematical modeling settled the modern concept of genes as formal objects with a functional purpose. To these days, operon and genes represent abstractions that are still regularly used by the experimental biologists, even if no one *saw* a gene per se – but its implication on different experiments. Notice also that Jacques Monod provided after the second world war another modeling based on ordinary differential equations to simulate the bacterial growth [150]. This modeling represented a simulation of biological quota variations over time and helped to design an experimental device, called chemostat, still used today [151].

Searching for self-organized structures The formalization of Life promoted a change of paradigm in biological studies. The use of modeling allows spanning multiple biological compounds (i.e., quantities or events) that one could integrate to examine collective behaviors. Such behaviors, as proposed by above the seminal works, produce an emerging property. Formally, one stated this property as being the resultant of the asymptotic behavior of the modeled system. In 2008, Eric Karsenti reviewed the implications of this (fruitful) biological paradigm when applied on molecules and the shapes of cells [107]. In particular, one should notice the joint use of mathematics and physics, magnified by computational simulations. Thus, as a consequence of this paradigm, different observations of the same biological entity are the result of bifurcation between different states. Bifurcations are the issue of (non-linear) collective behaviors that include elementary (physical) rules, such as feedback loop or diffusion processes. From a computer science viewpoint, one can discretize these processes for the sake of calculability. In particular, Stuart Kauffman, in his crucial book [110], discussed several implications of discrete events to simulate self-organized biological systems. Worth noticing here, this biological paradigm settles several studies that described biological architectures as a result of collective molecular behaviors, thus linking two critical biological concepts: evolution and biological developments such as embryogenesis (for examples, see [31]).

A multi-layered system Following the rise of the above biological concepts, biological sciences made a significant effort to increase the acquisition of experimental data. In the 1990s, large scale sequencing projects were set up and received massive fundings. The most popular one remains the sequencing of the human genome (i.e., identification of the genes and their organization over the different chromosomes). However, similar projects took benefits of high-throughput biotechnology developments to compile comprehensive catalogs of cell contents with gene expression, proteins, and metabolic reactions. Thus, in addition to genomics, one saw the emergence of projects in transcriptomics, proteomics, metabolomics, which settled a new biological discipline called *omics*. All omics projects focus on a given biological scales and require specific modelings and data analysis. Altogether, these descriptions contribute to defining biological systems as multi-layered, which increases, as well, the complexity of biological system descriptions compared to physical or chemical system standards. However, such new complexity is, maybe, along with the size of the modern biological dataset, one of the prime motivations for empowering computer sciences in biological studies (see Chapter 2 for further details).

1.2 Systems Biology: a definition

Among the above theories, at the interface of computer science, mathematics, and molecular biology, one saw the emergence of sub-disciplines. For the sake of illustration, Figure 1.2 describes them and their links with the above general concepts. First, Bioinformatics supports the development of biotechnologies. This sub-discipline focuses mostly on the process and analysis of high volumes of information, such as provided by high-throughput experiments in genetics, genomics, transcriptomic, and, more recently, in metagenomics or metabolomics. Thus, Bioinformatics supports data acquisition. Second, Computational Biology is a sub-discipline that summarizes developments in theoretical biology and biological modeling. It deals with the development of models to study biological systems [95]. Both sub-disciplines overlap. For instance, from a retrospective viewpoint, the work of Jacques Monod belongs to both sub-disciplines, either by promoting modern genetics or advancing in the concept of feedback loops and formalizing operons. Similarly, the self-organization concept contributes to both by fostering the cellular behavior from molecular collective behaviors [107] or identifying mechanistic behavior from physical statistic laws [190]. Such overlap, often happening with new experimental datasets, gave rise to a new sub-discipline: Systems Biology. The main goal of this discipline is to automatically extract emerging properties from biological systems depicted by high-throughput experiments. This later

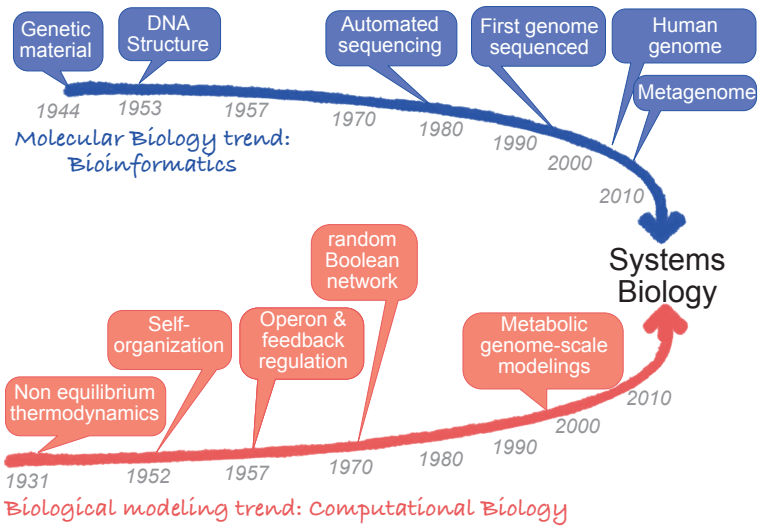


Figure 1.1: Evolution of molecular biology and biological modeling to create Systems Biology. The higher panel represents landmark discoveries in molecular biology that foster bioinformatics. The lower panel describes its pending in biological modelings later called computational biology. Both sub-disciplines combine to create the more recent Systems Biology. Figure adapted from [212].

sub-discipline takes an interest in system-level behavior of biological processes. The interaction of their molecular constituents describes such systems, conforming complex biological networks [113, 114]. Eighteen years ago, Hiroaki Kitano proposed a schema not denied since that articulates systems biology via four essential connected topics.

Conception of the biological system This topic consists of analyzing all species involved in the system, such as genes or proteins. The aim is herein to decipher species that one must consider as variables or constant within a model that will represent the system. Chapter 2 will further detail this topic.

Analysis of the biological system structure Variables are involved in a given referential and promote complex interactions that build a structure that is of interest. The complexity of this task relies on the necessity to deal with different referentials or scales such as genomic, transcriptomic, or metabolomic. This multi-scale nature implies the integration of diverse and heterogeneous variables within a unique modeling framework to capture the biological sys-

tem as a whole. This task, often called Integrative Biology, will be described in Chapter 3.

Simulating the dynamics of the biological system Once identified and integrated, one is interested in simulating the model to emphasize its emerging properties, such as dynamical behaviors for the sake of prediction. This task relies on formal abstractions employed in the previous tasks (e.g., the level of discretization for each variable). Chapter 4 presents different modelings and their practical use.

The control of the biological system The final task consists in controlling the system for the sake of *in silico* experiments. Whereas often neglected or associated with synthetic biology studies, one will resume this task in the chapter 4 for the sake of simplicity.

All these tasks are rarely published together but instead written via distinct studies. However, altogether, these tasks allow Systems Biology to identify the respective role of genes in the context of single-cell growth (see [20] for an illustration of our contribution), or integrate genome and metabolic networks to emphasize functional units that interplay at both experimental scales (see [16] for an illustration of our contribution). Achieving this goal is of biological interest but also challenging in Computer Sciences because biological knowledge is, by essence, incomplete and most heterogeneous. This interest and corresponding limitations will be discussed in the last chapter of this manuscript.

1.3 From Systems Biology to Systems Ecology

Systems Biology has been successful in analyzing cellular data sets at molecular scales, providing insights into underlying processes [113, 114]. Increasing computing capacity (i.e., data storage, computing time, but also formalization) and large dataset availability enabled Systems Biology to extend its application domain from small size reductionist networks to whole micro-organism systems [102] before considering metazoans [216, 148, 130]. However, despite the interest of studying sizable biological systems with different compartments, micro-organisms remained of primary interest. They are indeed the most diverse and abundant cellular life forms on Earth, with estimates from 25% to 50% of Earth total biomass [213]. Furthermore, one can not cultivate a majority of prokaryotes (>99%) in the laboratory [103], shoving the necessity to promote experimental expertise assessing *in situ* microbial diversity.

The rise of Environmental Genomics In parallel to genomics and bioinformatics, the last decade saw the rise of a new field in Biology, at the interface between Genomics and Ecology called *" Environmental Genomics"* [171]. One of its main challenges relies on understanding ecosystem behaviors and how microbial communities interact within their environment. Using biotechnological advances (i.e., high throughput DNA sequencing, RNA sequencing, or proteomics), one can today capture the whole microbial ecosystem composition and microbial behaviors, which represents a fantastic holistic viewpoint of ecological systems that take place in Nature. One will further discuss this point in the following chapters, but several fundamental ecological questions are today reachable. For instance, recent microbiology studies seek to understand how the presence of microbial communities could depollute water soils. Others show how one can engineer microbial communities as micro-factory, promoting synthetic ecology studies [58]. Others study the association of microbial communities in the Human gut with Human pathologies [83]. Overall, biotechnologies allow today to sample broader ecosystems at the gene level, targeting ecology and evolution of micro-organisms distributed around the world [109].

Motivations of this work Beyond the sole description of microbial communities that necessitates applications of bioinformatics techniques to isolate microbial genes within metagenomes, environmental genomics raises great computational questions. From a computational viewpoint, these questions consist of emphasizing ecosystem behaviors from complex microbial interactions, which relies on finding emerging properties from microbial ecosystems. These questions are similar to those proposed in Systems Biology, and one could be called the new subdiscipline applied on environmental genomics: *Systems Ecology* [117]. As previously demonstrated in Systems Biology, the role of Computer Sciences is herein, again, essential to support this new thematic, and beyond standard expectations (computing capacity, storage capacity), but rather by the need to formalize, automatic reasoning, mixing heterogeneous knowledge for analyzing interdependencies and inferring emerging properties. In this context, this manuscript advocates that one must pursue our systems biology efforts to design a modeling paradigm that will be suitable for studying ecosystems sampled at the gene level. Constraint Programming framework was of great help in systems biology, and one assumes here that it will propose guidance for dealing with environmental genomics. In particular, Constraint Programming is accurate to model partial or incomplete information [144]. Each piece of information is one constraint that concerns an investigated system. The set of constraints summarizes the system and delimits its solution space. Thus, the use of constraint programming techniques allows, via

optimization routines, either (i) to describe the set of solutions that satisfy all the constraints as previously stored, or (ii) to check whether a new constraint can be added to the set of constraints without modifying their overall consistencies (see previous work in systems biology [60] for an illustration). Once the problem modeled as a set of constraints, the computational elegance of such a paradigm relies on the use of state-the-art solvers that allow solving well-modeled problems rather than focusing on programming resolution techniques per se. Also, constraint-based models may be refined whenever additional biological knowledge becomes available, which allows one to make useful inferences even from partial and incomplete information. Therefore, constraint programming was considered to be an elegant modeling paradigm to face Systems Biology challenges as emphasized renown biologist such as B. Palsson (2000) in [54]: *"Because biological information is incomplete, it is necessary to take into account the fact that cells are subject to certain constraints that limit their possible behaviors. By imposing these constraints in a model, one can then determine what is possible and what is not, and determine how a cell is likely to behave, but never predict its behavior precisely."*.

This manuscript will present the adaptation systems biology techniques to environmental and ecological questions, with a strong emphasis on the use of optimization techniques on graphs and constraints. Three chapters, for each systems biology topic, constitute this summary. Computational implications and research perspectives will be discussed in the last chapter.

Chapter 2

Analysis of omics experiments

We cannot solve our problems with the same thinking we used when we create them

Albert Einstein

2.1 Introduction & Context

The biological data acquisition represents the cornerstone of many biological sciences for the sake of better biological knowledge. The generated data revert several aspects that must be stressed out herein. First, biological data are highly heterogeneous. As mentioned above, recent biotechnological progress allowed experimental scientists to drastically increase the focus of their investigation, which allows today to observe the biological systems at the molecular level via so-called omics data (i.e., mainly genomics, transcriptomics, metabolomics, proteomics). Omics data resume the state of a given biological system at a given condition. Analyzing these data from different states thus enables us to infer underlying molecular mechanisms. Second, the same biotechnological progress also proposed a significant increase in the quantity of data, which makes their analysis difficult or rebarbative without the use of dedicated algorithms. Considering this last point, it is therefore not surprising to consider biological data analysis as the primary scientific interface between biologists and more formal scientists. However, despite the new enthusiasm about it, it is worth to notice herein that the amount of biological data do not permit the qualification of Biology as a *"big data science"* per se. Indeed, biological data, even by considering the recent and sizeable genomic effort, do not revert, so far, the same magnitude as other sciences. For instance, Astronomy currently

proposes a data acquisition rate of 7.5 terabytes.s^{-1} when screening the space (i.e., Australian Square Kilometer Array Pathfinder (ASKAP) project) that is huge compared to one zetta-bases.$year^{-1}$ for the whole biological sequencing effort [187]. Nevertheless, despite a current debate (see [187] details), the rate of genomics data acquisition does not matter, mainly because the Big data attribute relies more on the complexity of heterogeneous biological data rather than on the quantity of data per se.

Biological data describe different abstractions of the multi-layered biological systems but are also of different types. From a historical viewpoint, the data acquisition protocol was primarily focusing on producing continuous data. The need to analyze this type of data explains the substantial impact of statistical analysis in biological sciences. Indeed, several tasks remain a routine for experimental biologists:

(i) comparing two biological datasets or a dataset with synthetic data produced by a statistical model,

(ii) finding structure(s) of interest within the dataset(s) via classification or clustering techniques,

(iii) understanding the data distribution compared to external parameters via, for instance, multivariate analysis.

An overview of these statistical approaches has been extensively discussed, in particular in ecology [129], where biological observations mainly concern populations which are, by essence, quantitatively measured. For the sake of ecosystems analysis, we performed several of these techniques: initially proposed in [209], we then extend this technique to more different molecular probes [22, 21], which will further used to study oil spill impact on microbial ecosystem [154]. When applied to cellular systems, the same techniques are of interest. For instance, we achieved similar analysis on transcriptomic knowledge (i.e., gene expression levels are semi-quantitative) [34], or at the proteomic level (i.e., the affinity of proteins are semi-quantitative) [145].

More recently, data acquisition protocols proposed to discretize the biological data. Thus, after normalization, gene expression levels have been discretized to build gene regulatory networks where genes are as Boolean variables that can be activated (i.e., +1) or repressed (i.e., 0) based on the activity of other genes [91]. The Section 4 will discuss the simulation of these networks. Whereas these data processing necessitates a discretization via the use of an expression threshold (above which the gene is considered as activated, not otherwise), other biological data are discrete by nature. For instance, DNA sequencing procedure outputs,

called reads, are small strings and represent the central biological knowledge produced so far in Biology [187]. For instance, sequencing a genome consists of cutting the genome sequence into smaller sequences, reads, that will be defined or sequenced by high-throughput sequencing techniques. As an output, a set of reads describes the genome sequence. The whole set of reads of the sequenced genome is composed of four letters (i.e., A, T, C, G) for each nucleotide and assumed to cover the whole original genome sequence [38]. Several bioinformatics protocols aim a reconstructing this original genome sequence. Most of them consist of assembling all reads by solving a general covering problem. The most modern and efficient techniques propose to assemble reads into another discrete structure called a de Bruijn graph [220, 51] that is a directed graph representing overlaps between sequences of symbols. Notice herein that, despite the discrete nature of this biological knowledge, its analysis remains difficult. For instance, comparing genomes is a difficult combinatorial problem, even by computing pairwise distances that revert the use of different metrics such as the number of transpositions or breakpoints (see [64] for review). We proposed the use of one of them, the common interval, to compare pairwise bacterial genome sequences [5].

2.2 Analysis of a graph that models a cellular system

The general use of discrete variables underlies the interest of discrete abstractions that are used to represent biological structures. In particular, the last decade saw the more extensive use of graphs to represent biological knowledge (see [7] for review). A graph is a discrete and formal abstraction that shows several interests in Biology. First, a graph is composed of nodes and edges. Based on the definition of its edges, a graph can be undirected (i.e., no direction on edges) or directed. One can also consider other edge attributes such as the use of weight on edges to produce a weighted graph or the use of weighted or color on nodes. These features promote a great expressivity to the graph abstraction for the sake of biological system modeling. This expressivity is in particular well-deserved by several programming languages dedicated to describing graphs and their representations (see [9, 119] for recent illustrations) It is worth to notice herein that graph expressivity is of particular interest for collaborating with biological collaborators. Indeed, most of the biological concepts, in particular in physiology, are resumed by graphs in textbooks. Graphs depict invisible biological system mechanisms or molecular descriptions of biological systems summarized for the sake of pedagogy. Thus complex datasets are interpreted and resumed, which makes the graph one of the favorite abstractions for biologists - and illustrates as well the fact that Biology is a modeling science - where students must learn state-of-the-art models rather than

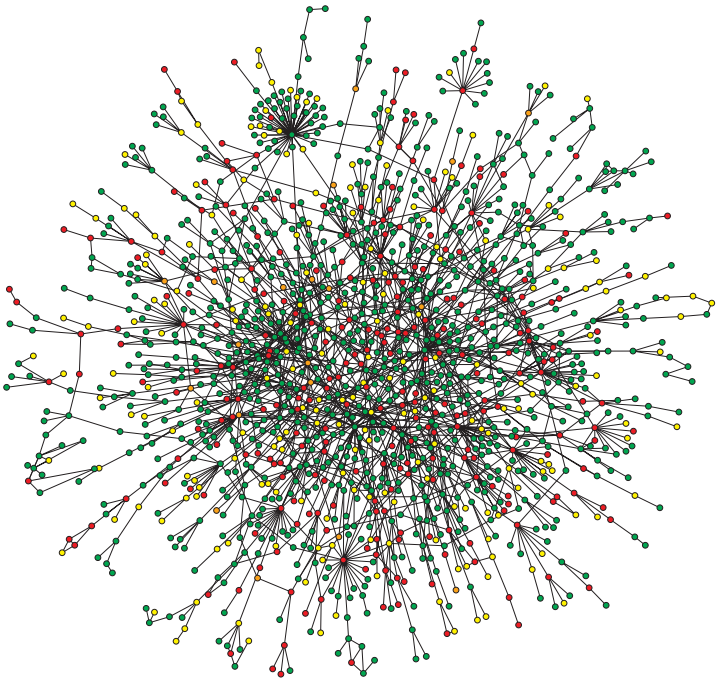


Figure 2.1: Illustration of a biological graph: yeast protein interaction network from Barabási and Oltvai [7]. This picture maps the main component of protein-protein interactions from *Saccharomyces cerevisiae*. Each node represents a protein, and edges represent interactions between two proteins as estimated by the first two-hybrid measurements in yeast. Accumulation of interactions describes a graph or called network by biologists. The node color indicates additional knowledge: red is a lethal protein, green is nonlethal, orange depicts a protein associated with slow growth, and yellow pinpoints no biological feature associated. Such a representation commonly called *hair-ball* shows the limit of a graph representation for the sake of biological investigation, which emphasizes the need for a dedicated analysis.

complex datasets. Also, the use of graph pinpoints the need to incorporate diversity or uncertainties within a formal and discrete description. For instance, the de Bruijn graph summarized several short reads by allowing sequencing mistakes or to consider single nucleotide polymorphisms [97].

Second, the use of graph abstraction allows the use of a large number of techniques to analyze and compare graphs (see [7] for review). For instance, these

techniques describe global or local properties, ability to cross the graph. Historically, several of these techniques have been applied to Protein-Protein Interaction networks (PPI); see Figure 2.2 for illustration. These graphs, built from experimental procedures, represent putative associations between different proteins that interplay (see Figure 2.2a). Interactions are determined based on pairwise physical and chemical protein properties, such as covalent bonds or stable or transient interactions. The set of interactions describes putative complex assemblies (see Figure 2.2b). Today several databases store PPI networks for several organisms, among which the renown STRING (see [191] for latest release). Early 2000's, Jeong and collaborators [100] pioneered the PPI network analysis by promoting the use of graph decomposition techniques. As a follow-up, Gagneur and coworkers [77] proposed the application of hierarchical decomposition of a graph called modular decomposition. This algorithm emphasizes sub-structures within the PPI, associated with protein complexes, via the use of a tree. By integrating the ComBi team in Nantes, Géraldine Del Mondo was a co-advised Master student with Irena Rusu on a similar subject. We advocated for a natural extension of hierarchical decomposition studies via the use of another algorithm called *modular decomposition*.

Géraldine Del Mondo, Damien Eveillard, and Irena Rusu. Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions. *Bioinformatics (Oxford, England)*, 25(7):926–932, April 2009

The homogeneous decomposition is a natural extension of the modular decomposition but remains more computationally challenging. Both modular and homogeneous decomposition proposes to build a decomposition tree, whose leaf nodes are proteins and internal nodes (called modules) are logical rules to combine the child nodes. Logical rules are herein the use of $\wedge$ (ex. $A \wedge B$) to describe the fact that protein A and B must be present within a module, or the use of $\vee$ (ex. $A \vee B$) to depict a selective choice between A or B to describe a module. Thus, once the graph is broken up into modules, one must be able to build the graph again by using only its module and the logical rules. Compare to the modular decomposition (see Figure 2.2c), the homogeneous decomposition further decomposes modular modules. In particular, we identify a new structure called W-graph via the use of new logical rules (see Figure 2.2d). When applied to realistic PPI, W-graph is of interest. This structure represents a hub that describes interplays between modules (or protein complexes). In particular, it shows that a protein could belong to several complex alternatively, emphasizing putative dynamics or regulation feature within the PPI. This new description thus, while being computationally challenging, allows describing hidden substructures within PPI datasets, in a similar manner than

multivariate analysis when working on continuous measurements.

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 7 2009, pages 926–932
doi:10.1093/bioinformatics/btp083

Systems biology

Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions

Géraldine Del Mondo, Damien Eveillard and Irena Rusu*

Computational Biology group (ComBi) - LINA, Université de Nantes, CNRS UMR 6241, 2 rue de la Houssinière, 44300 Nantes, France

Received on March 28, 2008; revised and accepted on February 10, 2009

Advance Access publication February 17, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Modules in biology appeared quickly as an accurate way for summarizing complex living systems by simple ones. Therefore, finding an appropriate relationship between modules extracted from a biological graph and protein complexes remains a crucial task. Recent studies successfully proposed various descriptions of protein interaction networks. These approaches succeed in showing modules within the network and how the modules interact. However, describing the interactions within the modules, i.e. intra-modular interactions, remains little analyzed despite its interest for understanding module functions.

Results: We overcome this weakness by adding a complementary description to the already successful approaches: a hierarchical decomposition named homogeneous decomposition. This decomposition represents a natural refinement of previous analyses and details interactions within a module. We propose to illustrate these improvements by three practical cases. Among them, we decompose the yeast protein interaction network and show reachable biological insights that might be extracted from a complex large-scale network.

Availability: A program is at disposal under CeCILL license at: www.lina.univ-nantes.fr/combi/DH/Home.html

Contact: irena.rusu@univ-nantes.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

By essence, biological systems are not fully explained. They appear as complex systems which emphasize our incapacity to understand the relation between inputs and outputs of the living system (Szallasi *et al.*, 2006). Evidence of modules in biology and utilities of such a concept quickly appeared as an accurate way to summarize complex living systems with simple ones. As an illustration, a molecular complex abstracts *numerous and complex interactions of proteins*. Using a module description implies to replace some part of the system with an abstraction that maintains a correct property with the given experimental data. This modeling approach introduces the concept of *modularity* such as it was expressed clearly by Hartwell *et al.* (1999). Applied on protein interaction networks, these *top-down approaches* emphasize molecular hubs or *functional components* within the network (Szallasi *et al.*, 2006). In other words, they find information of interest within a graph structure while describing its modules.

Many studies aim at discovering this kind of information within the structure of biological graphs (Jeong *et al.*, 2000). In particular, Spirin and Mirny (2003) give a strong support in such protein interaction network analyses. They show theoretical modules extracted from the network that correspond to protein complexes (splicing machinery, transcription factors, etc.) or dynamic functional units that can belong to the cell-cycle regulation. However, *in silico* techniques appear as very sensitive to the completeness of the protein interaction set. Finding biological modules is only efficient for already well-investigated protein–protein interaction graphs.

Due to intensive experimental investigations on *Saccharomyces cerevisiae*, the yeast protein interaction graph agrees with such a criterion. Thus, this graph quickly appeared as an accurate benchmark for testing protein interaction network analysis techniques (Guimerà *et al.*, 2004; Hart *et al.*, 2007; Ma *et al.*, 2004). Based on tandem affinity purification/mass spectrometry (TAP-MS) experiments (Puig *et al.*, 2001), various techniques aim at characterizing protein complexes [see Gavin *et al.* (2002, 2006); Ho *et al.* (2002) and Krogan *et al.* (2006) for illustration]. Among them, Hart *et al.* (2007) used the Markov Cluster Algorithm (MCL) technique developed by Enright *et al.* (2002). This is a statistical scoring-based approach that differentiates direct physical interactions from interactions mediated by other members of the complex. Based on various experiments, the authors emphasize the relevance of combining statistical analyses for inferring biological knowledge. In practice, their study clearly indicates a hierarchical organization of protein complexes in the cell and confirms that the yeast '*complex-ome*' is almost fully described.

In this context, protein complexes act as biological modules and the method gives a robust overview of how biological modules interact. Nevertheless, it also highlights other interesting questions about the impact of specific interactions in the modular description (He and Zhang, 2006), which is particularly relevant. However, the MCL technique fails to answer these questions intuitively.

To overcome this technique weakness, we follow the assumptions of Hart *et al.* (2007). We consider the yeast protein interaction network as a graph with an (almost) complete set of protein interactions. Consequently, each *unit* protein complex (that is, not including smaller protein complexes) appears as a fully connected

*To whom correspondence should be addressed.

part of the network. Note here that the reverse is not true: not every fully connected part of the network necessarily derives from an existing (unit or not) protein complex. Our approach aims at discovering *in silico* information that is hidden within the protein interaction network. It identifies unit protein complexes and the relations between them. *Hierarchical graph decompositions* present interesting features for describing complete and large-scale graphs such as the yeast protein interaction network. Therefore, we consider these decompositions as a natural theoretical framework for refining the description of protein complexes (equivalently, biological modules) obtained using the MCL technique. Following a similar assumption, Gagneur *et al.* (2004) apply the hierarchical decomposition named *modular decomposition* on protein interaction graphs. Various tests of this method show that theoretical modules obtained this way may correspond to protein complexes. They show as well that modular decomposition is not precise enough to capture several important features of biological systems. One main drawback of the modular decomposition is the existence of too large components when the modular decomposition is finished at all levels. Therefore, as observed using the MCL technique, important relationships between intra-modular components remain hidden in the network analysis.

Notwithstanding, we consider that the assumptions exposed by Gagneur *et al.* (2004) are convincing. Further investigations using hierarchical graph decompositions might complete the MCL results and show intra-modular interactions. We herein propose to extend the analysis involving modular description by using a natural refinement of the method, called *homogeneous decomposition*. We precisely explain that this decomposition improves the network partitioning (see Section 2). It hence allows us to go further into our biological purposes by (i) identifying smaller significative components of the network and (ii) showing up their detailed interactions with the other components. We illustrate these theoretical features by an application on various protein interaction networks (see Section 3). We first describe results on a theoretical protein interaction network. It shows various modular insights, emphasized by the homogeneous decomposition (see Section 3.1). Second, we illustrate the improvements obtained with homogeneous decomposition on known complexes, the transcriptional regulator complexes, already analyzed (Gagneur *et al.*, 2004) through modular decomposition (see Section 3.2). Finally, we apply the homogeneous decomposition on the yeast protein interactions (see Section 3.3) that represents an accurate realistic benchmark supporting our method.

2 METHODS

A *graph* is a data structure used for representing objects and their pairwise relationships. The objects are the *vertices* of the graph, while the pairwise (undirected) relationships between objects are the *edges*. We herein assume a protein–protein interaction network as a graph whose vertices and edges are, respectively, the proteins and their pairwise interactions. The structure of this graph provides a lot of information on the groups (or complexes) of proteins that act together to fulfill a biological function. To *discover* the organization of a graph, one usually uses *graph decompositions*. To *store* and *analyze* this organization, one uses *decomposition trees*.

Graph theory provides various ways to decompose a graph. Many of them are *single-level decompositions* because they partition the graph into several components that are not partitionable themselves. In contrast, *multi-level*

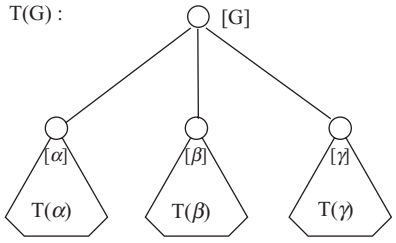


Fig. 1. Decomposition tree of an arbitrary graph G with modules α, β and γ.

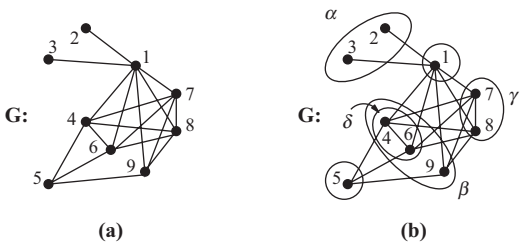


Fig. 2. (**a**) Example graph G and (**b**) its decomposition into modules $\alpha, 1, \beta, 5$ and γ.

(or hierarchical) decompositions have the major advantage of allowing an iterative study of the structure by fitting the components into each other.

The modular decomposition, also known as *substitution* (Möhring and Radermacher, 1984) or *X-join* (Habib and Maurer, 1979) decomposition, is probably the most well-known hierarchical decomposition of graphs. It was independently discovered several times [see Möhring and Radermacher (1984) for a review], and various very efficient algorithms [see for instance, McConnell and Spinrad (1994)] exist to compute it. As a natural refinement of the modular decomposition, Jamison and Olariu (1995) propose the homogeneous decomposition, for which Baumann (1996) describes an efficient algorithm. Both modular and homogeneous decompositions build a *decomposition tree* (See Fig. 1 for illustration), whose leaf nodes are proteins and whose internal nodes (called *modules*) represent *logical rules* to combine the child nodes.

Modular and homogeneous decomposition are explained below and illustrated on the graph G in Figure 2a. In this description, the notion of *adjacent vertices* (or simply *neighbors*) is paramount. It designates two vertices of G joined by an edge. A graph whose vertices are all neighbors to each other is called a *clique*.

2.1 Modular decomposition

Under the modular decomposition model, a module M is a graph inside the given graph G so that all the vertices in M have exactly the same neighbors outside M (let us call that the *neighborhood property*). The aim of the modular decomposition is to decompose a graph into non-trivial modules (at least two), and then to iterate the decomposition process on the resulting modules until all modules are made of one vertex (such modules are the leaves of the decomposition tree). Thus, the children of each node in the modular decomposition tree (Fig. 1) are its modules, whether they are internal vertices or leaves. Figure 2b shows one decomposition of G in Figure 2a into modules. Module β may be furtherly decomposed into modules with vertex sets $\{4, 6\}$ and $\{9\}$, while all the other modules have only *trivial decompositions* into 1-vertex modules (i.e. leaves).

When a graph is broken up into modules, one must be able to build the graph again using only its modules and a *logical rule* (otherwise the decomposition looses information). The logical rule is stored in the node of the tree corresponding to the graph as a character with values 0, 1 or P as

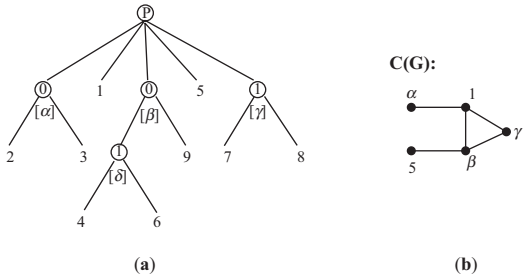


Fig. 3. The modular decomposition tree of the example graph G (**a**) and the characteristic graph $C(G)$ associated to the root (**b**), which is a P-node.

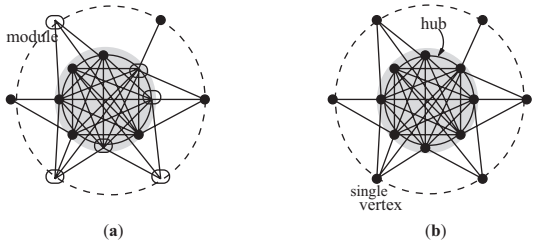


Fig. 4. (**a**) Wheel structure of a graph. (**b**) Characteristic graph with emphasized hub and single vertices. The dotted circle simply highlights the wheel structure, it does not indicate edges.

follows (see Fig. 3a for illustration):

- 0 means that G is the *union* of all its modules, without any edge joining vertices from different modules. In this case, G is a ⓪-*graph* (or ⓪-*module*) and its corresponding node is a ⓪-*node*. Modules α and β in Figure 2b are ⓪-modules.
  Biological interpretation: Gagneur *et al.* (2004) explain a ⓪-node as an alternative between its children: any of them successfully replaces the module G in the ⓪-node in any operation involving G.
- 1 means that G is the *join* of all its modules, obtained by adding an edge between every pair of vertices from two different modules. In this case, G is a ①-*graph* (or ①-*module*) and its corresponding node is a ①-*node*. Module γ of G and module δ with vertex set $\{4,6\}$ of β are ①-modules.
  Biological interpretation: following the parsimonious interpretation in Gagneur *et al.* (2004), a ①-node requires that *all* its children combine together to replace module G in a ①-node in any operation involving G.
- P means that G is obtained from its modules by performing, between any pair of modules, either a union or a join, according to the *characteristic graph* $C(G)$, whose vertices correspond to the modules, and whose edges correspond to the join operations. In this case, G is a P-*graph* (or P-*module*) and its corresponding node is a P-*node*. The whole graph G in Figure 2a is a P-module with modules $\alpha, 1, \beta, 5, \gamma$ (Fig. 2b), which have to be combined together according to the characteristic graph $C(G)$ in Figure 3b to build G again. Notice here that the characteristic graph of G is obtained by shrinking in G each module into a single vertex.
  Biological interpretation: there is no reasonable biological interpretation for a P-node, since such nodes may correspond to graphs which are very different one from another, and which may be very large [see, Gagneur *et al.* (2004) for example].

The uniqueness of such a decomposition (and thus of the interpretation one obtains when it is applied on a graph) is guaranteed by several simple rules you must apply when decomposing. To name a few, one must make sure that there are no edges between ⓪-nodes or between ①-nodes in the decomposition tree.

2.2 Homogeneous decomposition

The main drawback of the modular decomposition is its unability to further decompose the characteristic graphs associated to the P-modules. The homogeneous decomposition partially solves this problem by identifying P-modules with a specific structure, for which a further decomposition is proposed. It is not meant to *replace* the decomposition into modules but to *refine* it, once the modules have been computed and the characteristic graph has been built. The homogeneous decomposition offers, therefore, an obvious *qualitative* improvement to the modular decomposition, which is described here in an intuitive manner and is illustrated in examples. Notice here that we slightly modified the definition of the homogeneous decomposition tree

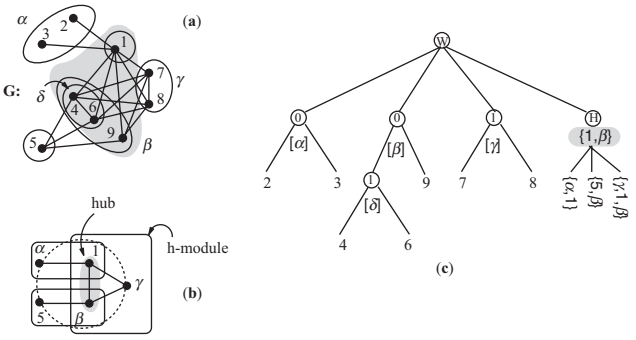


Fig. 5. (**a**) Wheel structure of G. (**b**) Characteristic graph $C(G)$ with hub (in gray) and its three h-modules. (**c**) Homogeneous tree of G.

compared with the original one by Jamison and Olariu (1995), so as to make it easier to handle and to explain.

A graph G furtherly decomposable by an homogeneous decomposition is called a W-*graph* (or W-*module*) in the remainder of the article. It has the *wheel* structure depicted in Figure 4a. Such a module has a characteristic graph (Fig. 4b) made of a *hub* (which is a clique) and of a set of *single* vertices around the hub that have neighbors in the hub but are not joined to each other. The *h-modules* of a W-module are the graphs (which are also cliques) made of a single vertex and all its neighbors in the hub. Note that h-modules do not have the neighborhood property as other modules do. These notions are illustrated on the example graph G in Figure 5a and b.

The homogeneous decomposition introduces two new logical rules in the decomposition tree (Fig. 5c), described by characters W and H used to label internal nodes:

- W means that the graph G is obtained from its modules (stored as children) and its h-modules (stored as a specific child which is an H-node) by recovering a wheel structure. This happens when G is a W-*graph* (or W-*module*) and in this case its corresponding node is called a W-*node*. Graph G in Figure 2a is a W-module whose corresponding W-node is shown in Figure 5c.
- H means that the internal node stores the h-modules of its father, which is necessarily a W-module, in the following form: each h-module is stored in a child labeled by a vertex set whose first element is the single vertex identifying the h-module and the other elements are its neighbors in the hub. In this case, the node is called an H-*node*. Although this is not necessary, in our figures and so as to simplify explanations, the vertex set of the hub labels the H-node. The h-modules of the characteristic graph $C(G)$ of the example graph G, identified in Figure 5b, are stored in the H-node in Figure 5c.

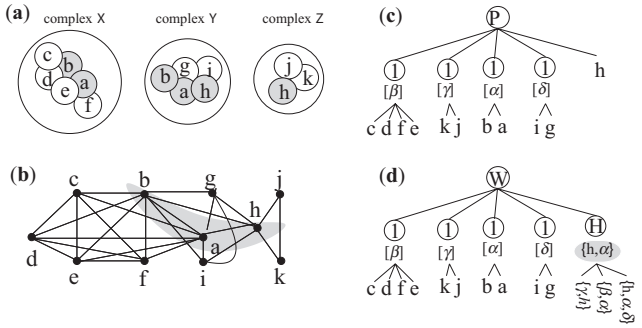


Fig. 6. Homogeneous decomposition of a theoretical protein interaction network from Wilhelm *et al.* (2003). (**a**) Protein complexes elucidated with TAP-MS. (**b**) The associated interaction network, where all possible interactions within the protein complexes are considered. (**c**) The corresponding modular decomposition tree. (**d**) The homogeneous decomposition tree that highlights modules and their intra-modular interactions. The gray area indicates the hub, stored in the H-*node*, that connects the other *h*-modules as leaves.

Biological interpretation. W-and H-nodes are always explained together. The *h*-modules stored in the H-node as well as in the hub itself (if it is not already included in an *h*-module) yield protein complexes. Its composition is described precisely through a careful interpretation of their vertices representing modules. In Figure 5c, the *h*-module with vertex set $\{\alpha, 1\}$ yields protein complexes $\{2, 1\}$ and $\{3, 1\}$, since module α is a ⓪-module. Therefore, proteins 2 and 3 are alternatives. The *h*-module with vertex set $\{\gamma, 1, \beta\}$ contains the hub, thus, the hub itself does not generate specific protein complexes. After the interpretation of modules γ and β, the protein complexes generated by $\{\gamma, 1, \beta\}$ are $\{7, 8, 1, 4, 6\}$ and $\{7, 8, 1, 9\}$.

3 RESULTS

Like many tools in many fields, modular and homogeneous decompositions might show very useful insights when applied in the appropriate context. The appropriate context, in this case, is an (almost) complete protein interaction network, where each unit protein complex is represented as a clique, due to the method used to infer the protein interaction network. Note here that exceptions to this constraint may either seriously or weakly damage the decomposition, depending on the nature of the exception. Network analysis aims at finding (i) the hierarchical structure of the network, (ii) the relations between complexes (inclusion, disjonction, overlapping), and eventually (iii) the proteins or groups of proteins within the network that play a central role in specific regions of the network. In this purpose, the decomposition tree has to clearly represent the complexes and highlight their relationships, which gives emphasis to the interpretation of each type of node in the decomposition tree.

3.1 Theoretical protein interaction network

We propose to illustrate the above notions on the small simplistic protein complex network shown by Wilhelm *et al.* (2003) (Fig. 6). This network comes from TAP and HMS-PCI techniques. Once the modules α, β, γ and δ are identified by modular decomposition, a wheel structure appears, with hub $\{h, \alpha\}$ and three *h*-modules. In practice, the hub does not represent a concrete protein complex (since the hub is part of the *h*-module $\{h, \alpha, \delta\}$), but shows the central role of its components in the intra-modular interaction description. Each *h*-module (which is a clique) yields one or more intra-modular complexes. As an example, the *h*-module $\{h, \alpha, \delta\}$ yields the larger clique $\{h, a, b, i, g\}$ after interpretation of modules α and δ. This clique, obtained by our theoretical approach, correctly identifies the complex *Y*. Similarly, the interpretation of the two other *h*-modules $\{\gamma, \text{h}\}$ and $\{\beta, \alpha\}$ conduces to recover cliques with vertex sets $\{k, j, h\}$ and $\{c, d, f, e, b, a\}$, that respectively correspond to the complexes *Z* and *X*. Therefore, our theoretical approach perfectly identifies, on this network, the known complexes, thus showing a great accuracy. Moreover, the homogeneous decomposition shows the relationships between these complexes. For instance, complex *X* and complex *Y* share the module α, that is the pair of proteins $\{a, b\}$. Despite the fact that they do not build a proper complex, these proteins always have to be considered together from a functional viewpoint (since α is a ①-module).

3.2 Transcriptional regulator complexes in yeast

We qualitatively validated our approach by comparing to the results of Gagneur *et al.* (2004), which used protein interactions from TAP-MS studies to define transcriptional regulatory complexes in yeast. It is composed of five complexes [see Cairns *et al.* (1994, 1996); Henry *et al.* (1994) and Kim *et al.* (1994) for details]: RSC, SWI/SNF (chromatin-remodeling complexes), TFIIF, TFIID (general transcription factor complexes) and Mediator (the mediator complex that mediates signals to RNA polymerase II). The modular decomposition tree (Fig. 7a) identifies several modules but fails to identify the relations between the children of the P-node.

In contrast, the homogeneous decomposition tree (Fig. 7b) singularly identifies a W-module that is structured as a wheel, according to the information stored in the H-node. The hub is composed by the protein Anc1 and the module α formed by Arp7 and Arp9. It plays a central role in the decomposition, since it generates all the interesting cliques in the network. This observation, exclusively based on our decomposition of the network, meets the biological knowledge, since either Anc1 or the module α belong to the five complexes experimentally known (Fig. 7c). Further analysis argue for the reliability of the homogeneous decomposition *and* our interpretation of it. The decomposition indicates three children of the H-node that correctly identify *all* the complexes of the network and their interactions. Each complex is herein associated with the *P*-value obtained after a Gene Ontology analysis using GO::TermFinder (Boyle *et al.*, 2004):

(1) The rightmost child of the H-node represents an interaction between Anc1, as a component of the hub, and the module β that is a ⓪-module. Consequently, Anc1 interacts alternatively with the components of β. This interpretation emphasizes three alternative cliques: either Anc1, Taf40, Taf19, Taf45, Taf61, Taf90, Taf47, Tsm1, Taf25, Taf67, Taf17, Taf60 (100% cluster frequency; i.e. how many genes from the clique are annotated to the GO term associated with the cluster; $P\text{-value} = 7.76e^{-28}$); or Anc1, Tfg1, Tfg2 (100%, $3.60e^{-10}$); or Anc1, Med4, Gal11, Nut2, Nut1, Med2, Med6, Med7, Pgd1, Cse2, Med8, Med11, Dmc1, Srb5, Srb4, Srb7, Srb6, Rgr1, Sin4, Srb2, Rox3 (90.5%, $5.08e^{-50}$). These three cliques correspond, respectively, to TFIID, TFIIF and Mediator complexes. Based on the TAP-MS experiments,

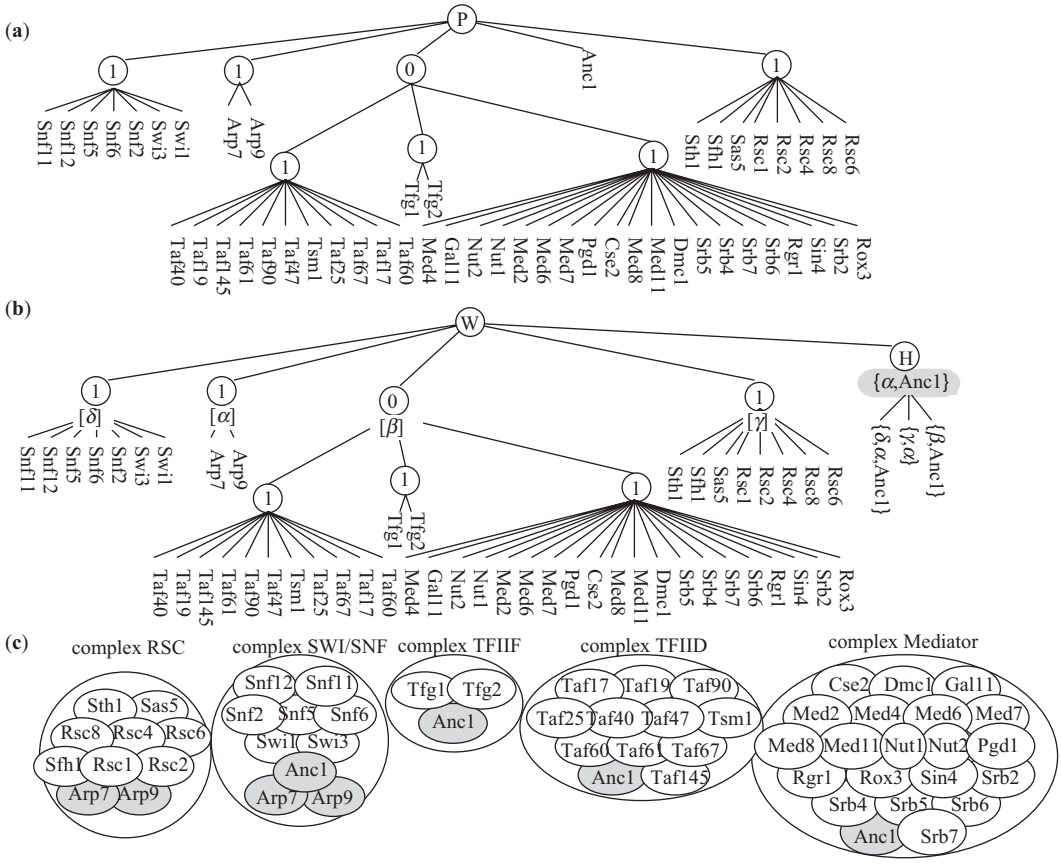


Fig. 7. Hierarchical decompositions of the transcriptional regulator complexes network in yeast. (**a**) An output of the modular decomposition. (**b**) An output of the homogeneous decomposition. (**c**) The experimental knowledge about the known complexes [adapted from Gagneur *et al.* (2004)]. The hub, i.e. H-*node*, and its components are highlighted in gray.

our decomposition predicts an alternative role for these complexes.

(2) The center child represents an interaction between the module α, as a component of the hub, and γ. The corresponding clique corresponds to the complex RSC (90%, $3.94e^{-23}$).

(3) The third leaf indicates interactions between δ, α and Anc1, which is interpreted as a clique that yields the SWI/SNF complex (100%, $1.50e^{-29}$).

These complexes are already obtained by using the MCL analysis on TAP-MS data (Hart *et al.*, 2007). However, such a statistical-based analysis does not show a precise description of the complexes and their sub-complexes. The homogeneous decomposition achieves to see the relations between their sub-complexes, just by taking a glance at the homogeneous decomposition tree.

The particular feature of the decomposition is the hub. It is composed by the module Arp7-Arp9 (namely α in Fig. 7b) and Anc1. A module like this one, not further decomposed, represents a sub-complex that is related to RNA polymerase II transcription factor activity (100%, $7.42e^{-6}$) and confirmed by experimental studies. Recently, Chen and Shen's (2007) experiments show that Arp7 and Arp9 compose a crucial subunit of the SWI/SNF complex. Moreover, Szerlong *et al.* (2003) demonstrate that Arp7 and Arp9 form a stable heterodimer with the properties of a functional module. In particular, they emphasize its impact in both restructuration of chromatin and interactions between transcriptional regulatory complexes, like SWI/SNF and RSC.

Anc1 is the other component of the hub. Anc1 connects three modules to the hub, corresponding to TFIID, TFIIF and the Mediator complex. Following the decomposition tree interpretation, these complexes are alternative. This interpretation implies that Anc1 plays a major regulatory function by interacting with either TFIID, TFIIF or the Mediator complex. Kabani *et al.* (2005) confirm such an assumption. Experimental evidences indicate indeed that Anc1 is the only non-essential subunit of TFIID. It is also associated with TFIIF, although it is not required for its activity. Anc1 thus appears as not really essential for the proper functions of complexes despite its overall importance on the whole network [based on its gene-deletion impact (Giaever *et al.*, 2002)].

The three proteins that compose the hub are interacting with another module named δ. It represents the complex SWI/SNF. The interpretation of the decomposition tree indicates that this last complex is functionally independent from other modules, but might be modulated by the combination of Anc1, Arp7 and Arp9. Interestingly, Kabani *et al.* (2005) indirectly confirm this assumption by not showing a clear interaction of Anc1 alone with the SWI/SNF complex. It intuitively implies the need of another component that

we assume, based on the decomposition tree interpretation, being the catalytic subunits of Arp7 and Arp9.

To sum up, the homogeneous decomposition emphasizes complexes that are in accordance with those observed using others techniques on TAP-MS data. These complexes, not depicted by the modular decomposition, present accurate statistical results when compared with biological functions and cellular compounds accessed via Gene Ontology information (see Supplementary Material). As additional features, it indicates how complexes interact using an hub. The function of proteins that belong to the hub is to connect complexes, which provides their regulation. In particular, experiments confirm the role of Anc1 as a regulatory function by modulating the activity of the respective catalytic subunits of the complexes mentioned above. From a topological viewpoint, the homogeneous decomposition identifies, in an optimized manner, proteins that possess the higher degree within the network, like those investigated by independent studies (Zotenko *et al.*, 2008).

3.3 A large-scale network: the yeast protein interactome

Previous examples show the qualitative accuracy of informations extracted by homogeneous decomposition, since the complexes identified *in silico* correspond to the already known biological ones. We propose an application on a more prospective network where the structure of the protein interaction network [that is, its (sub-)complexes] remains unknown.

Hart *et al.* (2006) decompose the large-scale network of the yeast using the MCL technique on TAP-MS experiments. They obtain 390 clusters or protein complexes, disjoint from each other. Since several clusters are too complex to be investigated in a precise manner, the homogeneous decomposition appears as a natural complement of the MCL technique. Indeed, the decomposition describes relationships within complexes emphasized by the MCL technique. As an illustration, we investigated the interactions inside each of the 103 complexes that present at least four proteins (see Supplementary Material). Among them, 61 complexes are too simplistic to require a homogeneous decomposition (they have no P-node), 21 complexes show identical modular and homogeneous decompositions and 21 complexes present an homogeneous decomposition that refines the modular one. Among these complexes, the modular decomposition identifies 52 sub-complexes, including 25 that give significant GO results (in average P-value $= 3.35e^{-04}$ and 69.9% of cluster frequency) when investigated with function ontologies (Boyle *et al.*, 2004). As evidences of a gain over the modular one, the homogeneous decomposition shows 75 sub-complexes, including 44 that show significant P-values after a GO analysis (in average P-value $= 8.44e^{-05}$ and 81.98% of cluster frequency, see Supplementary Material).

Figure 8 details both decompositions applied on a complex issued from the MCL analysis. The modular decomposition herein gives no information about the structure of the complex (Fig. 8b), whereas the homogeneous decomposition indicates potential protein sub-complexes (that is, unit complexes) within the complex (Fig. 8c). Indeed, the hub contains proteins Rfa2, Rad52, Rfa1 and Rfa3, which therefore play a central role in the description of complex interactions. The interpretation of the H-node and its children indicates five cliques that should be interpreted as sub-complexes: (Rfa2, Rim101, Rfa2, Rad52, Rfa1, Rfa3); (Rtt105, Rfa2, Rfa3); (DNA2, Rfa2, Rfa1); (Rad51, Rad52); and (Mec1, Rfa2). These deductions, based on the homogeneous decomposition, should be seen as a starting point for further intra-modular investigations.

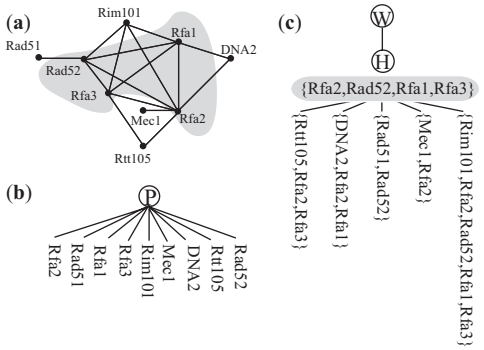


Fig. 8. Modular and homogeneous decompositions of a yeast complex presented by Hart *et al.* (2006). (**a**) Protein interaction network of proteins involved in the cell cycle and the DNA repair; (**b**) Modular decomposition tree; (**c**) Homogeneous decomposition tree. The components of the hub are highlighted in gray.

4 CONCLUSIONS

We herein studied the contribution of hierarchical decompositions of graphs to the analysis of protein interaction networks. The modular decomposition has been known for a while and showed great successes for investigating protein–protein interaction networks. This decomposition gives a representation of a graph as a tree of labeled nodes called modules. As a fundamental property, all the nodes within a module have the same neighborhood outside of the module. Based on the modular decomposition, Gagneur *et al.* (2004) interpret these modules as either: (i) ⓪-module, (ii) ①-module or (iii) P-module, the latter one designating undecomposable graphs. Unfortunately, the third case occurs on many protein interaction networks, which makes difficult the analysis of concrete biological networks, despite interesting internal structures of such subgraphs. Whereas recent techniques, like MCL analysis, overcome this problem by using clustering approaches, we here by propose to use another hierarchical decomposition that investigates the P-nodes: the homogeneous decomposition. Like the modular decomposition, the homogeneous one aims at finding the maximum number of cliques within a graph. In the protein–protein interaction network context, it gives the maximum number of complexes. As a major improvement, the homogeneous decomposition introduces two supplementary node types, namely W- and H-nodes. They give us the opportunity to identify a wheel structure around a hub, within certain P-modules. Such a structure *refines* the decomposition, thus allowing further investigation on the protein interaction network.

Compared with the modular decomposition, the homogeneous one efficiently stores the important cliques (i.e. the unit complexes). As a consequence, an easier analysis of the interactions both between complexes and within a complex is possible. The homogeneous decomposition hence represents a major contribution to infer (sub-)complexes and to identify particular features of (groups of) proteins. In particular, it emphasizes the presence/absence of specific proteins within complexes of interest, their impact or interactions on the overall network. To sum up, the homogeneous decomposition is a theoretical technique that extracts global and local information.

It should be used as a *complement* to experimental techniques that identify the complete set of interactions in a network. Further works should focus on developing new decomposition techniques, refining the current ones.

Conflict of Interest: none declared.

REFERENCES

Baumann,S. (1996) A linear algorithm for the homogeneous decomposition of graphs. *Technical Report M-9615*, Zentrum fur Mathematik, Technische Universität Munchen.

Boyle,E.I. *et al.* (2004) GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Cairns,B.R. *et al.* (1994) A multisubunit complex containing the SWI1/ADR6, SWI2/SNF2, SWI3, SNF5, and SNF6 gene products isolated from yeast. *Proc. Natl Acad. Sci. USA*, **91**, 1950–1954.

Cairns,B.R. *et al.* (1996) RSC, an essential, abundant chromatin-remodeling complex. *Cell*, **87**, 1249–1260.

Chen,M. and Shen,X. (2007) Nuclear actin and actin-related proteins in chromatin dynamics. *Curr. Opin. Cell Biol.*, **19**, 326–330.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Gagneur,J. *et al.* (2004) Modular decomposition of protein-protein interaction networks. *Genome Biol.*, **5**, R57.

Gavin,A.-C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin,A.-C. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Giaever,G. *et al.* (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, **418**, 387–391.

Guimerà,R. *et al.* (2004) Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E. Stat. Nonlin. Soft Matter phys.*, **70**, 025101.

Habib,M. and Maurer,M. (1979). On the X-join decomposition for undirected graphs. *Discrete Appl. Math.*, **1**, 201–207.

Hart,G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120.

Hart,G.T. *et al.* (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **8**, 236.

Hartwell,L. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402** (Suppl. 6761), C47–C52.

He,X. and Zhang,J. (2006) Why do hubs tend to be essential in protein networks? *PLoS Genet.*, **2**, e88.

Henry,N.L. *et al.* (1994) TFIIF-TAF-RNA polymerase II connection. *Genes Dev.*, **8**, 2868–2878.

Ho,Y. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

Jamison,B. and Olariu,S. (1995) P-components and the homogeneous decomposition of graphs. *SIAM J. Discrete Math.*, **8**, 448–463.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Kabani,M. *et al.* (2005) Anc1 interacts with the catalytic subunits of the general transcription factors TFIID and TFIIF, the chromatin remodeling complexes RSC and INO80, and the histone acetyltransferase complex NuA3. *Biochem. Biophys. Res. Commun.*, **332**, 398–403.

Kim,Y.J. *et al.* (1994) A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*, **77**, 599–608.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Ma,H.-W. *et al.* (2004) Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics*, **20**, 1870–1876.

McConnell,R.M. and Spinrad,J.P. (1994). Linear-time modular decomposition and efficient transitive orientation of comparability graphs. In *SODA '94: Proceedings of the Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA, pp. 536–545.

Möhring,R. and Radermacher,F. (1984) Substitution decomposition for discrete structures and connections with combinatorial optimization. *Ann. Discrete Math.*, **19**, 257–356.

Puig,O. *et al.* (2001) The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods*, **24**, 218–229.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.

Szallasi,Z. *et al*. (2006) In Szallasi,Z. *et al*. (eds) *System Modeling in Cellular Biology: from Concepts to Nuts and Bolts. Modules and Modularity*. The MIT Press, Cambridge, MA, pp. 41–50.

Szerlong,H. *et al.* (2003) The nuclear actin-related proteins Arp7 and Arp9: a dimeric module that cooperates with architectural proteins for chromatin remodeling. *EMBO J.*, **22**, 3175–3187.

Wilhelm,T. *et al.* (2003) Physical and functional modularity of the protein network in yeast. *Mol. Cell Proteomics*, **2**, 292–298.

Zotenko,E. *et al.* (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Comput. Biol.*, **4**, e1000140.

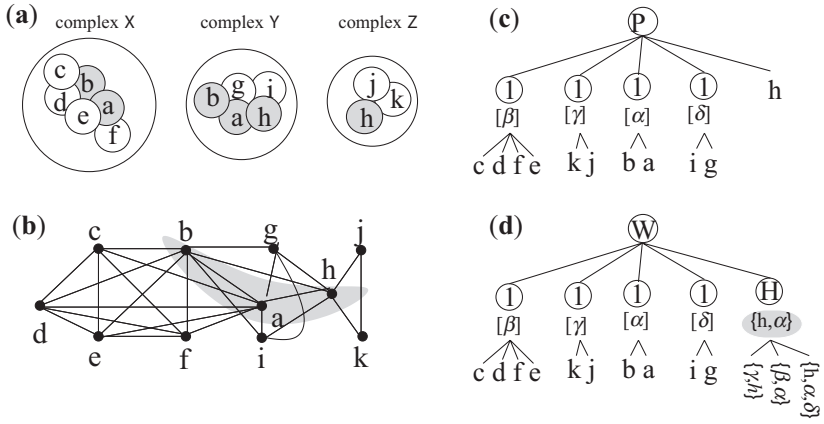


Figure 2.2: Homogeneous decomposition of a theoretical protein interaction network from Wilhelm et al. (2003). (a) Protein complexes elucidated with TAP-MS. (b) The associated interaction network, where all possible interactions within the protein complexes are considered. (c) The corresponding modular decomposition tree. (d) The homogeneous decomposition tree that highlights modules and their intra-modular interactions. The gray area indicates the hub, stored in the H-node, that connects the other h-modules as leaves.

2.3 Analysis of ecosystems from the omics lens

If new omics protocol changed the way to investigate the cellular systems and its biological abstraction, similar transitions have percolated with a ten-year delay at the population level. In particular, the early 2000s saw the rise of molecular methods to study one of the most fundamental issues in ecology: investigating the relationship between ecosystem processes and species richness in communities. Ecosystem processes concern all mechanisms that permit the chemical transformation of biological matter at the ecosystem level, whereas species richness studies the community composition. The estimation of diversity was particularly genomic-driven when applied to microbial organisms. First, because molecular techniques were mainly the only available technique, and second because these organisms were already well-documented about their respective role in biogeochemical cycles, with a particular emphasis in the nitrogen cycle [217]. Among other techniques, one can notice the significant impact of microarrays for detecting communities without the need for culturing microbial strains. Once the genomic knowledge is available (i.e., nucleic sequence of a specific gene), several studies proposed to design

probes related to nutrient cycles. These *functional gene arrays* detect the presence of genes within a given ecosystem and could be used to detect the same gene in a wide variety of microbial strains. Analysis of these probes permits to build a phylogenetic tree specific to functional genes (see Figure 2.3). In particular, the analysis of the amoA gene is of great interest because of its central role in the nitrogen cycle. amoA encodes for one of the ammonia monooxygenase subunits. This reaction catalyzes the transformation of ammonia (NH_4) into hydroxylamine (NH_2OH), which is then oxidized to nitrites (NO_2). The use of this reaction is in direct concurrency with another reaction driven by urease that transforms NH_4 into organic nitrogen compounds. Along the phylogenetic description, we proposed a complementary statistical approach to link the diversity of the functional gene amoA with these techniques - see Figure 2.4 (initially proposed in [209], then extended in [22] and [21]).

Complementary to this functional analysis, other molecular techniques were focusing on broader investigations. In a landmark study, [206] investigates the whole DNA or RNA that composes an ecosystem (first the Sargasso sea following by more samples from the global ocean via the Sorcerer II expedition [176]). Instead, on focusing on functional genes and prior knowledge, the protocol consists of focusing on the 16S ribosomal RNA fragment. This nucleic sequence is one of the most constrained over the tree of Life, one generally considers it as a good proxy for taxonomy. Thus, the gene sequence of the most abundant species will be more sequenced than the gene from rare species. Instead of assigning one sequence to a single species, one defines here the concept of Operational Taxonomic Units (OTU) that assigns multiple representative sequences to a given group called OTU. This assignation of a given sequence to an OTU was performed initially by hierarchical clustering of the nucleic sequence, followed by an arbitrary cut-off into categories. Despite the great interest of this analysis and knowledge extraction, several caveats remain. First, the choice of strict hierarchical clustering is highly sensitive, especially when one considers the speed of speciation not constant over the evolutionary time. Several other protocols were recently proposed to assess this limitation, among which the use of the swarm technique [140] that allows a dynamical clustering of sequences. The use of one assignation technique rather than another will not be discussed here, despite its significant impact on the following results. The gene copy number per organism is the second possible caveat. Indeed, some organisms, such as cyanobacteria, hold more that one 16S gene copy, which could modify the sequencing result by multiplying the gene count by a factor that corresponds to the number of gene copy. One magnifies this problem when the same sequencing technique concern metazoans where the number of cells by definition varies during their life cycle and individual growth rates.

These metagenomic results being by definition semi-quantitative, further anal-

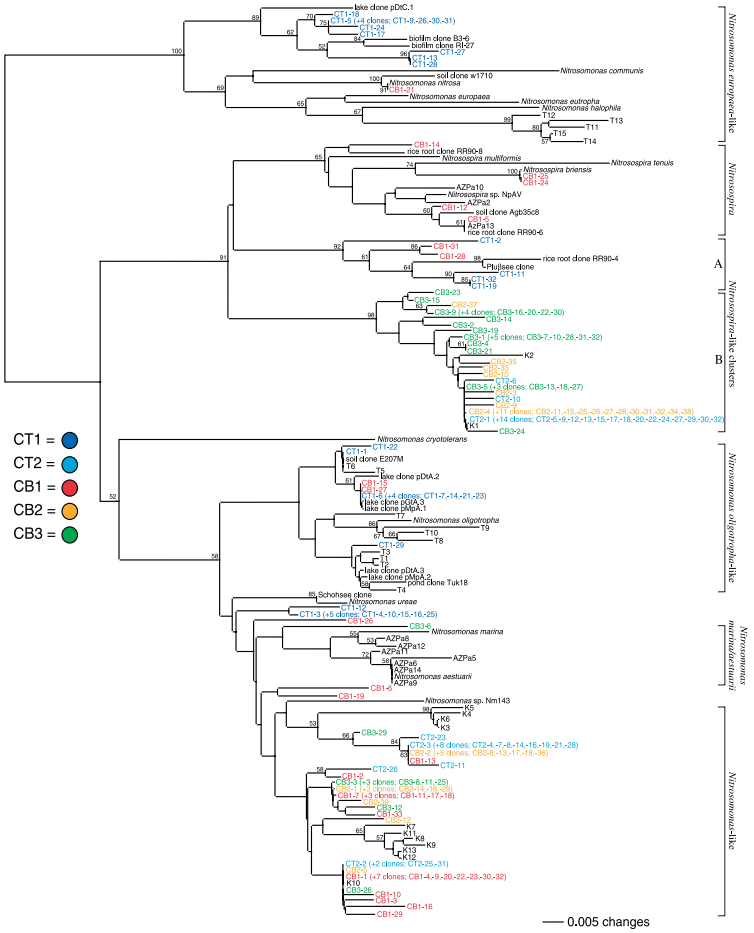


Figure 2.3: Neighbour-joining phylogenetic trees of amoA gene products and gene sequences from Chesapeake Bay sediments. Figure from Francis and collaborators [68]. The tree classifies 156 nucleic sequences of the amoA gene as extracted from the five Bay stations (see colour-coded key), together with sequences from cultivated ammonia-oxidizers and closely related environmental clones (black). Brackets highlight the different phylogenetic clusters

ysis have been proposed to avoid wrong interpretations of the gene copy number variations. Among others, *network analyses* proposes a change of computational abstraction to analyze these data. Considering that metagenomics techniques allow

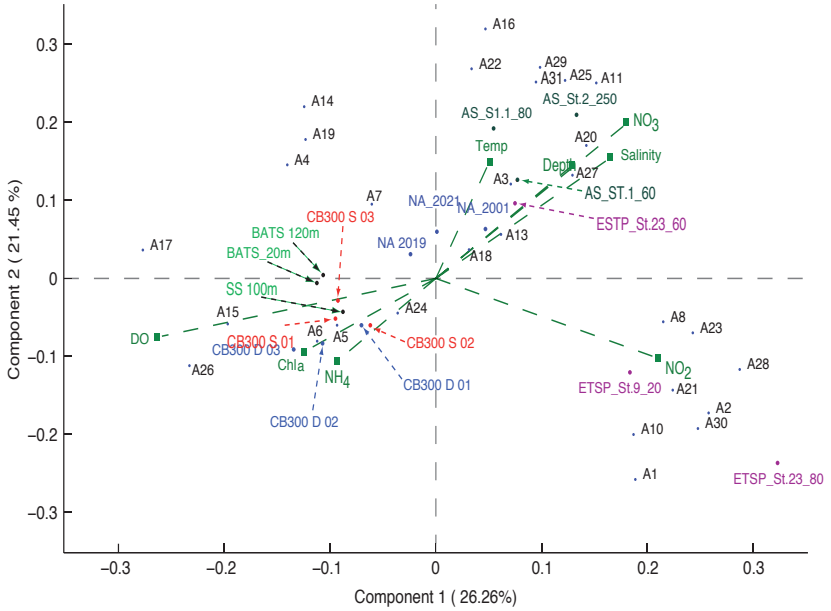


Figure 2.4: Principal components analysis based on a correlation matrix combining pre-correlated physicochemical and biological factors. Data for the plots are taken functional gene array data dedicated to amoA gene. Each probe data is represented in black. Data from the same geographic locations are grouped by color (i.e., red, green and blue). Physicochemical parameters that better explain 47.71% of the total variance are also represented by the following abbreviations: Temp, temperature; NH_4, ammonium; NO_2, nitrite; NO_3, nitrate; D.O, dissolved oxygen. Figure from Bouskill and collaborators [22]

extracting the whole DNA or RNA that composes the ecosystem, and by focusing on ribosomal RNA or specific functional genes, one counts the number of copy that belongs to a given species. An abundance matrix stores these count numbers where each line represents one OTU (*i.e.*, an approximation of species), and each column a sample site where DNA or RNA has been extracted (see Figure 2.5 for illustration). This experimental technique emphasizes *« who is there»* but also *« how many are they»* [165]. For each OTU, one then shows if one is significantly over (or under) abundant when environmental constraints challenge the ecosystem. Whereas preliminary studies focus on the description of the phylogenic distribution associated to this over or under abundance and how this diversity is related to environmental parameters (see for instance the seminal papers from [105] or [25]),

other studies propose to represent the abundance matrix as a network or graph (see [175, 82, 157] as a nonexhaustive list of studies). Roughly, when two given OTUs show abundance patterns that are correlated (and above a significant statistical threshold), one links both OTUs in a graph. This graph called a co-occurrence network is a weighted undirected graph where nodes are OTUs and edges represent significant correlations between them and weighted by a correlation score between the abundance signals associated with the two given OTUs. Herein the most critical step of this technique consists of finding an appropriate threshold above which the absolute value of the correlation is significant. Thus, as illustrated in Figure 2.5, the positive weight represents positive co-occurrence, whereas negative weights are negative co-occurrence between two OTUs. Representing the enumeration of omics data as graph significantly stimulated the recent environmental microbiology literature, and the scientific community quickly generalized the use of co-occurrence network inference protocol [23], which also opens the methodological question about the choice of the pairwise metric, *i.e.*, correlation-based measures [33, 62] versus mutual information-based measures [168]. For this purpose, the literature proposed several methods. CONET [63] proposes to combined different correlation metrics to decipher the most significant co-occurrences within the set of OTUs. SparCC [70] is dedicated to sparse datasets and proposes an adaptation of a linear Pearson correlation between log-transformed OTUs abundance signals. SparCC then considers an approximation of this correlation-like score that assumes the number of OTUs is large, and the correlation network is sparse. Both above favorite techniques, as well as state-of-the-art correlation scores, were applied on biological benchmarks for the sake of their comparison in [211]. More recently, [120] reinforces the interest on dedicated a co-occurrence inference method that handles sparse and compositional dataset with a method called SPIEC-EASI (Sparse InversE Covariance Estimation for Ecological Association Inference). This method and its recent adaptation [199] combines data transformation developed for compositional data analysis with a graphical model inference framework.

Following the global effort to investigate Earth microbiome (see for illustration [4]), several of these co-occurrence techniques have been used to summarize large metagenomic datasets: from the global ocean [133] or the human microbiome [63]. In particular, the CONET technique was also used to compare different biomes [61]. However, one must notice herein that such a comparison mostly relied on describing statistics and graph metrics of different built biome networks. Behind this methodological question lies the problem (i) of dealing with large networks and (ii) linking the network with broader biological questions. Indeed, the large number of edges makes challenging the standard functional analysis, as well as the identification of keystone species without just describing them by their centrality in the co-occurrence network using state-of-the-art graph centrality metrics.

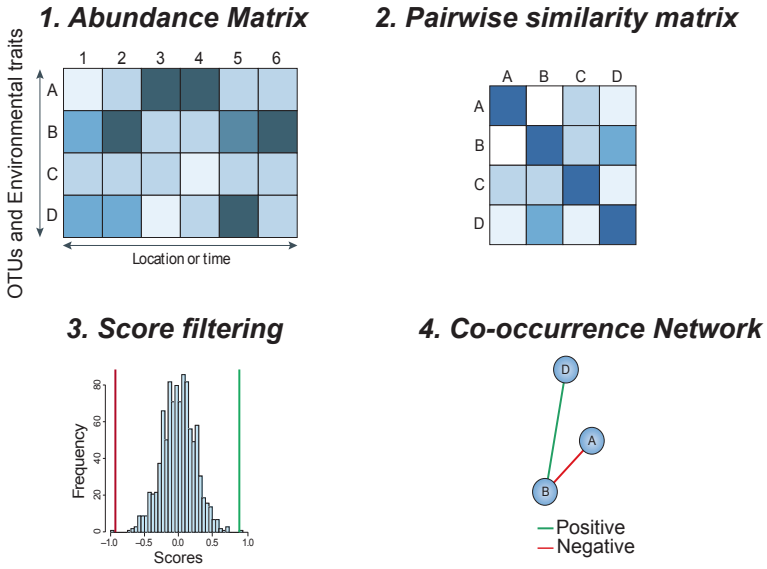


Figure 2.5: Protocol for building a co-occurrence network. From an abundance matrice that stores the relative abundances of OTUs for each sample, one can build a pairwise similarity matrix via the use of pairwise statistical scores. From the distribution of pairwise scores, compared to randomize scores, one filters the most discriminant pairs if their scores are above (or below) a given threshold with a confidence score. For each significant pair, one draws a graph where nodes are selected OTUs, and edges describe significant pairs. The pairwise score weights each edge. Figure adapted from Faust and Raes [62]

In such a broad metagenomic context, the study of the global ocean was exposed to similar questions [49], and the Tara Oceans expedition was presenting a great case study to tackle this computational problem. Launched in 2008, Tara Oceans has transformed our understanding of ocean ecology and diversity. Tara Ocean has federated a trans-disciplinary consortium around the schooner, Tara, to explore ocean diversity via two circum-global navigations [109]. During this navigation, the planktonic ecosystem was systematically sampled to cover the entire spectrum from virus, to prokaryotic, to eukaryotic organisms via meta-barcoding (*i.e.*, sequencing functional genes) [45, 189, 26], meta-genomic (*i.e.,* sequencing of the all genes in a given water mass) [189, 30] or meta-transcriptomic (*i.e.*, quantitative expression of all genes in a given water mass). For the past ten years, this project was driven by performing a holistic description of the plankton, major

biological component of the global ocean. This expedition enables many scientific advances, among others, (i) the most extensive collection of homogeneous eco-morpho-genetic samples and data from a biome, which today makes plankton the best described planetary ecosystem in terms of organisms and genomes, from viruses to animals ; (ii) The first referenced pan-ecosystemic database established at EBI for open and sustainable data sharing with the international community.

Beyond the description of planktonic data and the corresponding unveiled complexity, the Tara consortium performed other integrative studies. Villar and collaborators [207] linked the plankton diversity and putative planktonic functions with ocean circulation around the South Africa edge. Lima Mendez and collaborators [133] integrate the plankton relative abundance table as a co-occurrence network via the use of CONET. However, these techniques quickly underperformed to investigate a more general oceanographic question. In particular, the question of carbon export is of great interest in the context of climate change. The carbon export consists of quantifying the amount of carbon captured by the ocean. The physicochemical properties of the water can perform such capture. An increase of carbon dioxide in the atmosphere automatically increases the amount of carbon encapsulated within the water mass, which also decreases the pH of the water mass (*i.e.*, ocean acidification) automatically. This carbon capture belongs to a process called the physical pump that is responsible for most of the carbon flux (90%) from the atmosphere to the ocean. Not at the same order of magnitude (10%), complementary carbon sequestration involves the biological component of the global ocean. This sequestration occurs via trophic networks. Trophic networks are the result of all biotic interactions and magnified by individual biological behaviors (*i.e.*, predation, mutualism, commensalism). Roughly, these networks take place following the availability of solar energy; photon; that are used by photosynthetic systems to build organic matter that will be further assimilated by heterotrophic organisms. Thus, at the scale of the Earth system, the unique source of energy is therefore dissipated as (i) heat that contributes to the global ocean circulation or (ii) biological matter distributed within several trophic networks or biogeochemical cycles. Following fluxes of matter within trophic networks, the organic carbon will finally sink to reach ocean floors by taking the form of either bioproduct excretion (*i.e.*, mostly mucus that produces marine snow) or dead bodies potentially desegregated along with their sink toward the bottom floor. This process that leads to proper sequestration of carbon within sediments is called the carbon export, but its complexity remains challenging to investigate. Sediments traps usually measure the carbon export; a protocol that estimates the quantity of carbon per times; or more recently a computational proxy designed by colleagues from the "Laboratoire d'Océanographie de Villefranche" - LOV called UVP (Underwater Video Profiler) that estimates the carbon-based on particles distributions along with depth

profiles - distribution parametrized via video recording [159].

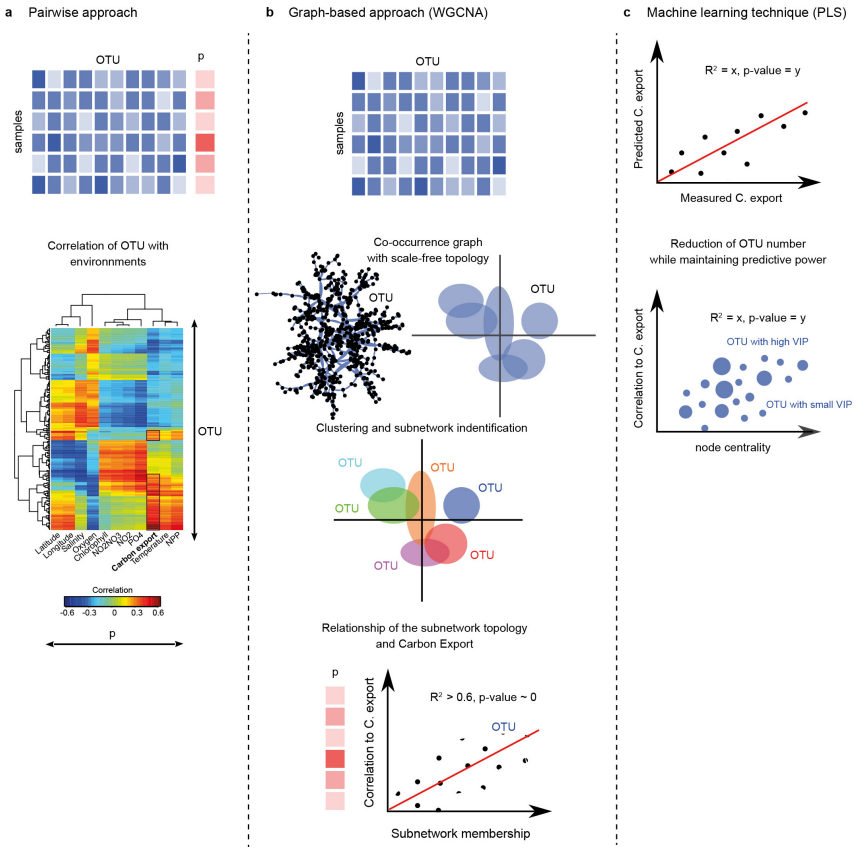


Figure 2.6: Overview of analytical methods used to decipher planktonic communities associated to carbon export. A. represents the standard pairwise analysis applied on a relative abundance matrix for s samples (s × OTUs (operational taxonomic units)) and its corresponding environmental matrix (s × p (parameters)). B. depicts the general protocol of WGCNA [122] when applied on OTUs co-occurrence network. C. illustrates the method used to reduce the number of OTUs of interest. The figure is adapted from [90].

However, even if estimating the carbon export is becoming a state-of-the-art measurement during oceanographical campaigns (but at high logistic costs), investigating its origin remains challenging because it implies to decipher trophic networks and eventually the whole planktonic system involved in the biological pump.

The Tara Oceans dataset could overcome such a limitation, and with colleagues from the Tara Oceans consortium, we design a study that focuses on delineating relationships between planktonic features and the carbon export. The standard analysis consists in searching for significant correlations between species or OTUs abundances and carbon export (see Figure2.6A). Several statistical methods exist to emphasize the most significant associations, among which sPLS as available in the Mixomics package [172]. However, behind their significant impact, these approaches highlight OTUs that explain the most significant variance, which does not necessarily describe communities that are involved. Complementary, we proposed the use of a graph-based approach that focuses on the whole co-occurrence network for the sake of community description. Among these putative communities, we will aim at deciphering the OTUs that drive the community. The analysis of the co-occurrence network extracted from the Tara Oceans dataset is complicated, and a previous study mainly focused on the graph description [133]. A large number of edges makes challenging the standard functional analysis, as well as the identification of patterns within the graph. To overcome this problem, we propose to apply a network analysis called WGCNA (Weighted Gene Correlation Network Analysis) that clusters the graph based on its overall topology. This technique, inspired from [122] shown its interest in the context of systems biology [94], and more recently display excellent efficiency compared to other module detection methods [177]. Using only a relative abundance matrix (s $\times$ OTUs) (see Figure Figure2.6B), WGCNA builds a graph where nodes are OTUs and edges represent significant co-occurrence. However, compared to the co-occurrence mentioned above techniques, WGCNA builds a weighted graph and focuses on such an abstraction to perform the analysis. Thus, co-occurrence scores between nodes are weights allocated to corresponding edges. WGCNA aims at detecting modules within the graph to emphasize a stronger group of OTUs that present a strong correlation between them. In this purpose, weights from the weighted graph are magnified by a power-law function until the graph becomes scale-free. This assumption relies on previous observations on biological networks [7]. The graph is then decomposed into subnetworks (groups of OTUs) that are analyzed separately. One subnetwork (a group of OTUs) is considered of interest when its topology is related to the trait of interest; in the current case, carbon export. For each subnetwork (for instance, the subnetwork related to carbon export), each OTU is spread within a feature space that plots each OTU based on its membership to the subnetwork (x-axis) and its correlation to the environmental trait of interest (that is, carbon export). The membership is estimated by the module eigenvalue [122]. A suitable regression of all OTUs emphasizes the putative relationship of the subnetwork topology and the carbon export trait. These modules are then considered as *trait-like* because the more a given OTU is crucial to define the subnetwork topology (*i.e.*, robust eigen-

value), the more it is correlated to the trait; the carbon export herein. Finally, to reduce the number of OTU to investigate (see Figure 2.6C), we applied a Partial Least Squares Regression on each module associated with the carbon export. This technique computes a score for each OTU that belong to the module. The score then refers to variable importance in projection and reflects the relative predictive power of a given OTU. Higher scores (that is, larger circles) emphasize the most essential OTUs for the sake of prediction. OTUs with a VIP score higher than one, are necessary for the predictive model. Considering a weighted graph to abstract the planktonic community shows its great impact in the following study where we investigate the plankton associated with the carbon export:

Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline, Jennifer R Brum, Jennifer Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Tara Oceans Consortium Coordinators, Lars Stemmann, Fabrice Not, Pascal Hingamp, Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stéphane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Peer Bork, Colomban de Vargas, Daniele Iudicone, Matthew B Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470, February 2016

This study discussed the interest of emphasized planktonic species. In particular, the same method was applied on distinct abundance matrices: (i) prokaryotic species, (ii) eukaryotic species, (iii) viruses proteins that reflect viral diversity, but also (iv) prokaryotic genes. Each matrix, abstracted as a weighted graph, produces one module associated to carbon export. Worth to notice, and for the sake of validation, we found consistencies between main species from different matrices, especially about the role of *Synechococcus sp.* and radiolarian, cryptic species difficult observe before the systematic use of environmental genomics.

ARTICLE

doi:10.1038/nature16942

Plankton networks driving carbon export in the oligotrophic ocean

Lionel Guidi[1,2]*, Samuel Chaffron[3,4,5]*, Lucie Bittner[6,7,8]*, Damien Eveillard[9]*, Abdelhalim Larhlimi[9], Simon Roux[10]†, Youssef Darzi[3,4], Stephane Audic[8], Léo Berline[1]†, Jennifer R. Brum[10]†, Luis Pedro Coelho[11], Julio Cesar Ignacio Espinoza[10], Shruti Malviya[7]†, Shinichi Sunagawa[11], Céline Dimier[8], Stefanie Kandels-Lewis[11,12], Marc Picheral[1], Julie Poulain[13], Sarah Searson[1,2], *Tara* Oceans Consortium Coordinators‡, Lars Stemmann[1], Fabrice Not[8], Pascal Hingamp[14], Sabrina Speich[15], Mick Follows[16], Lee Karp-Boss[17], Emmanuel Boss[17], Hiroyuki Ogata[18], Stephane Pesant[19,20], Jean Weissenbach[13,21,22], Patrick Wincker[13,21,22], Silvia G. Acinas[23], Peer Bork[11,24], Colomban de Vargas[8], Daniele Iudicone[25], Matthew B. Sullivan[10]†, Jeroen Raes[3,4,5], Eric Karsenti[7,12], Chris Bowler[7] & Gabriel Gorsky[1]

The biological carbon pump is the process by which CO_2 is transformed to organic carbon via photosynthesis, exported through sinking particles, and finally sequestered in the deep ocean. While the intensity of the pump correlates with plankton community composition, the underlying ecosystem structure driving the process remains largely uncharacterized. Here we use environmental and metagenomic data gathered during the *Tara* Oceans expedition to improve our understanding of carbon export in the oligotrophic ocean. We show that specific plankton communities, from the surface and deep chlorophyll maximum, correlate with carbon export at 150 m and highlight unexpected taxa such as Radiolaria and alveolate parasites, as well as *Synechococcus* and their phages, as lineages most strongly associated with carbon export in the subtropical, nutrient-depleted, oligotrophic ocean. Additionally, we show that the relative abundance of a few bacterial and viral genes can predict a significant fraction of the variability in carbon export in these regions.

Marine planktonic photosynthetic organisms are responsible for approximately 50% of Earth's primary production and fuel the global ocean biological carbon pump[1]. The intensity of the pump is correlated with plankton community composition[2,3], and controlled by the relative rates of primary production and carbon remineralization[4]. About 10% of this newly produced organic carbon in the surface ocean is exported through gravitational sinking of particles. Finally, after multiple transformations, a fraction of the exported material reaches the deep ocean where it is sequestered over thousand-year timescales[5].

Like most biological systems, marine ecosystems in the sunlit upper layer of the ocean (denoted as the euphotic zone) are complex[6,7], characterized by a wide range of biotic and abiotic interactions[8–10] and in constant balance between carbon production, transfer to higher trophic levels, remineralization, and export to the deep layers[11]. The marine ecosystem structure and its taxonomic and functional composition probably evolved to comply with this loss of energy by modifying organism turnover times and by the establishment of complex feedbacks between them[6] and the substrates they can exploit for metabolism[12]. Decades of ground-breaking research have focused on identifying independently the key players involved in the biological carbon pump. Among autotrophs, diatoms are commonly attributed to being important in carbon flux because of their large size and fast sinking rates[13–15], while small autotrophic picoplankton may contribute directly through subduction of surface water[16] or indirectly by aggregating with larger settling particles or consumption by organisms at higher trophic levels[17]. Among heterotrophs, zooplankton such as crustaceans impact carbon flux via production of fast-sinking fecal pellets while migrating hundreds of meters in the water column[18,19]. These observations, focusing on just a few components of the marine ecosystem, highlight that carbon export results from multiple biotic interactions and that a better understanding of the mechanisms involved in its regulation requires an analysis of the entire planktonic ecosystem.

Advanced sequencing technologies offer the opportunity to simultaneously survey whole planktonic communities and associated

[1]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. [2]Department of Oceanography, University of Hawaii, Honolulu, Hawaii 96822, USA. [3]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [4]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [5]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [6]Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005, Paris, France. [7]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France. [8]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, 29680 Roscoff, France. [9]LINA UMR 6241, Université de Nantes, EMN, CNRS, 44322 Nantes, France. [10]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA. [11]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. [12]Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1, 69117 Heidelberg, Germany. [13]CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [14]Aix Marseille Université, CNRS, IGS, UMR 7256, 13288 Marseille, France. [15]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris CEDEX 05, France. [16]Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [17]School of Marine Sciences, University of Maine, Orono, Maine 04469, USA. [18]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [19]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [20]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [21]CNRS, UMR 8030, CP 5706 Evry, France. [22]Université d'Evry, UMR 8030, CP 5706 Evry, France. [23]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona E0800, Spain. [24]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [25]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.
†Present addresses: Department of Microbiology, The Ohio State University, Columbus, Ohio 43210, USA (S.R., J.R.B.); Department of Microbiology, and Department of Civil, Environmental and Geodetic Engineering, The Ohio Sate University, Colombus, Ohio 43210, USA (M.B.S.); Aix Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography (MIO) UM 110, 13288, Marseille, France (L.B.); Biological Oceanography Division, CSIR-National Institute of Oceanography, Dona Paula, Goa 403 004, India (S.M.).
*These authors contributed equally to this work.
‡A list of authors and affiliations appears at the end of the paper.

molecular functions in unprecedented detail. Such a holistic approach may allow the identification of community- or gene-based biomarkers that could be used to monitor and predict ecosystem functions, for example, related to the biogeochemistry of the ocean[20–22]. Here, we leverage global-scale ocean genomics data sets from the euphotic zone[10,23–25] and associated environmental data to assess the coupling between ecosystem structure, functional repertoire, and carbon export at 150 m.

Carbon export and plankton community composition

The *Tara* Oceans global circumnavigation crossed diverse ocean ecosystems and sampled plankton at an unprecedented scale[20,26] (see Methods). Hydrographic data were measured *in situ* or in seawater samples at all stations, as well as nutrients, oxygen and photosynthetic pigments (see Methods). Net primary production (NPP) was derived from satellite measurements (see Methods). In addition, particle size distributions (100 μm to a few millimetres) and concentrations were measured using an underwater vision profiler (UVP) from which carbon export, corresponding to the carbon flux (Fig. 1a) at 150 m, was calculated to range from 0.014 to 18.3 mg m^{-2} d^{-1} using methods previously described (see Methods). One should keep in mind that fluxes are calculated from images of particles. These estimates are derived from an approximation of Stokes' law relating the equivalent spherical diameter of particles to carbon flux (see Methods). This exponential approximation is reasonable assuming similar particle composition across all sizes, as highlighted by the standard deviations of parameters in equation (5) (see Methods). Furthermore, because of instrument and method limitations, particles <250 μm were not used, which may underestimate total carbon fluxes. Finally, these fluxes are instantaneous because they do not integrate space and time as sediment traps would. However, the approach allowed us to assemble the largest homogeneous carbon export data set during a single expedition, corresponding to more than 600 profiles over 150 stations. This data set is of similar magnitude to the body of historical data available in the literature that includes the 134 deep sediment trap-based carbon flux time series[27] from the JGOFS program and the 419 thorium-derived particulate organic carbon (POC) export measurements[28].

From 68 globally distributed sites, a total of 7.2 terabases (Tb) of metagenomics data, representing ~40 million non-redundant genes, around 35,000 operational taxonomic units (OTUs) of prokaryotes (Bacteria and Archaea) and numerous mainly uncharacterized viruses and picoeukaryotes, have been described recently[23,25]. In addition, a set of 2.3 million eukaryotic 18S rDNA ribotypes was generated from a subset of 47 sampling sites corresponding to approximately 130,000 OTUs[24]. Finally, 5,476 viral 'populations' were identified at 43 sites from viral metagenomic contigs, only 39 (<0.1%) of which had been previously observed[25] (see Methods). These genomics data combined across all domains of life and viruses together with carbon export estimates (Fig. 1a) and other environmental parameters were used to explore the relationships between marine biogeochemistry and euphotic plankton communities (see Methods) in the top 150 m of the oligotrophic open ocean. Our study did not include high-latitude areas owing to the current lack of available molecular data and results should not be extrapolated to deeper depths.

Using a method for regression-based modelling of highly multidimensional data in biology (specifically a sparse partial least square analysis (sPLS)[29], Extended Data Fig. 1), we detected several plankton lineages for which relative sequence abundance correlated with carbon export and other environmental parameters, most notably with NPP, as expected (Fig. 1b and see Supplementary Table 1). These included diatoms, dinoflagellates and Metazoa (zooplankton), lineages classically identified as key contributors to carbon export.

Plankton networks associated with carbon export

While the analysis presented in Fig. 1b supports previous findings about key organisms involved in carbon export from the euphotic zone[14,15,17–19], it is not able to capture how the intrinsic structure of the planktonic community relates to this biogeochemical process. Conversely, although other recent holistic approaches[10,30,31] used species co-occurrence networks to reveal potential biotic interactions, they do not provide a robust description of sub-communities driven by abiotic interactions. To overcome these issues, we applied a systems biology approach known as weighted gene correlation network analysis (WGCNA)[32,33] to detect significant associations between the

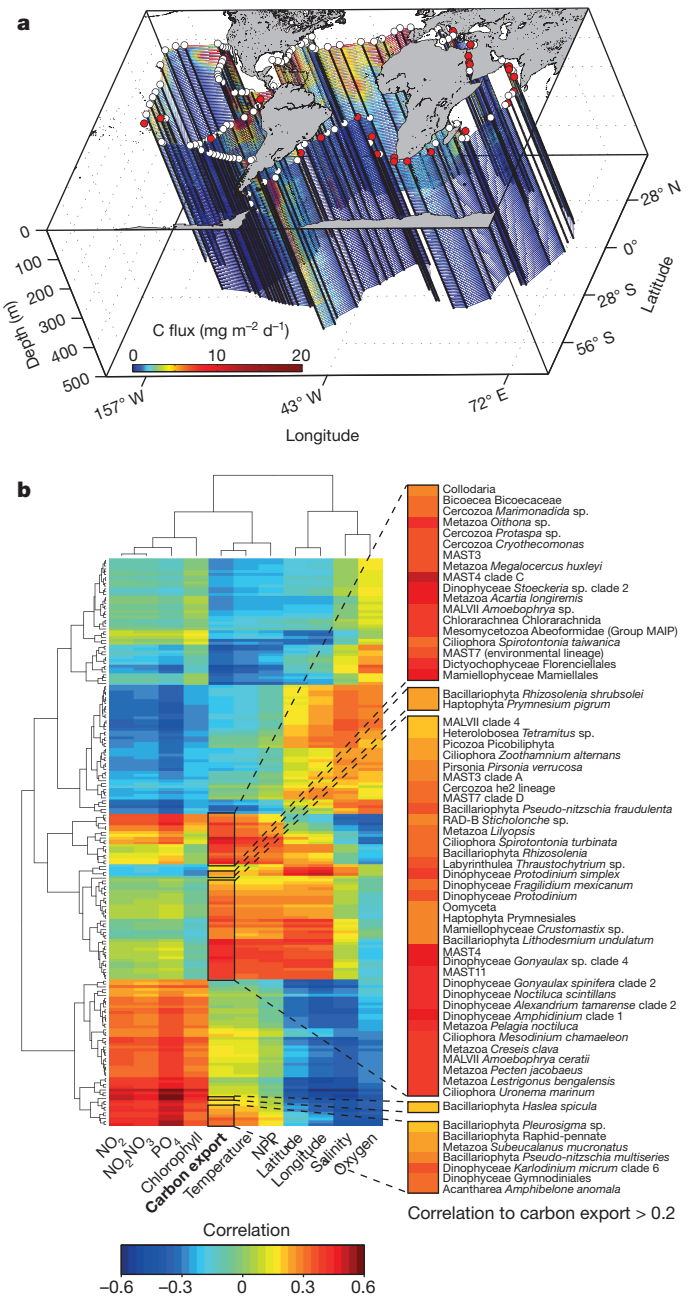


Figure 1 | Global view of carbon fluxes along the *Tara* Oceans circumnavigation route and associated eukaryotic lineages. a, Carbon flux in mg m^{-2} d^{-1} and carbon export at 150 m estimated from particle size distribution and abundance measured with the underwater vision profiler (UVP). Stations at which environmental data are available (Supplementary Table 9) are depicted by white dots. Stations at which eukaryotic samples are available are coloured in red (Supplementary Tables 10 and 12). **b**, Eukaryotic lineages associated to carbon export as revealed by standard methods for regression-based modelling (sPLS analysis). Correlations between lineages and environmental parameters are depicted as a clustered heat map and lineages with a correlation to carbon export higher than 0.2 are highlighted (detailed results in Supplementary Table 1).

Tara Oceans genomics data and carbon export. This method delineates communities in the euphotic zone that are the most associated with carbon export rather than predicting organisms associated with sinking particles.

In brief, the WGCNA approach builds a network in which nodes are features (in this case plankton lineages or gene functions) and links are evaluated by the robustness of co-occurrence scores. WGCNA then clusters the network into modules (hereafter denoted subnetworks) that can be examined to find significant subnetwork–trait relationships. We then filtered each subnetwork using a partial least square (PLS) analysis that emphasizes key nodes (based on the variable importance in projection (VIP) scores; see Methods and Extended Data Fig. 1). These particular nodes are mandatory to summarize a subnetwork (or community) related to carbon export. In particular, they are of interest for evaluating: (i) subnetwork robustness; and (ii) predictive power for a given trait (see Methods and Extended Data Fig. 1).

We applied WGCNA to the relative abundance tables of eukaryotic, prokaryotic and viral lineages[23–25] and identified unique subnetworks significantly associated with carbon export within each data set (see Methods and Supplementary Tables 2–4). The eukaryotic subnetwork (subnetwork–trait relationship to carbon export, Pearson correlation $r = 0.81$, $P = 5 \times 10^{-15}$) contained 49 lineages (Extended Data Fig. 2a and Supplementary Table 2) among which 20% represented photosynthetic organisms (Fig. 2a and Supplementary Table 2). Surprisingly, this small subnetwork's structure correlates very strongly to carbon export ($r = 0.87$, $P = 5 \times 10^{-16}$, Extended Data Fig. 2d) and it predicts as much as 69% (leave-one-out cross-validated (LOOCV), $R^2 = 0.69$) of the variability in carbon export (Extended Data Fig. 2g). Only ~6% of the subnetwork nodes correspond to diatoms and they show lower VIP scores than dinoflagellates (Supplementary Table 2). This is probably because our samples are not from silicate-replete conditions where diatoms were blooming. Furthermore, our analysis did not incorporate data from high latitudes, where diatoms are known to be particularly important for carbon export, so this result suggests that dinoflagellates have a heretofore unrecognized role in carbon export processes in subtropical oligotrophic 'type' ecosystems. More precisely, four of the five highest VIP scoring eukaryotic lineages that correlated with carbon export at 150 m were heterotrophs such as Metazoa (copepods), non-photosynthetic Dinophyceae, and Rhizaria (Fig. 2a and Supplementary Table 2). These results corroborate recent metagenomics analysis of microbial communities from sediment traps in the oligotrophic North Pacific subtropical gyre[34]. Consistently, *in situ* imaging surveys have revealed Rhizarian lineages, made up of large fragile organisms such as the Collodaria, to represent an until now under-appreciated component of global plankton biomass (T. Biard *et al.*, submitted), which here also appear to be of relevance for carbon export. Another 14% of lineages from the subnetwork correspond to parasitic organisms, a largely unexplored component of planktonic ecosystems when studying carbon export.

The prokaryotic subnetwork that associated most significantly with carbon export at 150 m (subnetwork–trait relationship to carbon export, $r = 0.32$, $P = 9 \times 10^{-3}$) contained 109 OTUs (Extended Data Fig. 2b and Supplementary Table 3), its structure correlated well to carbon export ($r = 0.47$, $P = 5 \times 10^{-6}$, Extended Data Fig. 2e) and it could predict as much as 60% of the carbon export variability (LOOCV, $R^2 = 0.60$) (Extended Data Fig. 2h). By far the highest VIP score within this community was assigned to *Synechococcus*, followed by *Cobetia*, *Pseudoalteromonas* and *Idiomarina*, as well as *Vibrio* and *Arcobacter* (Fig. 2b and Supplementary Table 3). Noteworthy, the genus *Prochlorococcus* and SAR11 clade fall out of this community, while the significance of *Synechococcus* for carbon export could be validated using absolute cell counts estimated by flow cytometry ($r = 0.64$,

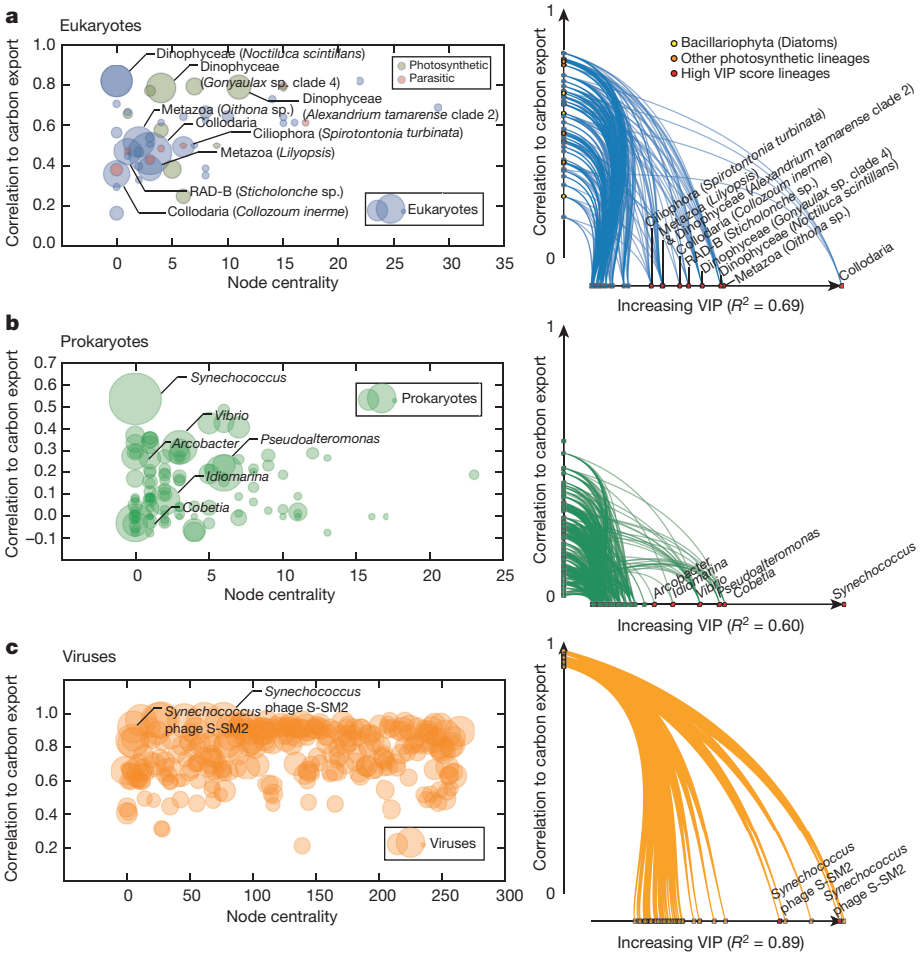


Figure 2 | Ecological networks reveal key lineages associated with carbon export at 150 m at global scale. The relative abundances of taxa in selected subnetworks were used to estimate carbon export and to identify key lineages associated with the process. **a**, The selected eukaryotic subnetwork ($n = 49$, see Supplementary Table 2) can predict carbon export with high accuracy (PLS regression, LOOCV, $R^2 = 0.69$, see Extended Data Fig. 2g). Lineages with the highest VIP score (dot size is proportional to the VIP score in the scatter plot) in the PLS are depicted as red dots corresponding to three Rhizaria (Collodaria, *Collozoum inerme* and *Sticholonche* sp.), one copepod (*Oithona* sp.), one siphonophore (*Lilyopsis*), three Dinophyceae and one ciliate (*Spirotontonia turbinata*). **b**, The selected prokaryotic subnetwork ($n = 109$, see Supplementary Table 3) can predict carbon export with good accuracy (PLS regression, LOOCV, $R^2 = 0.60$, see Extended Data Fig. 2h). **c**, The selected viral population subnetwork ($n = 277$, see Supplementary Table 4) can predict carbon export with high accuracy (PLS regression, LOOCV, $R^2 = 0.89$, see Extended Data Fig. 2i). Two viral populations with a high VIP score (red dots) are predicted as *Synechococcus* phages (see Supplementary Table 4).

$P = 4 \times 10^{-10}$, Extended Data Fig. 2k). Moreover, *Prochlorococcus* cell counts did not correlate with carbon export ($r = -0.13$, $P = 0.27$, Extended Data Fig. 2j) whereas the *Synechococcus* to *Prochlorococcus* cell count ratio correlated positively and significantly ($r = 0.54$, $P = 4 \times 10^{-7}$, Extended Data Fig. 2l), suggesting the relevance of *Synechococcus*, rather than *Prochlorococcus*, to carbon export. Notably, *Pseudoalteromonas*, *Idiomarina*, *Vibrio* and *Arcobacter* (of which several species are known to be associated with eukaryotes[35]) have also been observed in live and poisoned sediment traps[34] and display very high VIP scores in the subnetwork associated with carbon export. Additional genera reported as being enriched in poisoned traps (also known as being associated with eukaryotes) include *Enterovibrio* and *Campylobacter*, and are present as well in the carbon export associated subnetwork.

Interestingly, the viral subnetwork (involving 277 populations) most related to carbon export at 150 m ($r = 0.93$, $P = 2 \times 10^{-15}$, Extended Data Fig. 2c) contained particularly high VIP scores for two *Synechococcus* phages (Fig. 2c and Supplementary Table 4), which represented a 16-fold enrichment (Fisher's exact test $P = 6.4 \times 10^{-9}$). Its structure also correlated with carbon export ($r = 0.88$, $P = 6 \times 10^{-93}$, Extended Data Fig. 2f) and could predict up to 89% of the variability of carbon export (LOOCV, $R^2 = 0.89$) (Extended Data Fig. 2i). The significance of these convergent results is reinforced by the fact that sequences from these data sets are derived from organisms collected on distinct filters with different mesh sizes (see Methods), and further implicates the importance of top-down processes in carbon export.

With the aim of integrating eukaryotic, prokaryotic, and viral communities in the euphotic zone with carbon export at 150 m, we synthesized their respective subnetworks using a single global co-occurrence network established previously[10]. The resulting network focused on key lineages and their predicted co-occurrences (Fig. 3). Lineages with high VIP values (such as *Synechococcus*) are revealed as hubs of the co-occurrence network[10], illustrating the potentially strategic key roles within the integrated network of lineages under-appreciated by conventional methods to study carbon export. Associations between the hub lineages are mostly mutually exclusive, which may explain the relatively weak correlation of some of these lineages with carbon export when using standard correlation analyses, as shown in Fig. 1b.

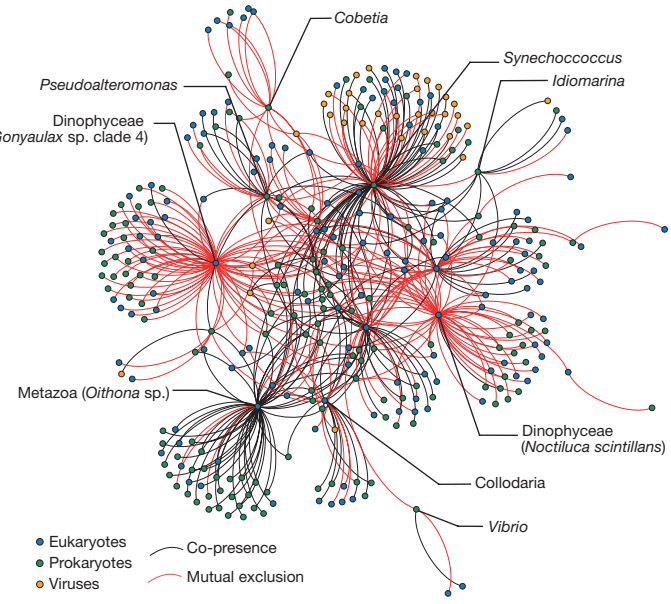


Figure 3 | Integrated plankton community network built from eukaryotic, prokaryotic and viral subnetworks related to carbon export at 150 m. Major lineages were selected within the three subnetworks (VIP > 1) (Supplementary Tables 2, 3 and 4). Co-occurrences between all lineages of interest were extracted, if present, from a previously established global co-occurrence network (see Methods). Only lineages discussed within the study are pinpointed. The resulting graph is composed of 329 nodes, 467 edges, with a diameter of 7, and average weighted degree of 4.6.

Gene functions associated with carbon export

Given the potential importance of prokaryotic processes influencing the biological carbon pump[22], we used the same analytical approaches to examine the prokaryotic genomic functions associated with carbon export at 150 m in the annotated Ocean Microbial Reference Gene Catalogue from *Tara* Oceans[23]. We built a global co-occurrence network for functions (that is, orthologous groups of genes (OGs)) from the euphotic zone and identified two subnetworks of functions that are significantly associated with carbon export (light and dark green subnetworks; FNET1 and FNET2, respectively, see Extended Data Fig. 3a–c).

The majority of functions in FNET1 and FNET2 correlate well with carbon export (FNET1: mean $r = 0.45$, s.d. = 0.09 and FNET2: mean $r = 0.34$, s.d. = 0.10). Interestingly, FNET2 functions ($n = 220$) encode mostly (83%) core functions (that is, functions observed in all euphotic samples, see Methods) while the majority of FNET1 functions ($n = 441$) are non-core (85%) (see Supplementary Tables 5 and 6), highlighting both essential and adaptive ecological functions associated with carbon export. Top VIP scoring functions in the FNET1 subnetwork are membrane proteins such as ABC-type sugar transporters (Extended Data Fig. 3c). This subnetwork also contains many functions specific to the *Synechococcus* accessory photosynthetic apparatus (for example, relating to phycobilisomes, phycocyanin and phycoerythrin; see Supplementary Table 5), which is consistent with the major role of this genus for carbon export inferred from the prokaryotic subnetwork (Fig. 2b). In addition, functions related to carbohydrates, inorganic ion transport and metabolism, as well as transcription, are also well represented (Fig. 4), suggesting overall a subnetwork of functions dedicated to photosynthesis and growth.

The FNET2 subnetwork contains several functions encoded by genes taxonomically assigned to *Candidatus pelagibacter* and *Prochlorococcus*, known as occupying similar oceanic regions as *Synechococcus*, but overall most of its relative abundance (74%) is taxonomically unclassified (Extended Data Fig. 3e). Top VIP scoring functions in FNET2 are also membrane proteins and ABC-type sugar transporters, as well as functions involved in carbohydrate breakdown such as a chitinase (Extended Data Fig. 3c). These features highlight the potential roles of bacteria in the formation and degradation of marine aggregates[36]. Notably, 77% and 58%, of OGs with a VIP score >1 in FNET1 and FNET2, respectively, are functionally uncharacterized[37,38] (Fig. 4), pointing to the strong need for future molecular work to explore these functions (see Supplementary Tables 5 and 6).

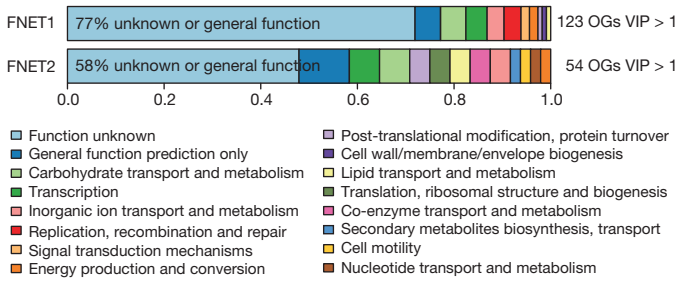


Figure 4 | Key bacterial functional categories associated with carbon export at 150 m at global scale. A bacterial functional network was built based on orthologous group/gene (OG) relative abundances using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters. Two functional subnetworks (FNET1 ($n = 220$) and FNET2 ($n = 441$), respectively, Extended Data Fig. 3a) are significantly associated with carbon export (FNET1: $r = 0.42$, $P = 4 \times 10^{-9}$ and FNET2: $r = 0.54$, $P = 7 \times 10^{-6}$, see Extended Data Fig. 3b). Higher functional categories are depicted for functions with a VIP score >1 (PLS regression, LOOCV, FNET1 $R^2 = 0.41$ and FNET2 $R^2 = 0.48$, see Extended Data Fig. 3d) in both subnetworks.

As for plankton communities, the relevance of the identified bacterial functions to predict carbon export was also confirmed by PLS regression (Extended Data Fig. 3d). The functional subnetworks predict 41% and 48% of carbon export variability (LOOCV, $R^2 = 0.41$ and 0.48 for FNET1 and FNET2, respectively) with a minimal number of functions (Fig. 4, 123 and 54 functions with a VIP score >1 for FNET1 and FNET2, respectively). Finally, higher predictive power was obtained using subnetworks of viral protein clusters (Extended Data Fig. 4a–c), predicting 55% and 89% of carbon export variability (LOOCV $R^2 = 0.55$ and 0.89 for VNET1 and VNET2, respectively; Extended Data Fig. 4d, Supplementary Tables 7 and 8), suggesting a key role of not only bacteria, but also their phages in processes sustaining carbon export at a global level.

Discussion

In this work we reveal the potential contribution of unexpected components of plankton communities, and confirm the importance of prokaryotes and viruses for carbon export in the nutrient-depleted oligotrophic ocean. Carbon export at 150 m has been estimated from particle size distribution in a global data set, but should be taken with caution, as the estimates do not account for particle composition. In addition, these export estimates evaluate how much carbon leaves the euphotic zone, but they are not related and should not be extrapolated to sequestration, which occurs after remineralization, deeper in the water column, and over longer timescales. Nonetheless, the use of the UVP was the only realistic method to evaluate carbon flux over the 3-year expedition because deployment of sediment traps at all stations would have been impossible. While our findings are consistent with the numerous previous studies that have highlighted the central role of copepods and diatoms in carbon export[14,15,17–19], they place them in an ecosystem context and reveal hypothetical processes correlating with the intensity of export, such as parasitism, infection and predation. For example, while viruses are commonly assumed to lyse cells and maintain fixed organic carbon in surface waters, thereby reducing the intensity of the biological carbon pump[39], there are hints that viral lysis may increase carbon export through the production of colloidal particles and aggregate formation[40]. Our current study suggests that these latter roles may be more ubiquitous than currently appreciated. The importance of aggregation and cell stickiness as inferred from gene network analysis should be further explored mechanistically to investigate the biological significance of these findings.

The future evolution of the oceanic carbon sink remains uncertain because of poorly constrained processes, particularly those associated with the biological pump. With current trends in climate change, the size and biodiversity of phytoplankton are predicted to decrease globally[41,42]. Furthermore, in spite of the potential importance of viruses revealed in this study, they have largely been ignored because of limitations in sampling technologies. Consequently, as oligotrophic gyres expand and global mean NPP decreases[43], the field is currently unable to predict the consequences for carbon export from the ocean's euphotic zone. By pinpointing key lineages and key microbial functions that correlate with carbon export at 150 m in these areas, this study provides a framework to address this critical bottleneck. However, the associations presented do not necessarily suggest a causal effect on carbon export, which will require further investigation.

One of the grand challenges in the life sciences is to link genes to ecosystems[44], based on the posit that genes can have predictable ecological footprints at community and ecosystem levels[45–47]. The *Tara* Oceans data sets have allowed us to predict as much as 89% of the variability in carbon export from the oligotrophic surface ocean with just a small number of genes, largely with unknown functions, encoded by prokaryotes and viruses. These findings can be used as a basis to include biological complexity and guide experimental work designed to inform climate modelling of the global carbon cycle. Such statistical analyses, scaling from genes to ecosystems, may open the way to the development of a new conceptual and methodological framework to better understand the mechanisms underpinning key ecological processes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 May; accepted 18 December 2015.
Published online 10 February 2016.

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281,** 237–240 (1998).
2. Boyd, P. W. & Newton, P. Evidence of the potential influence of planktonic community structure on the interannual variability of particulate organic-carbon flux. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **42,** 619–639 (1995).
3. Guidi, L. *et al.* Effects of phytoplankton community on production, size, and export of large aggregates: a world-ocean analysis. *Limnol. Oceanogr.* **54,** 1951–1963 (2009).
4. Kwon, E. Y., Primeau, F. & Sarmiento, J. L. The impact of remineralization depth on the air-sea carbon balance. *Nature Geosci.* **2,** 630–635 (2009).
5. IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* (Cambridge University Press, 2013).
6. Kitano, H. Biological robustness. *Nature Rev. Genet.* **5,** 826–837 (2004).
7. Suweis, S., Simini, F., Banavar, J. R. & Maritan, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* **500,** 449–452 (2013).
8. Chow, C. E. T., Kim, D. Y., Sachdeva, R., Caron, D. A. & Fuhrman, J. A. Top-down controls on bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists. *ISME J.* **8,** 816–829 (2014).
9. Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459,** 193–199 (2009).
10. Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348,** (2015).
11. Giering, S. L. C. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507,** 480–483 (2014).
12. Azam, F. Microbial control of oceanic carbon flux: the plot thickens. *Science* **280,** 694–696 (1998).
13. Agusti, S. *et al.* Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nature Commun.* **6,** 7608 (2015).
14. Sancetta, C., Villareal, T. & Falkowski, P. Massive fluxes of rhizosolenid diatoms – a common occurrence. *Limnol. Oceanogr.* **36,** 1452–1457 (1991).
15. Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic north Pacific gyre at station ALOHA. *Mar. Ecol. Prog. Ser.* **182,** 55–67 (1999).
16. Omand, M. M. *et al.* Eddy-driven subduction exports particulate organic carbon from the spring bloom. *Science* **348,** 222–225 (2015).
17. Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean. *Science* **315,** 838–840 (2007).
18. Steinberg, D. K. *et al.* Bacterial vs. zooplankton control of sinking particle flux in the ocean's twilight zone. *Limnol. Oceanogr.* **53,** 1327–1338 (2008).
19. Turner, J. T. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Prog. Oceanogr.* **130,** 205–248 (2015).
20. Karsenti, E. *et al.* A holistic approach to marine eco-systems biology. *PLoS Biol.* **9,** (2011).
21. Strom, S. L. Microbial ecology of ocean biogeochemistry: a community perspective. *Science* **320,** 1043–1045 (2008).
22. Worden, A. Z. *et al.* Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347,** 1257594 (2015).
23. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348,** 1261359 (2015).
24. de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348,** 1261605 (2015).
25. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
26. Bork, P. *et al. Tara* Oceans studies plankton at planetary scale. *Science* **348,** 873 (2015).
27. Honjo, S., Manganini, S. J., Krishfield, R. A. & Francois, R. Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983. *Prog. Oceanogr.* **76,** 217–285 (2008).
28. Henson, S. A., Sanders, R. & Madsen, E. Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean. *Glob. Biogeochem. Cycles* **26,** (2012).
29. Lê Cao, K. A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* **7,** 35 (2008).
30. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20,** 947–959 (2010).

31. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature Rev. Microbiol.* **10,** 538–550 (2012).
32. Aylward, F. O. *et al.* Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc. Natl Acad. Sci.* **112,** 5443–5448 (2015).
33. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9** (2008).
34. Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front. Microbiol.* **6,** (2015).
35. Thomas, T. *et al.* Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. *PLoS ONE* **3,** (2008).
36. Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nature Rev. Microbiol.* **5,** 782–791 (2007).
37. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459,** 266–269 (2009).
38. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* **5,** e16 (2007).
39. Suttle, C. A. Marine viruses – major players in the global ecosystem. *Nature Rev. Microbiol.* **5,** 801–812 (2007).
40. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28,** 127–181 (2004).
41. Finkel, Z. V. *et al.* Phytoplankton in a changing world: cell size and elemental stoichiometry. *J. Plankton Res.* **32,** 119–137 (2010).
42. Sommer, U. & Lewandowska, A. Climate change and the phytoplankton spring bloom: warming and overwintering zooplankton have similar effects on phytoplankton. *Glob. Change Biol.* **17,** 154–162 (2011).
43. Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444,** 752–755 (2006).
44. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
45. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106,** 1374–1379 (2009).
46. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277,** 1300–1302 (1997).
47. Wymore, A. S. *et al.* Genes to ecosystems: exploring the frontiers of ecology with one of the smallest biological units. *New Phytol.* **191,** 19–36 (2011).
48. Picheral, M. *et al.* Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the *Tara* Oceans expedition 2009-2013. *PANGAEA* http://dx.doi.org/10.1594/PANGAEA.836319 (2014).
49. Picheral, M. *et al.* Vertical profiles of environmental parameters measured from physical, optical and imaging sensors during *Tara* Oceans expedition 2009-2013. *PANGAEA* http://dx.doi.org/10.1594/PANGAEA.836321 (2014).
50. Chaffron, S. *et al.* Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009–2013). *PANGAEA* http://dx.doi.org/10.1594/PANGAEA.840718 (2014).
51. Pesant, S. *et al.* Open science resources for the discovery and analysis of *Tara* Oceans data. *Sci. Data* **2,** 150023 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders, Rega Institute, KU Leuven, The French Ministry of Research, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218, SAMOSA, ANR-13-ADAP-0010), European Union FP7 (MicroB3/No.287589, ERC Advanced Grant Award to C.B. (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790 and #2631) and the UA Technology and Research Initiative Fund and the Water, Environmental, and Energy Solutions Initiative to M.B.S., the Italian Flagship Program RITMARE to D.I., the Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to S.G.A., TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca to S.G.A., JSPS KAKENHI grant number 26430184 to H.O., and FWO, BIO5, Biosphere 2 to M.B.S. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. This article is contribution number 34 of *Tara* Oceans.

Author Contributions L.G., S.C., Lu.B. and D.E. designed the study and wrote the paper. C.D., M.P., J.P. and Sa.S. collected *Tara* Oceans samples. S.K.-L. managed the logistics of the *Tara* Oceans project. L.G. and M.P. analysed oceanographic data. S.C. and Lu.B. analysed taxonomic data. S.C., Lu.B., D.E. and S.R. performed the genomic and statistical analyses. A.L., Y.D., L.G., S.C., Lu.B. and D.E. produced and analysed the networks. E.K., C.B. and G.G. supervised the study. M.S., J.R., E.K., C.B. and G.G. provided constructive comments, revised and edited the manuscript. *Tara* Oceans coordinators provided constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

Author Information Data described herein is available at European Nucleotide Archive under the project identifiers PRJEB402, PRJEB6610 and PRJEB7988, PANGAEA[48–50], and a companion website (http://www.raeslab.org/companion/ocean-carbon-export.html). The data release policy regarding future public release of *Tara* Oceans data is described in ref. 51. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.G. (lguidi@obs-vlfr.fr), S.C. (samuel.chaffron@vib-kuleuven.be), Lu.B. (lucie.bittner@upmc.fr), D.E. (damien.eveillard@univ-nantes.fr), J.R. (Jeroen.Raes@vib-kuleuven.be), E.K. (karsenti@embl.de), C.B. (cbowler@biologie.ens.fr) or G.G. (gorsky@obs-vlfr.fr).

***Tara* Oceans Consortium Coordinators**

Silvia G. Acinas[1], Peer Bork[2,3], Emmanuel Boss[4], Chris Bowler[5], Colomban de Vargas[6], Michael Follows[7], Gabriel Gorsky[8], Nigel Grimsley[9], Pascal Hingamp[10], Daniele Iudicone[11], Olivier Jaillon[12,13,14], Stefanie Kandels-Lewis[15,16], Lee Karp-Boss[4], Eric Karsenti[5,16], Fabrice Not[6], Hiroyuki Ogata[17], Stephane Pesant[18,19], Jeroen Raes[20,21,22], Christian Sardet[23], Mike Sieracki[24]†, Sabrina Speich[25], Lars Stemmann[8], Matthew B. Sullivan[26]†, Shinichi Sunagawa[15], Patrick Wincker[12,13,14]

[1]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Marítim de la Barceloneta 37-49, Barcelona E0800, Spain. [2]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. [3]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [4]School of Marine Sciences, University of Maine, Orono, Maine 04469, USA. [5]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France. [6]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [8]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Université Paris 06, CNRS, Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique de Banyuls, 66650 Banyuls-sur-Mer France, France. [10]Aix Marseille Université, CNRS, IGS, UMR 7256, 13288 Marseille, France. [11]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [12]CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [13]CNRS, UMR 8030, CP5706 Evry, France. [14]Université d'Evry, UMR 8030, CP5706 Evry, France. [15]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany. [16]Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1, 69117 Heidelberg, Germany. [17]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [18]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, 28359 Bremen, Germany. [19]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [20]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [21]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [22]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [23]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire de biologie du développement (LBDV), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. [24]Bigelow Laboratory for Ocean Science, East Boothbay ME 04544, USA. [25]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris CEDEX 05, France. [26]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.

†Present addresses: National Science Foundation, Arlington, 22230 Virginia, USA (M.S.); Department of Microbiology, and Department of Civil, Environmental and Geodetic Engineering, The Ohio Sate University, Colombus, Ohio 43210, USA (M.B.S).

2.4 Conclusions

Nevertheless, beyond the biological interest of WGCNA to investigate the plankton, two methodological points briefly discussed in the above study must be highlighted here. First, these studies [46, 90] emphasized the interest of choosing a great abstraction to explore the data without considering experimental data per se. In particular, we have shown herein that the use of the graph could be of great help to summarize a large quantity of homogeneous knowledge. However, such an abstraction necessitates discretizing the knowledge, via, for instance, the choice of a threshold. This particular step is critical and must be considered carefully, which justifies recent methodological efforts. The interest of Computer Science is to give herein access to several abstractions (from discrete to continuous via probabilistic ones). These abstractions are formal objects on which optimization techniques are performed, designed following parsimonious assumptions. These techniques thus aim at solving the original problem in an abstraction domain that becomes computationally tractable. The choice of one formal abstraction rather than another belongs to the first step of computational modeling, a step often neglected because of the broader access to sophisticated techniques via repository plate-forms. Beyond the interest of solving a biological question, an appropriate abstraction must produce an emerging property that could later on considered as another biological abstraction of interest if validated. For the sake of illustration, following the above studies, the module is becoming of interest in investigating the planktonic knowledge. By extension, one considers the modules as biological units that allow tackling biogeography questions. Figure 2.7 represents the contribution of eukaryotic taxa modules (resp. Fig. 2.7A) and prokaryotic genes modules (resp. Fig. 2.7B) associated to iron. These eight modules are the direct extension of WGCNA but are becoming as well an abstraction that reduces the whole Tara Oceans dataset complexity for the sake of global analysis [1].

Second, the use of graph abstraction proposes a new focus on the dataset. Instead of focusing on the most abundant species or the most variant species, the graph emphasizes species or genes or proteins that are central. Such a centrality is computed by different metrics that remain at the discretion of the scientists. Computer Sciences drive this change of paradigm and not standard statistical techniques, which could benefit to state-of-the-art ecological concepts such as niche [87], or keystone species [69]. However, the analysis of the biological graph remains at its infancy despite landmark studies. In particular, considering graphs to abstract biological data opens several methodological locks that the perspective

[1] extracted from Caputi L., Carradec Q., Eveillard D. et. al. Community-level response to the natural perturbation in open ocean planktonic ecosystems. Submitted

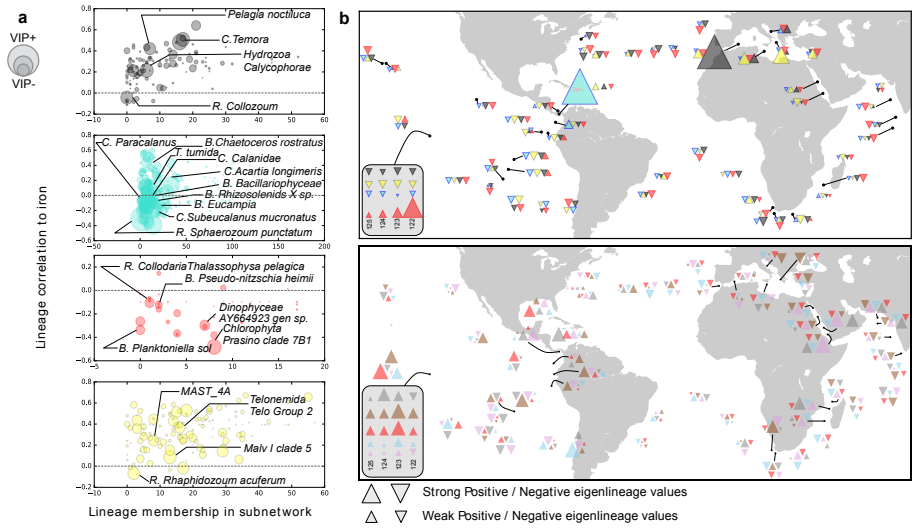


Figure 2.7: Planktonic Iron-Associated Assemblages (IAAs) in the global ocean and the Marquesas Islands stations. (A) Description of eukaryotic modules associated with iron. Relative abundances and co-occurrences of eukaryotic lineages were used to decipher modules. Four modules can predict iron with high accuracy. For each IAA, lineages are associated with their score of centrality (x-axis), to their correlation with iron concentrations (y-axis), and their VIP score (circle area). Circles depicted representative lineages within each module and named (C: Copepoda, B: Bacillariophyta, R: Rhizaria). (B) Top panel: contribution of Tara Oceans stations to the global variance of IAAs of eukaryotic lineages. For each IAA, we represent the projection of stations on the first principal component (upper panel). Lower panel: projection of the relative contribution of the Tara Oceans stations to the global variance of iron-associated prokaryotic gene assemblages, as revealed by WGCNA. For each prokaryotic gene module associated with iron, we represent the projection of stations on the first principal component, proportional to triangle sizes for each module. The inset shows the behavior of each IAA in the Marquesas archipelago stations.

section will further discuss.

Synthesis of publications that contribute to the analysis of omics experiments

- Luigi Caputi, Quentin Carradec, Damien Eveillard, Amos Kirilovsky, Eric Pelletier, Juan J Pierella Karlusich, Fabio Rocha Jimenez Vieira, Emilie Villar, Samuel Chaffron, Shruti Malviya, Eleonora Scalco, Silvia G Acinas,

Adriana Alberti, Jean-Marc Aury, Anne Sophie Benoiston, Alexis Bertrand, Tristan Biard, Lucie Bittner, Martine Boccara, Jennifer R Brum, Christophe Brunet, Greta Busseni, Anna Carratalà, Hervé Claustre, Luis Pedro Coelho, Sébastien Colin, Salvatore D'Aniello, Corinne Da Silva, Marianna Del Core, Hugo Doré, Stéphane Gasparini, Florian Kokoszka, Jean-Louis Jamet, Christophe Lejeusne, Cyrille Lepoivre, Magali Lescot, Gipsi Lima-Mendez, Fabien Lombard, Julius Lukeš, Nicolas Maillet, Mohammed-Amin Madoui, Elodie Martinez, Maria Grazia Mazzocchi, Mario B Néou, Javier Paz Yepes, Julie Poulain, Simon Ramondenc, Jean-Baptiste Romagnan, Simon Roux, Daniela Salvagio Manta, Remo Sanges, Mario Sprovieri, Vincent Taillandier, Atsuko Tanaka, Leila Tirichine, Camille Trottier, Julia Uitz, Alaguraj Veluchamy, Jana Veselá, Flora Vincent, Sheree Yau, Stefanie Kandels-Lewis, Sarah Searson, Céline Dimier, Marc Picheral, Peer Bork, Lionel Guidi, Paolo Sordino, Matthew B Sullivan, Alessandro Tagliabue, Adriana Zingone, Laurence Garczarek, Fabrizio d'Ortenzio, Pierre Testor, Maurizio Ribera d'Alcalà, Patrick Wincker, Chris Bowler, Tara Oceans coordinators, Emmanuel Boss, Colomban Vargas, Michael J Follows, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Lee Karp-Boss, Eric Karsenti, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen Raes, Emmanuel G Reynaud, Christian Sardet, Mike Sieracki, Sabrina Speich, Lars Stemmann, Shinichi Sunagawa, Didier Velayoudon, and Jean Weissenbach. Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Global Biogeochemical Cycles*, 4(30):10,438, March 2019

- Dinka Mandakovic, Claudia Rojas, Jonathan Maldonado, Mauricio Latorre, Dante Travisany, Erwan Delage, Audrey Bihouée, Géraldine Jean, Francisca P Díaz, Beatriz Fernández-Gómez, Pablo Cabrera, Alexis Gaete, Claudio Latorre, Rodrigo A Gutiérrez, Alejandro Maass, Verónica Cambiazo, Sergio A Navarrete, Damien Eveillard, and Mauricio González. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific reports*, 8(1):5875, April 2018

- Stanislas Thiriet-Rupert, Gregory Carrier, Camille Trottier, Damien Eveillard, Benoit Schoefs, Gael Bougaran, Jean-Paul Cadoret, Benoit Chénais, and Bruno Saint-Jean. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of Tisochrysis lutea. *Algal Research*, 30:59–72, March 2018

- Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline,

Jennifer R Brum, Jennifer Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Tara Oceans Consortium Coordinators, Lars Stemmann, Fabrice Not, Pascal Hingamp, Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stéphane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Peer Bork, Colomban de Vargas, Daniele Iudicone, Matthew B Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470, February 2016

- Silvia E Newell, Damien Eveillard, Mark J McCarthy, Wayne S Gardner, Zhanfei Liu, and Bess B Ward. A shift in the archaeal nitrifier community in response to natural and anthropogenic disturbances in the northern Gulf of Mexico. *Environmental Microbiology Reports*, 6(1):106–112, February 2014

- Richard A Long, Damien Eveillard, Shelli L M Franco, Eric Reeves, and James L Pinckney. Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. *FEMS microbiology ecology*, 83(1):74–81, January 2013

- Nicholas J Bouskill, Damien Eveillard, Diana Chien, Amal Jayakumar, and Bess B Ward. Environmental factors determining ammonia-oxidizing organism distribution and diversity in marine environments. *Environmental Microbiology*, 14(3):714–729, March 2012

- Nicholas J Bouskill, Damien Eveillard, Gregory O'mullan, George A Jackson, and Bess B Ward. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environmental Microbiology*, 13(4):872–886, April 2011

- Géraldine Del Mondo, Damien Eveillard, and Irena Rusu. Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions. *Bioinformatics (Oxford, England)*, 25(7):926–932, April 2009

- Sébastien Angibaud, Damien Eveillard, Guillaume Fertin, and Irena Rusu. Comparing bacterial genomes by searching their common intervals. In *Bioinformatics and Computational Biology*, pages 102–113. Springer, 2009

- Abalo Chango, Afif Abdel Nour, Souad Bousserouel, Damien Eveillard, Pauline M Anton, and Jean-Louis Guéant. Time course gene expression in

the one-carbon metabolism network using HepG2 cell line grown in folate-deficient medium. *The Journal of nutritional biochemistry*, 20(4):312–320, April 2009

- Bess B Ward, Damien Eveillard, Julie D Kirshtein, Joshua D Nelson, Mary A Voytek, and George A Jackson. Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology*, 9(10):2522–2538, October 2007

Chapter 3

Integrative Biology: Understanding Biological Systems through the integration of heterogeneous data

Nature is ever at work building and pulling down, creating and destroying, keeping everything whirling and flowing, allowing no rest but in rhythmical motion, chasing everything in endless song out of one beautiful form into another.

John Muir

Computer Sciences research applied to integrative biology is still in its infancy for solid reasons. So far, integrative biology appears as a *meta-biology*, in which all aspects of the biological system, previously studied as separate objects, are treated together, while one acquires their different knowledge at different levels of advancement. Because the description of the biological system is sparse, computer sciences developed fragmented and heterogeneous methods for individual aspects. However, the current demand for integrative biology is to move to a higher level of integration of biological abstractions, which does not consist of a simple parameterization of more, but a real integration of concepts, models, bioinformatics methods themselves. This formalization must be done step by step, aspect by aspect,

and must rely on several kinds of knowledge to produce integrative knowledge. Once again, modeling reverts particular importance. Biological abstractions used for fragmented biological knowledge must be either transformed into another one from automatic treatments that compare an abstraction with a knowledge database (see Section 3.1 for illustration) or merged with other abstractions to produce another distinct one (see Section 3.2 or 3.3). Overall, these transformations that could be linked to the field of model engineering as promoted for software engineering must promote the emergence of new properties that were not available without such integration. Thus, the primary objective of integrative biology consists in deciphering potentially new functional features, that could take the form of new phenotypic descriptions. This chapter will present how the use of optimization assumptions could help to detangle new biological properties from different experimental descriptions. Note herein integrative biology also saw the rise of linked data techniques [14, 221], which will not be discussed in this chapter for the sake of concision.

3.1 Inferring metabolic networks: integration of genomics contents with biological & chemical knowledge

The metabolic network relies on the chemical knowledge that identifies the role of different enzymes to catalyze the transformation of molecules into others. In the 2000s, as discussed in the previous chapter, the development of high-throughput sequencing techniques promote the acquisition of extensive data on the genes content of organisms (i.e., a genome) or even a community of organisms (i.e., a metagenome). By (almost) direct translation, one assumes the presence of a gene that encodes for an enzyme as a signal for the alleged use of a metabolic reaction within the organism. In particular, these computational techniques that use state-of-the-art gene database help to identify, for a given organism, a catalog of catalytic proteins potentially produced by this organism. In the context of metabolism, metabolic reactions are feasible because of the presence of the catalyzer, which allows the consumption of given compounds that will be transformed by the production of other ones (see Figure 3.1 for illustration). The checking of gene content represents a step called metabolic mapping.

However, since the genetic information is highly incomplete despite substantial sequencing depth, and the chemical knowledge do not necessarily cover all biochemical reaction specificities in a given organism (in particular on the balance of co-factors), the sole metabolic mapping remains limited to characterize the full metabolic capability of an organism without further analysis. Modeling such a metabolic network can take either the form of a graph or a stoichiometric matrix.

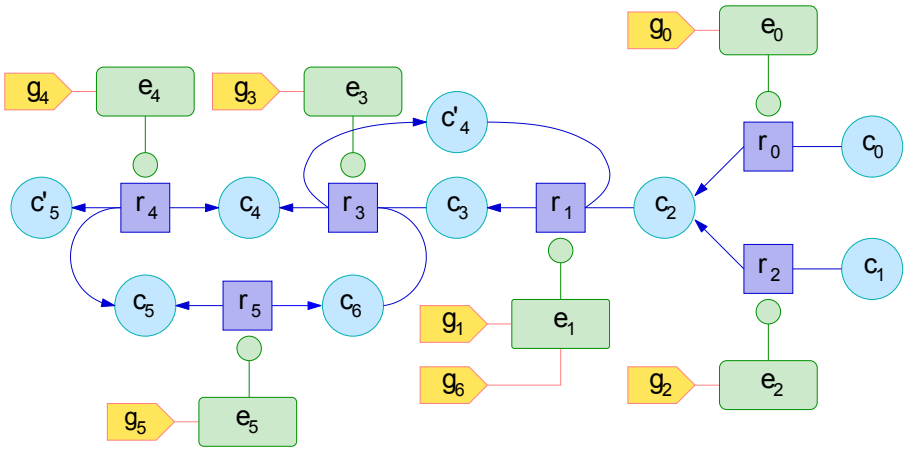


Figure 3.1: Schema of a metabolic network as built from genomic knowledge. Each gene (yellow arrow or g_x) encodes for one enzyme (green rectangle or E_x). An enzyme can be encoded by several genes when several protein subunits are necessary (ex. g_1 and g_2 necessary for producing E_1). The presence of enzymes allows metabolic reactions to take place (purple square or r_x). Chemical database depicts thermodynamical constraint that results in reversible or irreversible reactions if one takes reactions in both directions or only one. The same database indicates chemical compounds (blue circle or c_x) that are produced or consumed by each reaction. Two reactions are linked to each other when the product of one reaction is the substrate of the other one. The interplay between reactions thus describes a metabolic network, that is a bipartite graph directed by the chemical knowledge ©Philippe Bordron

This latter form will be discussed later in Chapter 4.3, whereas one will focus on the graph in this Chapter. Despite this methodological antagonism, Thiele and Palsson present in their work [193] a global method of metabolic network reconstruction by trying to homogenize the different protocols for the creation of these networks. Another review [92] proposes a complementary methodology based on previous work, where the authors compare various reconstruction tools applied in particular to prokaryotic networks. Overall, one could divide the metabolic reconstruction into four stages:

1. construction of a draft metabolic network following the metabolic mapping
2. improvement of the draft
3. conversion of the network into a format usable by automatic methods

4. evaluation of the quality of the final network
5. reiterate above steps until biological satisfaction

Each of these steps represents itself a field of research, but today steps two and three are often done together because of recently improved methods that perform these steps automatically. Also, the final step usually results from expert biological annotations and determines if further repetitions of these steps are necessary before releasing the model in the public repository (for instance, Bigg [112]). Beyond the sole validation of the network, this final step is also fundamental because it drives the purpose of the model. Indeed, without starting a philosophical discussion here, the resulting metabolic model aims at reproducing a given phenotype in given environmental conditions (i.e., a quantity of substrates). However, without extensive consideration of this final step, one could over-interpret a metabolic model. Indeed, the model simulation outside of its validation conditions may lead to false simulations. Certainly, no model is universal, and one must consider with secure care the following steps to avoid model extrapolations.

Creating a draft metabolic network

The first step builds a draft of the metabolic network by extracting information from two distinct biological sources. On the one hand, one considers functional annotations of the genome, such as:

- EC number, which makes it possible to classify the enzymes according to the reaction which they catalyze,
- GO term, related to the ontology of genes in species,
- Generic names of reactions as proposed by biological expertise

As seen above, all of these data, when available, pinpoint which enzymes are potentially available, thus determining the reactions that take place in a metabolic network, but also which proteins or metabolic compounds are consumed or produced. On the other hand, one could retrieve complementary information from neighboring species. In particular, when gene information is missing, one could search for gene homologies from HMMs (Hidden Markov Models) profiles for finding enzymes for non-annotated genes. In particular, one could recover this information from databases such as KEGG [104] or METACYC [32]. However, because one performs these functional annotations via alignment scores with a reference database that stores knowledge of reference species, the draft will be more accurate if the species that one investigates are sufficiently phylogenetically close to reference species.

Refinement of the metabolic network and conversion to an analyzable format

Once drafted, the metabolic network should be verified and completed such that one can use it in further analysis. Refinement consists of verifying biological properties from the network. For instance, one can investigate the topology of the bipartite graph that represents the metabolic network by checking if there exists a path that links metabolites that are nutrients of the given organism (i.e., called source) to those that are produced by the same organism (i.e., called targets). By extension, one often considers these metabolites as elements involved in the organism's growth [173]. Another quantitative technique can be used here via the use of Flux Balance Analysis, later described in Sec. 4.3. When the draft network fails to reach phenotypic expectations, it is then necessary to add reactions in order to obtain the production of all the targets. This task could be made manually or automatically via gap-fill techniques. They consist of finding missing reactions that one must add in the draft. Choosing these reactions consists of optimizing the shortest path problem [106]. In collaboration with the team DYLISS in Rennes, we recently proposed a new gap-filling method that focuses on searching all the filled draft metabolic networks that satisfy biological expertise queries and merged them into a unique network [36, 164]. The draft network is filled using Answer Set Programming framework (ASP) to satisfy specific in silico queries, e.g., relating to its capacity to produce given molecules, from given substrates. Among others, automatic tools perform some of these analyses as soon as the network and its associated knowledge are formatted in a standardized way [136]. In particular, one of the most widely used formats is the SBML format, for Systems Biology Markup Language, which first lists all the metabolites involved in the network and then lists the reactions with their substrates and products.

Notice herein that the group of Patil recently proposed a complementary technique at EMBL that aims to carve a generic prokaryotic metabolic network to fit the genomic content while maintaining generic properties [135]. This technique is in the opposite direction of standard gap-fill techniques but presents great promises for modeling more exotic bacterial strains, such as those that cannot be cultivated but observed from in situ metagenomic data.

3.2 Operons & Regulons as emerging features of metabolic and genomic knowledge integration

Above mentioned biological abstractions allow us to understand two particular characteristics of living systems: (i) the genotype or distribution of genes in a genome and (ii) the phenotype or functional characteristics of living organisms

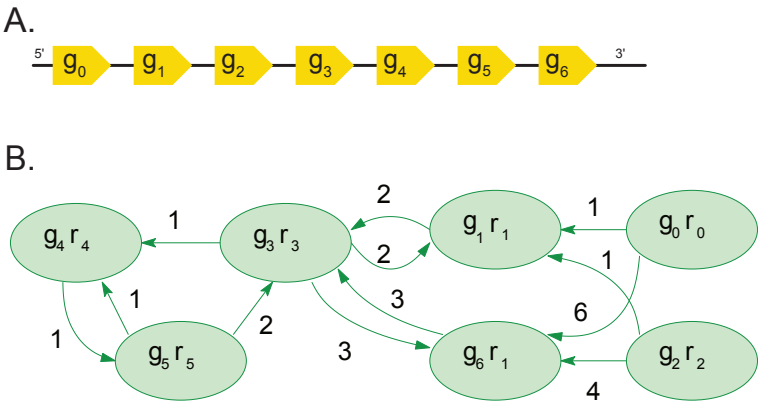


Figure 3.2: Schema of an integrated graph that resumes genomic and metabolic knowledge. A set of genes (yellow arrow or g_x) is ordered to build a genome (or chromosome). Each of these genes is separated by intervals that can gene measured as gene count or pair base count. From the metabolic network depicted in Figure 3.1, one can build a directed weighted graph in B., where a node represents the dual genomic and metabolic knowledge (g_x, r_y) with g and r stand for genes and their encoded reactions respectively. Directions of edges correspond to the directions between two reactions as imposed by thermodynamical constraints. The weights on edges represent the numbers of gene interval between two genes pointed as sources and targets of the given edge. ©Philippe Bordron

such as metabolisms. To integrate these abstractions, one must consider two characteristics that have been previously studied. First, as mentioned in Section 3.1, a metabolism, modeled as a metabolic network, can be described as a set of paths of biological signal travels in an oriented graph. For prokaryotes, the distance traveled by the signal from one reaction to another is then a phenotypic characteristic of importance that follows a parsimonious assumption. Second, the genomic content of a given prokaryotic system describes a potential phenotype (i.e., a set of biological functions). However, one could use as well the gene order within the genome (or chromosome) to compare one organism with other prokaryotic strains [65]. In particular, the number of gene order rearrangement is a distance metric that compares bacteria following (again) a parsimonious assumption. Comparing both distances, i.e., associated with both phenotypic and genotypic abstractions, is a general issue in biology at the era of omics data. This comparison is not directly accessible by standard experimental techniques and requires a bioinformatics contribution. From the applicative point of view, it is worth noticing that this

issue is also at the heart of modern comparative genomics, which can no longer afford to focus on genome sequences alone. However, it must already integrate data from macromolecular networks such as protein-protein interactions or metabolic networks [24]. Our contribution to compare both abstractions relies on a formal description of the problem for the sake of new algorithm design. The problem data are (i) a network-oriented graph, whose arcs are labeled with protein names, and (ii) a sequence of characters representing the corresponding genome, each character of which is a gene generating a protein from the oriented graph. When two arcs are fixed in the graph, labeled with two genes g_x and g_y, several oriented paths with both arcs as ends exist. The diversity of these paths is a functional index that must be further studied, and that could be used to compare bacterial species on a phenotypic basis. In the following study enclosed below:

Philippe Bordron, Damien Eveillard, and Irena Rusu. Integrated analysis of the gene neighbouring impact on bacterial metabolic networks. *IET systems biology*, 5(4):261–268, July 2011

we propose to investigate all the paths that link two given genes (or corresponding encoded metabolic reactions). Among them, following a parsimonious assumption, only the shortest ones are discussed and compared with, at the time, biological knowledge. To our surprise, one could associate the shortest genome segments with already known operons or regulons (i.e., functional units for microbial systems) with 54% of accuracy. This result is of particular interest because one considers only topological knowledge, whereas most efficient state-of-the-art operon prediction techniques use machine learning approaches that involve training datasets. From a computational viewpoint, one computes these shortest genome segments, called SGS in the following, via the brute force algorithm [16]. For the sake of computational efficiency, in [15], we proposed another implementation encoded as a Logic programing via Answer Set Programming [79]. As mentioned above, when building metabolic networks, the benefit of this programming paradigm relies on the design of the problem rather than its resolution *per se*. The problem, once formulated, will be translated for being solved by dedicated and efficient solvers.

Worth the notice, this biological motivation consists of solving a general graph covering problem since a genome, being a succession of genes, is also a linear graph. Considering the general problem, one could extend the SGS identification to other graphs. For instance, we recently proposed an extension of this problem to the integration of genomic and transcriptomic knowledge [195]. For this purpose, one abstracts gene expressions in several experimental conditions as a Gene Regulatory Network (GRN). This new kind of graph is distantly related to co-occurrence networks discussed above and will be more discussed in Sec. 4. GRN and genome are two graphs that one could integrate to emphasize functional units. They will

represent sets of genes that are nearby in the genome while being connected to the GRN. Here, SGS cannot be directly related to operons but could foster the investigation of the set of functional units that are associated with given conditions such as distinct growth media.

Published in IET Systems Biology
Received on 26th July 2010
Revised on 24th March 2011
doi: 10.1049/iet-syb.2010.0070

ISSN 1751-8849

Integrated analysis of the gene neighbouring impact on bacterial metabolic networks

P. Bordron D. Eveillard I. Rusu

Computational Biology Group (ComBi) – LINA, Université de Nantes, CNRS UMR 6241, 2 rue de la Houssinière, 44300 Nantes, France
E-mail: philippe.bordron@univ-nantes.fr

Abstract: Different levels of abstraction are needed to represent a living system. Unfortunately information of different nature is not superposable in an obvious way, but requires a dedicated framework. Because biological abstractions, i.e., genomic or metabolic information, can be easily respresented as graphs, it is intuitive to integrate them into a unique graph, in which one can perform graph analysis for investigating a given biological assumption. This study follows such a philosophy and completes a genome and metabolome combination. In a such integrated framework and as illustration, we applied a graph analysis that automatically investigates impacts of the gene adjacency to predict functional relationships between genes and reactions. Our approach, called SIPPER, creates a weighted graph, in which the weights rely on the given relationship between genes, and computes (alternative) chains of reactions catalysed by genes. This method, as a generalisation of methods already published, can be easily adapted to several biological assumptions, properties or measures. This paper evaluates SIPPER on *Escherichia coli*. We automatically extract subgraphs, called *k*-SIPs, and quantify their interest in both genomic and metabolic contexts by showing functional compounds like operons or functional modules.

1 Introduction

A living system is represented via different types of information, that result from distinct experiments. They depict the system with distinct detail degrees or at different biological abstraction levels (i.e. genome, metabolome etc.). The integration of data coming from these different sources is an unavoidable approach to identify modules supported by several data types, thus predicting protein functions and interactions. Unfortunately, information of different nature is not superposable in an obvious way, and dedicated approaches are usually developed for specific types of information. However, the need to combine heterogeneous information and to analyse it as a whole increases constantly. For example, the functional interpretation of genomes depicts the potential function of a given species, but not its phenotype expressed under particular environmental conditions. Bridging this gap between function and genome can be made via the integration of metabolic knowledge and the relationship between genes that notably describe operons. Therefore generic methods to (i) gather information together and to (ii) represent it in a suitable automatic way for exploration are necessary today.

Several fields of biological investigations follow such an integrative philosophy, each focusing on a dedicated assumption. Like this, comparative genomics uses the hypothesis that conserved groups of contiguous genes among prokaryotic genomes are conserved during evolution [1]. It suggests that the species survival partially depends on the relationship between the genes within a conserved group, which may monitor a major function. The study of the gene co-expression also shows that genes are more or less strongly linked together, under given environmental conditions for a given behaviour [2]. Such relationship or biological assumptions are taken into account according to different manners. Some studies focus on the notion of connectivity [3], which informs us about the existence of a relationship between biological compounds. Some others [4] are based on the notion of distance, which informs about how strong the relationship between biological compounds is.

When one is interested in integrating heterogeneous biological information, a generalisation of these methods relies on the use of the concept of connectivity between biological compounds, which is easily handleable in networks. The approach offers herein an implementation, called SIPPER [5], of such a concept. It generalises previous works aiming at integrating, for instance, genomic and metabolic data [4, 6, 7], or genomic, co-expression and metabolic data [8–11]. As a result, SIPPER provides a graph that integrates heterogeneous biological knowledge. Each edge of this graph is then weighted by a given connectivity measure or a given distance. Beyond this new, but natural and automatic abstraction of the biological system, such an approach provides as well a wide range of analyses for investigating the integrated graph.

It is worthwhile to underline here that devising general methods that integrate heterogeneous information has genuine limits imposed by the exceptional diversity of the data and its interpretations. Our approach is independent of

the distance, but interpretations of the results are not. To perform a complete explanation of our framework, and for the sake of clarity, we choose herein to illustrate it on the integration of two kinds of biological knowledge, the genome and the metabolism of *Escherichia coli*; and a particular distance between genes, the 'gene neighbouring distance' defined as the number of intermediate genes along the genome between two given genes, plus 1. This choice relies on two points. Firstly, the genome and the metabolism of *E. coli* are, by far, the most studied information of a well-investigated species. Moreover, as we work on prokaryotes, a linear relationship between the bacterial genes and protein activities allow a natural integration between genome and metabolism (via the enzymes, products of genes) within a unique graph. Secondly, the neighbouring of genes takes on a major importance for investigating the prokaryotic functions. It is indeed commonly accepted that the gene order in bacterial genomes is far from random [12, 13], which represents a way to compare genomes, finding functional modules or predicting operons using automatic approaches [14–21].

As a concrete application result of SIPPER, this article illustrates our approach by first (Section 2.1) showing the integration of metabolic and bacterial genome information from *E. coli*. We thus introduce the gene neighbouring distance and how such a distance contributes to build a weighted graph. The resulting integrated model is suitable for standard graph analyses, as depicted in Section 2.2. Their use will extract subgraphs of interest, called *k*-shortest integrated paths (*k*-SIP). The sequel will propose to investigate the biological meaning of *k*-SIPs compared to functional knowledge like operons or KEGG (Kyoto Encyclopedia of Genes and Genomes) modules (Section 2.3). As described in Section 3 and discussed later (Section 4), SIPPER will emphasise the set of the biological compounds that convey biological meanings, like operons but also the couples and triples of operons associated for functional reasons, or functional modules but also couples of functional modules. From a methodological point of view, and compared to other approaches, we will discuss in particular how SIPPER permits a decent prediction of operons and functional modules despite its genericity. Finally, we will describe and discuss the impact of the gene neighbouring measure for identifying functional units. In particular, the sequel will show that this intergenic distance is not self-sufficient for an accurate functional prediction, confirming other independent studies.

2 Material and methods

SIPPER is a generic method, whose full algorithmic details can be found in [5]. However, for the sake of clarity, we intend in the sequel to illustrate SIPPER by its step-by-step application on a concrete biological system and a given measure. Owing to the linear relationship between the gene activity and its encoded protein production, SIPPER is automatically applied on the integration of a microbial genome and its corresponding metabolic information. Among the microbial phylum, *E. coli* represents the most-studied organism for which genome and metabolome are at our disposal. Moreover, the neighbouring of genes which gives information about the proximity of genes on a chromosome, is one of the most used biological measures to investigate genomes and determine ab initio functional modules.

2.1 Integrated model

A bacterial genome is represented as a linear or circular sequence of genes (see Fig. 1*a*). Each gene produces one or multiple proteins. Each protein catalyses one or several metabolic reactions. Thus, the gene's products impact on the chain of reactions that occur in the metabolic network (Fig. 1*b*), that one considers as a directed graph. This observation leads us to define an integrated genomic metabolic network, denoted $\mathcal{G}_{\text{int}}$. The network $\mathcal{G}_{\text{int}}$ is a directed graph, whose vertices are all the pairs (gene g, reaction r) such as the gene g produces an enzyme (identified herein by its EC number) that catalyses the reaction r. An arc goes from vertex (g_1, r_1) to vertex (g_2, r_2) whenever a product of r_1 is a substrate of r_2. Its weight w expresses the studied relationship between g_1 and g_2, which is the gene neighbouring distance in Fig. 1*c*. The coupling between each gene and reaction in $\mathcal{G}_{\text{int}}$ is an important property. It is more informative than creating an enzyme network [22]. Indeed, an enzyme can be produced

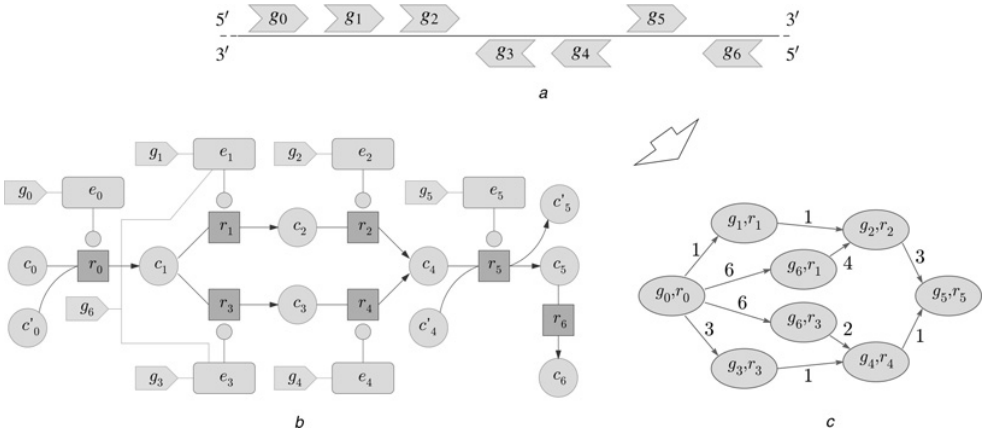


Fig. 1 *Integration of genomic and metabolic information into* $\mathcal{G}_{\text{int}}$

a Bacterial monochromosomic genome is a linear or circular sequence of genes (fat arrows)
b Its corresponding metabolic network (in SGBN standard): compounds (circles) are substrates and/or products of reactions (squares), that are catalysed by enzymes (rounded boxes) produced by genes (fat arrows)
c Resulting integrated network $\mathcal{G}_{\text{int}}$, where each arc is weighted by the gene neighbouring distance between the two genes in its endpoints

doi: 10.1049/iet-syb.2010.0070

by homologous genes, which implies different genomic relationship between the same enzymes.

E. coli (K12 MG1655, 31 March 2008 release, [23]), as mentioned above, is an accurate benchmark for evaluating SIPPER. At the time of the study, a set of 4242 genes composes its circular monochromosomic genome (NCBI/GenBank). Its corresponding metabolism is taken from KEGG PATHWAYS database (KEGG PATHWAYS, 21 October 2008 release, [24]) which proposes the metabolic network as a set of pathways maps. Note that in each pathways map, common ('hub') metabolites, that is compounds that participate in more than 25 reactions, have been removed to avoid bringing reactions together in a way that is not biochemically informative [25]. The whole KEGG metabolic network is then reconstructed by taking, in each map, the reactions and associated biological compounds. The connection between a reaction and a compound exists in the whole metabolic network when the reaction and the compound are connected in at least one KEGG pathway map. At the end, the whole bacterial metabolic network is composed of a set of 2971 biochemical compounds involved in 1131 reactions catalysed by 647 enzymes. Among them, only 558 are encoded by identified genes, as indicated by NCBI/GenBank. For obvious technical reasons, SIPPER only takes into account the reactions catalysed by these enzymes for the generation of $\mathcal{G}_{\text{int}}$ concerning *E. coli*. The resulting $\mathcal{G}_{\text{int}}$ of *E. coli* is composed of 2343 vertices and 13 288 arcs, which correspond to 1049 metabolic reactions (92.75% of the *E. coli* reactions) and 779 genes (18.36% of the bacterial genome).

Additionally, one must consider a biological property, or assumption, for which we are interested in evaluating the impact. As suggested above, the adjacency property, also called herein gene neighbouring, is abstracted by the gene neighbouring distance, which is the number of intermediate genes between the two genes along the genome, plus 1 [26]. By convention, the genomic distance between a gene and itself is null. In a circular genome, the distance consists of the minimum one obtained from the right-hand and left-hand traversal. Evaluating the impact of the gene neighbouring property consists in considering the genomic distance between two genes as the weight between their respective vertices in $\mathcal{G}_{\text{int}}$.

2.2 Shortest integrated paths

Intuitively, a shortest (directed) path from $\mathcal{G}_{\text{int}}$ represents a reaction chain from the metabolic network whose genes encoding for the reactions are strongly related according to the defined property between them. Such a biological property is measured by a distance between vertices, herein the gene neighbouring measure. Finding these paths takes on a particular interest from a functional viewpoint, since the involved genes are linked by both a metabolic feature and a distance within the genome.

Given any pair of reactions involved in the vertices of $\mathcal{G}_{\text{int}}$, which we identify as the source reaction and the destination reaction, we are interested in the (directed) paths in $\mathcal{G}_{\text{int}}$ that start with a vertex containing the source reaction and end with a vertex containing the destination reaction. We then define the 'neighbouring coefficient $\bar{w}$' of a given path as the ratio between its total weight and the number of its distinct reactions. Intuitively, the neighbouring coefficient measures the average genomic distance between two genes involved in successive reactions from the path. A path in $\mathcal{G}_{\text{int}}$ with the smallest $\bar{w}$, that is neighbouring coefficient, is called the – '1-SIP' of the two reactions. This path represents a way to join two given reactions while preserving the minimum genomic distance along the genome. Fig. 2*a* shows an example of 1-SIP.

By extension, for a fixed positive integer k, the subnetwork of $\mathcal{G}_{\text{int}}$ obtained as the union of the k distinct paths with smallest $\bar{w}$ joining two given reactions is called the 'k-SIP' of the two reactions. It represents alternative ways to join the two given reactions. Fig. 2*b* shows an example of a 10-SIP. All the k-SIPs (for a given k) from $\mathcal{G}_{\text{int}}$, over all possible pairs of source and destination reactions, are computed. A k-SIP is generated by using a heuristic that we obtained by modifying Yen's algorithm [27]. It calculates the k minimum weighted circuit-free paths in a weighted directed graph. It has a running time of $O(kn(m + n \times \log n))$, where n is the number of vertices and m is the number of arcs in the network.

2.3 Evaluating the biological interest of a k-SIP

2.3.1 Benchmark datasets: In a wide analysis context, each given k-SIP must be compared with precise collections of biologically meaningful entities. Genes that belong to a unique transcriptional unit are, by nature, contiguous along a given bacterial genome and share a common biological function [28]. Operons of *E. coli* are taken from RegulonDB (Release: 6.3, 30 January 2009) [29]. They are transcriptional units that are composed of, at least, two genes. We selected 'metabolic operons' that describe transcriptional units of, at least, two genes that catalyse metabolic reactions. Herein, a k-SIP is said to match exactly one or many operons, when the set of genes of the k-SIP equals the set of genes of one or many operons. The resulting benchmark contains 135 metabolic operons (16.2% of the whole set in *E. coli*).

The KEGG database provides small pieces of biochemical pathways (i.e. modules) manually defined as either molecular complexes, consecutive reaction steps, regulatory, phylogenetic or functional units [30]. A k-SIP matches exactly one or many modules when the set of reactions of the k-SIP is the set of reactions of one or many KEGG modules. We considered KEGG modules that exist in

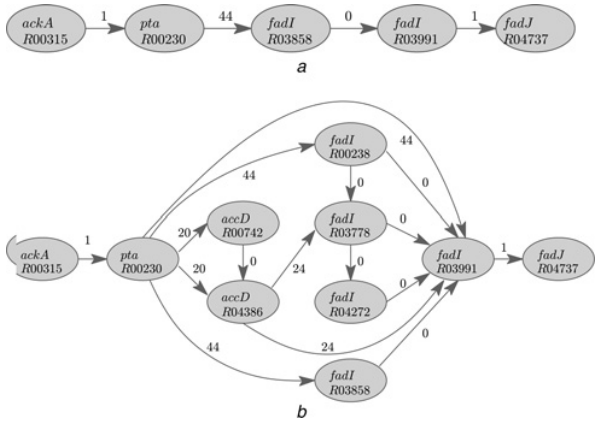


Fig. 2 *Example for a given pair of reactions (from R00315 to R04737) in E. coli*

a 1-SIP ($\bar{w} = 9.2$)
b 10-SIP ($\bar{w} = 19.8$)
It is easy to notice here that paths in the 10-SIP are mainly variants of the 1-SIP that share many common vertices and arcs

E. coli. They are composed of, at least, two enzymatic reactions. They constitute a benchmark of 99 metabolic modules (among 689 modules – 14.39%) of the whole KEGG database for all listed species).

2.3.2 Benchmark measures: In order to compare a *k*-SIP with the above biological datasets, we used 'Jaccard's measure'. As an illustration, applied to the comparison of a *k*-SIP and a given operon, this measure consists in the number of common genes shared by the *k*-SIP and the operon, divided by the total number of distinct genes within the *k*-SIP and the operon. Jaccard's measure takes values between 0 and 1. The higher the value of Jaccard's measure is, the more similar are the *k*-SIP and the operon. When Jaccard's measure is equal to 1, the *k*-SIP and the operon match exactly. When the measure equals 0, the *k*-SIP and the operon are distinct.

As a complementary measure, the coverage of a given operon by a *k*-SIP is the number of common genes shared by the *k*-SIP and the operon, divided by the number of distinct genes in the operon. The coverage takes values between 0 and 1. When the coverage equals 1, the operon is fully included in the *k*-SIP. The *k*-SIP is called then fully covering the operon. When the covering measure equals 0, the operon is not covered by the *k*-SIP. Otherwise, when the covering measure is between 0 and 1, the operon is partially covered by the *k*-SIP.

We complete our study by performing random shuffling experiments independently on the genome (by randomly modifying the gene order) and the metabolic network (according to [31], by randomly shuffling the endpoints of the arcs in a compound-free representation of the metabolic network called representation graph). Note that such an approach might decrease the number of arcs in the network, and consequently some vertex degrees.

3 Results

3.1 *k*-SIP of *E. coli*

The application of SIPPER on *E. coli* produces 439 382 *k*-SIPs for each *k*. As an illustration, Fig. 2 shows the 1-SIP and the 10-SIP for the reactions R00315 and R04737. In average (and respective standard deviation), a 1-SIP and a 10-SIP contain respectively 11.8 ($\pm$4.5) and 13.8 ($\pm$4.7) genes; and 13.9 ($\pm$5.6) and 17.6 ($\pm$6.2) reactions. Such numbers highlight the fact that adding alternative paths (i.e. from 1 to 10) reveals a weak impact on the number of genes and reactions used in a *k*-SIP.

3.2 Comparing a *k*-SIP with the operon dataset

3.2.1 Impact of the neighbouring coefficient $\bar{w}$: We study the evolution of the rate of *k*-SIPs that exactly match operons (or 'operonic confidence rate') regarding the $\bar{w}$ of the *k*-SIPs. This analysis was performed for several *k* values. For the sake of illustration, Fig. A.1 in Supplementary material, depicts the evolution of this rate when *k* is equal to 1. As a synthesis of all the results, Table 1 shows the comparison between the rate of operons exactly matching a *k*-SIP ($k = 1$ and 10) in *E. coli* and in the shuffled data. Clearly, a random gene order (row *k*-SIPs with shuffled genome) significantly changes the genomic distances between the genes belonging to the same operon of *E. coli*, leading to a very important increase of $\bar{w}$ for the 1-SIPs that exactly match the operon. These 1-SIPs (and the corresponding operons) still exist (the metabolic network did not change), but one cannot identify them. A random metabolic network (row *k*-SIPs with randomised metabolism) significantly changes the paths with respect to the network of *E. coli*, and the result is a very small number of operons exactly matching a 1-SIP only by chance. In *E. coli*, *k*-SIPs for both $k = 1$ and 10 show similar results: the rate of exactly matched operons raises and describes a plateau when $\bar{w}$ increases. As a matter of fact, even if $\bar{w}$ is the SIPPER's primer search criterion, minimising $\bar{w}$ is not a sufficient criterion for optimising the selection of *k*-SIPs that exactly match operons.

Table 1 Summary of the exact matches between *k*-SIPs ($k = 1$ and 10) and the operons from *E. coli*, in function of $\bar{w}$ values

Dataset	*K*	Rate of operons exactly matched by *k*-SIPs, %		
		$\bar{w} \leq 1.0$	$\bar{w} \leq 5.0$	$\bar{w} \leq 200.0$
k-SIPs	1	23.71	24.44	24.44
	10	8.15	12.59	12.59
k-SIPs with	1	0	0	2.22
shuffled genome	10	0	0	0.74
k-SIPs with	1	1.48	1.48	1.48
randomised metabolism	10	0	0	0

3.2.2 Impact of the parameter k and prediction of operons: Impact of k: Since $\bar{w}$ is not a self-sufficient criterion to automatically predict operons, we propose now to focus on the parameter *k* that represents the number of alternative paths within a *k*-SIP. For each *k* from 1 to 10, we compared each operon with each *k*-SIP. In all, 24.4% of the operons match exactly a 1-SIP each (namely, 33 over 135 operons, see Table A.1 in the Supplementary material for the detailled listing – and the ◇-plot in Fig. 3 for $k = 1$). It means that each of such operon produces, using all its genes, all the enzymes that are needed to catalyse the corresponding reaction chain. We also found that 49.63% of the operons are fully covered by at least one 1-SIP. These rates drop to 12.59% and rise to 64.44%, respectively, when $k = 10$. As *k* becomes larger, the *k*-SIPs also become

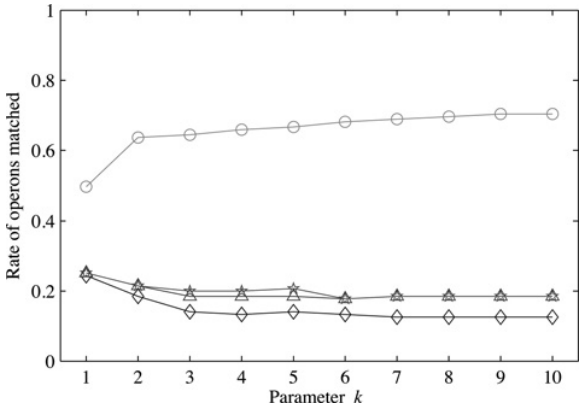


Fig. 3 *Operonic interest of all k-SIPs for distinct k in E. coli*

○-line represents the rate of operons that are fully covered by at least one *k*-SIP. The ◇-line (the Δ-line and the ★-line, respectively) resumes the rate of the single operons that are exactly matched by an *k*-SIP (rate of the single and the couple of operons that are matched exactly by a *k*-SIP and rate of the single, the couple and the triple of operons that are matched exactly by a *k*-SIP, respectively)

doi: 10.1049/iet-syb.2010.0070

larger, and thus tend to associate sets of genes that strictly contain operons rather than exactly match an operon. It is worth noticing here that 8.15% of the operons (11 of them) exactly match a k-SIP for all values of k from 1 to 10, mainly because of the fact that no alternative path exists to the 1-SIP and that the 10-SIP is thus identical to the 1-SIP. We then repeated the comparison between each k-SIP and each couple (and respectively triple) of operons. We found that 14 couples (2 triples, respectively) of operons match exactly one k-SIP each, for various $k \geq 1$ (see Table A.2 in the Supplementary material for details).

Operon prediction: After investigating the respective impacts of $\bar{w}$ and k, one can compare SIPPER with other techniques used to automatically predict operons. SIPPER prediction results are reported in Table 2. One must notice that even if the exact matching decreases with an increase of k, the prediction accuracy (i.e. true positive and negative values) of SIPPER is clearly improved. We compared our results with those obtained with one of the methods that predict metabolic operons. In particular, the method proposed in [33] has a sensitivity of 89% and a specificity of 87%. SIPPER, with $k = 10$, is almost as accurate. Note that in this previous approach, as in [6, 7], a plasticity parameter is introduced to skip a few reactions or genes in order to take into account the potential missing information. Unlike other approaches, SIPPER's plasticity is inherent to the method and relies on the k parameter. Moreover, SIPPER takes on an advantage that no other prediction method permits: it not only provides a description of a unique operon per k-SIP, but also descriptions of couples or triples of operons (see Table A.2 in the Supplementary material). Analysing their GO p-values confirms the functional interest of such operonic sets.

3.2.3 Impact of the gene neighbouring distance to predict operons:

Unlike other operon prediction methods, SIPPER is not a probabilistic-like approach. It consists on analysing, in a precise manner, what the impact of the biological property used to identify operons is. The gene neighbouring distance is well used, but operon structure lies in gene adjacencies. We put forward herein a criterion, called 'genomic density', which helps to discriminate paths that strictly use contiguous genes (not like $\bar{w}$ that combines both length of reaction chains and gene neighbouring distances). The genomic density of a given k-SIP is the ratio between the number of genes of the k-SIP and the number of genes of the smallest contiguous sequence of genes within the genome that fully covers the whole set of genes of the k-SIP. Fig. 4 illustrates the relationship between the genomic density of all 10-SIPs and their operon matching (i.e. by using the Jaccard's measure). The 10-SIPs are grouped by classes of genomic density. One has noticed an increase in the operon prediction accuracy when the genomic density increases. The class of a density between 0.7 and 0.9 indicates the 10-SIPs that most predict the operons. Notice again that for higher densities, the prediction ability of SIPPER significantly decreases (see Fig. A.2 in Supplementary material for further illustrations). This evolution shows that k-SIPs with a genomic density above 0.9 are either incomplete metabolic operons or exact metabolic operons with additional genes. These last are either functionally irrelevant or interesting, but not already referenced. The k-SIPs with a genomic density equal to 1 are strictly formed by contiguous genes. As these k-SIPs are not the best candidates for matching operons, it emphasises the fact that operons cannot be automatically predicted by focusing on the gene adjacency only. Note that such a result is biologically counter-intuitive. Searching contiguous genes is an important criterion, but not the optimal one when used alone for predicting metabolic operons. As discussed in the sequel, this observation confirms independent studies.

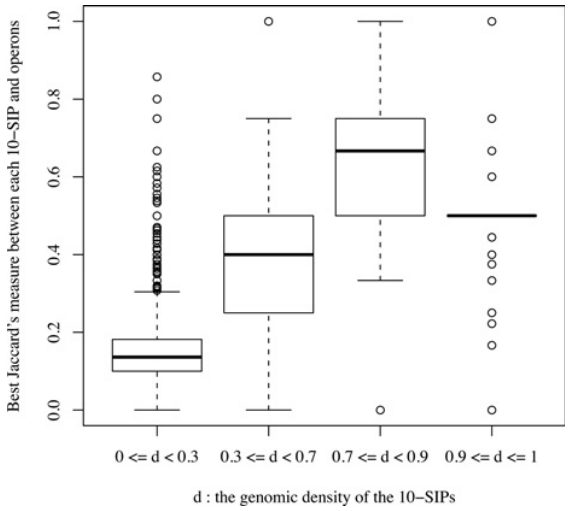


Fig. 4 *Rate of matching between 10-SIPs and operons in E. coli by using Jaccard's measure*

10-SIPs are grouped in genomic density classes (the X-axis) and box is plot to sum up the best Jaccard matching between each 10-SIP of each genomic density class and the operons (the Y-axis). In each genomic density class, the bold line is the median Jaccard's measure value of the corresponding 10-SIPs, the up and down limits of the box are the 0.25 and 0.75 quartile values, and the whiskers are the most extreme 10-SIPs which are no more than 1.5 times the height of the box. Plots outside of these extends are 10-SIPs that differ too much from those in the box

Table 2 Rates of operons identified by k-SIPs

	Exact matching, %	Sensitivity, %	Specificity, %
1-SIPs	24.44	61.93	77.26
10-SIPs	12.59	83.75	89.86

Exact matching measure is calculated using the Jaccard's measure. Results are also depicted using standard measures used for the automatic prediction of operons [32]: sensitivity is the rate of within operon gene pairs correctly predicted, specificity is the rate of transcriptional unit border gene pairs correctly predicted

3.3 Biological processes investigation

A similar analysis was performed in the metabolic context by comparing k-SIPs with functional modules, as described in the KEGG database. The comparison shows, for each k from 1 to 10, which proportion of functional modules are fully covered and exactly matched by the k-SIPs. We found that 33.33% of the modules (namely, 33 over 99 modules, see Table A.3 in Supplementary material) match exactly one 1-SIP each (see the ◇-plot in Fig. 3 for $k = 1$). We found that 51.52% of the modules are fully covered by at least one 1-SIP (○-plot in Fig. 5, with $k = 1$). These two rates, respectively, drop to 26.26% and rise to 60.60% when $k = 10$. Moreover, 18.18% of the KEGG modules (18 of them) exactly match a k-SIP even for all k from 1 to 10.

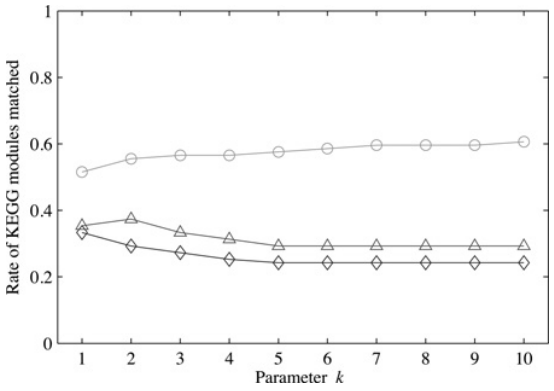


Fig. 5 *Modular interest of all k-SIPs for distinct k in E. coli*

○-line represents the rate of KEGG modules that are fully covered by at least one *k*-SIP. The ◇-line (the Δ-line respectively) resumes the rate of single (or couple of respectively) KEGG modules that are exactly matched by some *k*-SIPs

A closer look at these results (see Fig. A.3 in Supplementary materials for details) does not show a particular value of $\bar{w}$ that permits to match exactly the KEGG modules.

For a qualitative investigation, we then compared couples of modules and *k*-SIPs. We found (see Tables A.4 and in the Supplementary material for details) that 16 couples of modules are matched exactly by one *k*-SIP each, for $k \geq 1$. In Fig. 5, the Δ-line shows the increase in the rate of exactly matched KEGG modules, when *k*-SIPs that match couple of modules are considered additionally to those that match single module. No result was found for three or more modules.

4 Discussion

Applying SIPPER on *E. coli* with the gene neighbouring distance, in both the genomic and metabolic contexts (i.e. operons and KEGG modules), points out the existence, and often the biological significance, of metabolic paths directed by the genes localisation proximity (Section 3.1). These metabolic paths are *k*-SIPs found in the integrated network $\mathcal{G}_{\text{int}}$. A part of them revealed to be about 25% of the operons (see Section 3.2 and Fig. 3) and about 35% of the KEGG modules (see Section 3.3 and Fig. 5). The quality of these particular *k*-SIPs is confirmed by the shuffled tests made on genome and metabolism. The first test shows that *k*-SIPs that match operons are linked to the chosen gene neighbouring distance. This result is in accordance with the fact that changing the gene order within genome breaks operons. The second test shows that *k*-SIPs that match KEGG modules depend on the chaining of reactions, which is in agreement with the fact that KEGG modules are mostly composed of successive reactions. It is worth emphasising here that operons and KEGG modules are not a priori similar. This idea is supported by the fact that the *k*-SIPs that match operons, and those that match KEGG modules are very distinct (i.e. there is only one reaction chain in the metabolic network that exactly matches both an operon, fadBA, and a KEGG module, M00059).

Previous works [6, 7, 33] assume that the intra-operonic gene order has to rely on the role of encoded enzymes in the bacterial metabolism. Although they consider undirected metabolic networks and, thus, accept neighbouring relationship between reactions, which is not allowed in our approach, these studies show the great need for evidences to support that assumption. Kovács *et al.* [12] first propose a systematic study, by considering pairs of (not necessarily successive) genes within the same operon and asking whether their operonic order reflects the functional order of the encoded enzymes, as recorded by their participation to a common biochemical pathway. Such pairs are so-called 'colinear', and they fulfil constraints involving the order of genes and reactions only, and not the immediate succession of the genes along the genome or of the reactions along some reaction chain. The study shows that approximately 60% of the gene pairs in *E. coli* are colinear. Turning back to our approach – whose constraints involve both the order and the succession of genes and reactions – *k*-SIPs are topologically (and not biochemically) defined. However, they do match or include operons. Firstly, this fact shows the tendency of operonic genes to participate together in the same process in the metabolic network. In all, 24.4% of operons encode precisely the set of enzymes that are necessary to catalyse a reaction chain, whereas 25.2% of them (which gives a total of 49.6% of operons) are associated with other genes, not obligatorily close, to catalyse a reaction chain. Secondly, this combined genomic and metabolic proximity does not necessarily imply colinearity since the gene order within the genome may be entirely reversed (no colinear pair) or merged (some pairs, but not all pairs, are colinear) with respect to the order of reactions along the reaction chain. Indeed, for example, operons kbl.tdh, cyn.TSX, csiD.lhgO.gabDTP, otsBA, dgoRKADT are reversed, whereas the operon fadIJ is found both in right and reversed order, and operons rhaBAD, glgCAP, araBAD, rfbBDACX have their gene set merged along the reaction chain. However, some operons show a 'strong' (topological) colinearity, since all their genes appear exactly in the same order on the genome (see Table A.1 in the Supplementary material for details) and, via their encoded enzymes, along the reaction chain, even when the reaction chain is much longer than the operon. For instance, the fadBA operon contains two genes that encode enzymes catalysing a six reaction chain (fadB and fadA, respectively, encode for the first two reactions and the last four reactions). Similarly, operon pdhR.aceEF contains three genes that encode the enzymes catalysing a four reaction chain (aceE encodes the first two reaction enzymes, whereas aceF and lpd encode the last two reaction enzymes). Some other short operons (two genes) contribute to the catalysis of short reaction chains (two genes) thus being colinear pairs (as considered in [12]), but with the notable property of being made of successive genes corresponding to successive reactions (which is not required by a colinear pair and emphasises again our approach interest).

The genomic density of a *k*-SIP is an interesting measure since, for a small *k* (i.e. $k \leq 4$), a dense *k*-SIP emphasises an operon. Note that, from a quantitative viewpoint, although talA-tktB [34] and sucABCD [35] were not referenced in RegulonDB at the time of the analysis, they are highlighted by their gene density (density between 0.9 and 1), which argues the interest of SIPPER to discover potential metabolic operons. This observation remains true when considering alternative paths (*k* equals 10), reinforced by the accuracy improvement of the operon prediction. However, Section 3.2.3 shows that it is more difficult to discriminate the whole operon set using the genomic density only when alternative paths are added. The use of a gene neighbouring measure alone or with the genomic density as discriminant criteria does not seem self-sufficient

doi: 10.1049/iet-syb.2010.0070

to identify operon. This observation is confirmed by an independent study [36] that quantifies the operon prediction gain when other knowledge than gene arrangement are used.

Our approach provides a novel emphasis. *k*-SIPs identify the couples or triples of operons (Table A.2 in the Supplementary material for details). Some of them are already known in a regulatory context. For instance, fadBA, fadIJ share the dual common repressor ArcA, fadR [37]. Among couples of operons, cysDNC, cysJIH has already been identified as an über-operon [38]. Other operons have already been associated with a unique metabolism of interest, like atoDAEB and fadIJ that participate in the fatty acid degradation [39]. Some *k*-SIPs emphasise as well operons that share homologous genes (ascFB, bglGFB), which gives an insight into a reaction chain that is encoded in distinct locations on the genome, showing an abstraction of robustness as proposed by Kitano [40]. Notice here that those operons alone show a worse GO *p*-values than when they are associated [41].

Methods described in [6, 7, 33] authorise to skip a few reactions or genes to provide a flexible way to deal with missing knowledge. In particular, Boyer *et al.* [6] highlights the fact that operon identification is better when small skips are allowed. Conversely, the *k*-SIPs that match several modules or operons (see Sections 3.2.2 and 3.3) are a clear illustration of the interest of investigating imbricated functional modules as already highlighted by independent studies [42, 43].

In a similar way, some *k*-SIPs emphasise modules of interest. Some indicate functional units that are closely related to operons. For instance, M00037 represents an almost complete transcriptional unit that encodes for the histidine biosynthesis pathway. This *k*-SIP omits the hisL that controls the response in histidine availability [44], but remains accurate to describe a functional process that involves gene regulations. Other *k*-SIPs indicate couples of KEGG modules, like the ones that rely on the amino acid metabolism: M00033–M00034, M00033–M00035 or M00035–M00036. Each couple is transcriptionally regulated and represents a biological process that transforms chorismate into amino acids. The combination of these *k*-SIPs constitutes a set of reactions that have been called superpathways of chorismate [37]. Thus, these *k*-SIPs confirm herein their interest by showing a hierarchical description of the metabolism. As observed before for the couples of operons, the GO *p*-values confirm again this fact by showing higher *p*-values for modules alone compared to their associations (details in Table A.4 in Supplementary material). This supports the assumption of previous studies that decompose the metabolic network for further investigations [42, 43, 45].

5 Conclusion

We suggest herein a general framework that integrates gene neighbouring information from one or several data sources into a metabolic network. This integration is obtained using a generic notion of distance between genes, that is projected on the metabolic network via the encoded enzymes. The resulting integrated network, called $\mathcal{G}_{\text{int}}$, is analysed by computing the *k*-SIPs ($k \geq 1$), or *k*-SIPs, between two given reactions involved in this network, where the optimisation uses the generic distance between genes. The collection of *k*-SIPs obtained for a given *k* is then filtered according to an appropriate criterion to extract further information about a given genomic or metabolic context.

Our method allows us to observe interesting *k*-SIPs that are biologically relevant entities: *k*-SIPs that match either an operon, or a small group of operons, or a module, or a small group of modules. We also observe that alternative ways to transform metabolites is a corner stone feature in the search for functional entities, even when we consider intermediate non-neighbouring genes in the associated process. Further applications of our approach should involve alternative measures between genes and these are the main lines of our future works.

6 References

1 Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: 'Combinatorics of genome rearrangements' (MIT Press, 2009)

2 Zhang, B., Horvath, S.: 'A general framework for weighted gene co-expression network analysis', *Stat. Appl. Genet. Mol. Biol.*., 2005, **4**, article 17

3 Galperin, M.Y., Koonin, E.V.: 'Who's your neighbour? New computational approaches for functional genomics', *Nat. Biotechnol.*, 2000, **18**, pp. 609–613

4 Simeonidis, E., Rison, S.C.G., Thornton, J.M., Bogle, I.D.L., Papageorgiou, L.G.: 'Analysis of metabolic networks using a pathway distance metric through linear programming', *Metab. Eng.*, 2003, **5**, (3), pp. 211–219

5 Bordron, P., Eveillard, D., Rusu, I.: 'SIPPER: a flexible method to integrate heterogeneous data into a metabolic network'. 2011 IEEE First Int. Conf. on Computational Advances in Bio and Medical Sciences (ICCABS), 2011, pp. 40–45

6 Boyer, F., Morgat, A., Labarre, L., Pothier, J., Viari, A.: 'Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data', *Bioinformatics*, 2005, **21**, (23), pp. 4209–4215

7 Ogata, H., Fujibuchi, W., Goto, S., Kanehisa, M.: 'A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters', *Nucleic Acids Res.*, 2000, **28**, pp. 4021–4028

8 Williams, E.J.B., Bowles, D.J.: 'Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*', *Genome Res.*, 2004, **14**, pp. 1060–1067

9 Ihmels, J., Bergmann, S., Barkai, N.: 'Defining transcription modules using large-scale gene expression data', *Bioinformatics*, 2004, **20**, (13), pp. 1993–2003

10 Ihmels, J., Levy, R., Barkai, N.: 'Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*', *Nat. Biotechnol.*, 2004, **22**, (1), pp. 86–92

11 Wei, H., Persson, S., Mehta, T., *et al.*: 'Transcriptional coordination of the metabolic network in *Arabidopsis thaliana*', *Plant Physiol.*, 2006, **18**, pp. 762–774

12 Kovács, K., Hurst, L.D., Papp, B.: 'Stochasticity in protein levels drives colinearity of gene order in metabolic operons of *Escherichia coli*', *Plos Biol.*, 2009, **7**, (5), p. e1000115

13 Rocha, E.P.C.: 'The organization of the bacterial genome', *Ann. Rev. Genetics*, 2008, **42**, pp. 211–233

14 Chua, H.N., Sung, W.K., Wong, L.: 'An efficient strategy for extensive integration of diverse biological data for protein function prediction', *Bioinformatics*, 2007, **23**, (24), pp. 3364–3373

15 Enright, A.J., Ilipoulos, I., Kyrpides, N.C., Ouzounis, C.A.: 'Protein interaction maps for complete genomes based on gene fusion events', *Nature*, 1999, **402**, pp. 86–90

16 Gelfand, M.S., Koonin, E.V., Mironov, A.A.: 'Prediction of transcription regulatory sites in Archaea by a comparative-genomic approach', *Nucleic Acids Res.*, 2000, **28**, pp. 695–705

17 Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D.: 'A combined algorithm for genome-wide prediction of protein function', *Nature*, 1999, **402**, pp. 83–86

18 Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D.: 'Detecting protein function and protein-protein interactions from genome sequences', *Science*, 1999, **285**, pp. 751–753

19 Mironov, A.A., Koonin, E.V., Roytberg, M.A., Gelfand, M.S.: 'Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes', *Nucleic Acids Res.*, 1999, **27**, pp. 2981–2989

20 Overbeek, R., Fonstein, M., DSouza, M., Pusch, G.D., Maltsev, N.: 'The use of gene clusters to infer functional coupling', *Proc. Natl. Acad. Sci.*, 1999, **96**, pp. 2896–2901

21 Snel, B., Bork, P., Huynen, M.A.: 'The identification of functional modules from the genomic association of genes', *Proc. Natl. Acad. Sci.*, 2002, **99**, (9), pp. 5890–5895

22 Yamanishi, Y., Vert, J.P., Kanehisa, M.: 'Supervised enzyme network inference from the integration of genomic data and chemical information', *Bioinformatics*, 2005, **21**, (Suppl 1), pp. i468–i477

23 Blattner, F.R., Plunkett, G., Bloch, C.A., *et al.*: 'The complete genome sequence of *Escherichia coli K-12*', *Science*, 1997, **277**, (5331), pp. 1453–1462

24 Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A.: 'The KEGG databases at GenomeNet', *Nucleic Acids Res.*, 2002, **30**, (1), pp. 42–46

25 Croes, D., Couche, F., Wodak, S.J., van Helden, J.: 'Metabolic PathFinding: inferring relevant pathways in biochemical networks', *Nucleic Acids Res.*, 2005, **33**, (Web server issue), pp. W326–W330

26 Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., Eisenberg, D.: 'Prolinks: a database of protein functional linkages derived from coevolution', *Genome Biol.*, 2004, **5**, (5), p. R35

27 Yen, J.Y.: 'Finding the *k* shortest loopless paths in a network', *Manag. Sci.*, 1970, **17**, pp. 712–716

28 Jacob, F., Perrin, D., Sanchez, C., Monod, J.: 'L'opéron : groupe de gènes à expression coordonnée par un opérateur', *C.R. Hebd. Acad. Sci. Paris*, 1960, séance 250, pp. 1727–1729

29 Salgado, H., Gama-Castro, S., Peralta-Gil, M., *et al.*: 'RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions', *Nucleic Acids Res.*, 2006, **34**, (Database issue), pp. D394–D397

30 Kanehisa, M., Araki, M., Goto, S., *et al.*: 'KEGG for linking genomes to life and the environment', *Nucleic Acids Res.*, 2008, **36**, (Database issue), pp. D480–D484

31 Maslov, S., Sneppen, K.: 'Specificity and stability in topology of protein networks', *Science*, 2002, **296**, (5569), pp. 910–913

32 Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., Collado-Vides, J.: 'Operons in *Escherichia coli*: genomic analyses and predictions', *Proc. Natl. Acad. Sci. USA*, 2000, **97**, (12), pp. 6652–6657

33 Zheng, Y., Szustakowski, J.D., Fortnow, L., Roberts, R.J., Kasif, S.: 'Computational identification of operons in microbial genomes', *Genome Res.*, 2002, **12**, (8), pp. 1221–1230

34 Lacour, S., Landini, P.: 'σS-dependent gene expression at the onset of stationary phase in *Escherichia coli*: function of σS-dependent genes and identification of their promoter sequences', *J. Bacteriol.*, 2004, **186**, (21), p. 7186

35 Buck, D., Spencer, M.E., Guest, J.R.: 'Cloning and expression of the succinyl-CoA synthetase genes of *Escherichia coli* K12', *J. Gen. Microbiol.*, 1986, **132**, (6), pp. 1753–1762

36 Chuang, L.Y., Tsai, J.H., Yang, C.H.: 'Binary particle swarm optimization for operon prediction', *Nucleic Acids Res.*, 2010, **38**, (12), p. e128

37 Keseler, I.M., Bonavides-Martnez, C., Collado-Vides, J., *et al.*: 'EcoCyc: a comprehensive view of *Escherichia coli* biology', *Nucleic Acids Res.*, 2009, **37**, (Database issue), pp. D464–D470

38 Che, D., Li, G., Mao, F., Wu, H., Xu, Y.: 'Detecting über-operons in prokaryotic genomes', *Nucleic Acids Res.*, 2006, **34**, (8), pp. 2418–2427

39 Jenkins, L.S., Nunn, W.D.: 'Genetic and molecular characterization of the genes involved in short-chain fatty acid degradation in *Escherichia coli*: the ato system', *J. Bacteriol.*, 1987, **169**, (1), pp. 42–52

40 Kitano, H.: 'Biological robustness', *Nat. Rev. Genetics*, 2004, **5**, (11), pp. 826–837

41 Boyle, E.I., Weng, S., Gollub, J., *et al.*: 'GO: TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes', *Bioinformatics*, 2004, **20**, (18), pp. 3710–3715

42 Gagneur, J., Jackson, D.B., Casari, G.: 'Hierarchical analysis of dependency in metabolic networks', *Bioinformatics*, 2003, **19**, (8), pp. 1027–1034

43 Ma, H.W., Zhao, X.M., Yuan, Y.J., Zeng, A.P.: 'Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph', *Bioinformatics*, 2004, **20**, (12), pp. 1870–1876

44 Alifano, P., Carlomagno, M.S., Bruni, C.B.: 'Location of the *hisGDCBHAFI* operon on the physical map of *Escherichia coli*', *J. Bacteriol.*, 1992, **174**, (11), pp. 3830–3831

45 Schilling, C.H., Schuster, S., Palsson, B.O., Heinrich, R.: 'Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era', *Biotechnol. Prog.*, 1999, **15**, (3), pp. 296–303

doi: 10.1049/iet-syb.2010.0070

3.3 Building a functional metabolic network of a microbial ecosystem

The study of the microbial ecosystems followed similar aspirations than those in cellular biology. Complementary to extensive microbial community descriptions, ecological sciences seek to understand the functional features of a given microbial ecosystem. To study the microbial ecosystem using genomic techniques, one proposed a protocol similar to what we proposed for investigating single organisms. A first step consists of generating sequence data for all bacterial components that constitute the ecosystem. Then, one aims to extract the whole set of chemical reactions involved within the ecosystems. One or several microbial species will encode these reactions. As previously shown, the combination of these reactions designs a metabolic network. Contrary to cellular systems analysis, the complexity herein consists in considering that (i) similar genes could be carried out by several microbial organisms - emphasizing a putative functional redundancy - sometimes called core genes or reactions. Also, (ii) reversely part of metabolic pathways are specific, which indicates a putative interplay between the microbial strains for a full metabolic network to take place. Ideally, each component of the ecosystem is sequenced separately. This task is difficult. One achieves this by the isolation of bacterial cultures before nucleic acid extraction, which allows a clear functional separation of different components of the ecosystem. For an application to non-cultivable microbes, a parallel meta-genome or meta-transcriptome approach is advisable. Meta-genome or meta-transcriptome reads are mapped against the isolated bacterial genomes to remove known components. Worth the notice, new computational techniques proposed to delineate genomes from metagenomes. The techniques are very similar to those used for building co-occurrence networks and take the benefit of graphical LASSO algorithms. When genes co-occur in different samples, these techniques assume these genes are belonging to the same Metagenome species (MGS) [155]. Complementary, one could perform an extension of the MGS identification by binning all the genes and full-filling their assembly. The result, called metagenome-assembled genomes (MAGS), is a reasonable approximation of a genome that depicts a non-model organism (i.e., mostly uncultivable organism) [48].

Once identified and sequenced individually (or extracted from MAGS), the different components of the microbial community are resumed by their metabolic networks. For this purpose, one selects the genes that encode for enzymes that permit a biochemical reaction. The manual generation of these networks is too labor-intensive to be applied to assemblages of several organisms and requires dedicated tools [219]. One commonly used tool for this purpose is MetaPatthway [118]. In-

spired from Pathway Tools [106], this technique proposes network reconstructions based on functional annotations. It can also easily be combined with increasingly powerful automatic annotation platforms such as MaGE [204] or RAST [147], thus generating first draft metabolic networks for the different components of the ecosystem. These networks are then merged to constitute a draft network of the whole ecosystem, keeping track of which components contributed which reactions. A significant benefit of this procedure is that it bridges the gap between studies that focus on a single species and studies that represent an entire ecosystem as a unique metabolic network without functional distinction between different components [161]. Worth to notice, in [123] proposed a complementary technique to build a metabolic network for an ecosystem that consists in, both, maximizing the completion of single species network (i.e., intraspecies network or subnetwork), and minimizing the use of reactions between species (i.e., interspecies network).

The resulting network constitutes an approximation of the right ecosystem metabolic network designed to decipher a given physiological question. Following a schema similar to above, this metabolic knowledge could be then integrated with genomic knowledge when one considers the genomes of all microbial species that interplay within the ecosystem. The technique CANOE assumes this hypothesis to identify the functions of unknown prokaryotic genes [186]. In a similar trend, the following study proposes another extension to identify SGS (called metabolon in [186]) that defines the putative interactions of functional units that promote metabolic pathways at the ecosystem level.

Philippe Bordron, Mauricio Latorre, Maria Paz Cortés, Mauricio González, Sven Thiele, Anne Siegel, Alejandro Maass, and Damien Eveillard. Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen*, 5(1):106–117, February 2016

MicrobiologyOpen

Open Access

ORIGINAL RESEARCH

Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach

Philippe Bordron[1,2], Mauricio Latorre[1,2,3], Maria-Paz Cortés[1,2], Mauricio González[2,3], Sven Thiele[4], Anne Siegel[5,6], Alejandro Maass[1,2,7] & Damien Eveillard[8]

[1]Mathomics, Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile
[2]Center for Genome Regulation (Fondap 15090007), Universidad de Chile, Santiago, Chile
[3]Laboratorio de Bioinformática y Expresión Génica, INTA, Universidad de Chile, Santiago, Chile
[4]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[5]IRISA, UMR 6074, CNRS, Rennes, France
[6]INRIA, Dyliss Team, Centre Rennes-Bretagne-Atlantique, Rennes, France
[7]Department of Mathematical Engineering, Universidad de Chile, Santiago, Chile
[8]LINA, UMR CNRS 6241, EMN, Université de Nantes, Nantes, France

Keywords
Environmental microbiology, in silico analysis, metabolic pathways, molecular microbial ecology

Correspondence
Philippe Bordron, Mathomics, Center for Mathematical Modeling, Universidad de Chile, Beauchef 851, Piso 6, Of. 613, Santiago, Chile. Tel: +56 2 2978 4551; E-mail: phbordron@dim.uchile.cl

Funding Information
This work was supported by grants Fondap 15090007, Basal program PFB-03 CMM, IntegrativeBioChile INRIA Assoc. Team, ANR-10-BLANC-0218 and CIRIC-INRIA Chile (line Natural Resources).

Received: 16 June 2015; Revised: 12 October 2015; Accepted: 19 October 2015

doi: 10.1002/mbo3.315

Abstract

Following the trend of studies that investigate microbial ecosystems using different metagenomic techniques, we propose a new integrative systems ecology approach that aims to decipher functional roles within a consortium through the integration of genomic and metabolic knowledge at genome scale. For the sake of application, using public genomes of five bacterial strains involved in copper bioleaching: *Acidiphilium cryptum*, *Acidithiobacillus ferrooxidans*, *Acidithiobacillus thiooxidans*, *Leptospirillum ferriphilum*, and *Sulfobacillus thermosulfidooxidans*, we first reconstructed a global metabolic network. Next, using a parsimony assumption, we deciphered sets of genes, called *Sets from Genome Segments* (SGS), that (1) are close on their respective genomes, (2) take an active part in metabolic pathways and (3) whose associated metabolic reactions are also closely connected within metabolic networks. Overall, this SGS paradigm depicts genomic functional units that emphasize respective roles of bacterial strains to catalyze metabolic pathways and environmental processes. Our analysis suggested that only few functional metabolic genes are horizontally transferred within the consortium and that no single bacterial strain can accomplish by itself the whole copper bioleaching. The use of SGS pinpoints a functional compartmentalization among the investigated species and exhibits putative bacterial interactions necessary for promoting these pathways.

Introduction

The ecosystems behavior, as observed today by experiments, is the immediate result of microbial interactions between several organisms. By themselves, these interactions explain several steps of natural biogeochemical cycles as well as biological processes of economical interest. Interestingly, recent high-throughput genomic data have shown for different ecological systems a microbial diversity that was far greater than expected (Roesch et al. 2007; Hingamp et al. 2013; de Vargas et al. 2015) and promise to improve the understanding of microbial ecosystems behaviors from the molecular viewpoint. To this aim, recent studies advocate for the use of metagenomic techniques to intensively investigate different microbial ecosystems (DeLong 2005; Karsenti et al. 2011; Ranjard et al. 2013).

Advances in bioinformatics have also improved the analysis of next-generation sequencing data that characterize microbial communities, addressing the question "*who is there and who is not*" (Raes et al. 2011). However, this description remains insufficient to depict functional behaviors of microbial ecosystems if no other complementary knowledge is considered. Recent studies overcame this weakness by combining all biotechnological resources available within a modeling framework. In particular, one must notice the success of techniques that target potential cross-feedings within microbial consortium. Without being exhaustive, they focus on either solely community metabolic network (via graph or constraint based approaches) (Borenstein et al. 2008; Zengler and Palsson 2012; Zomorrodi et al. 2014), or co-occurrence graph techniques that interconnect covariation of microbial abundance (Faust and Raes 2012; Faust et al. 2012) and environmental features (Ruan et al. 2006, 2010; Brown et al. 2009; Chaffron et al. 2010; Patel et al. 2010). Some other techniques advocate for hybrid approaches that combine heterogeneous knowledge. Microbial cross-feedings are for instance investigated by integrating phylogenetic and environmental knowledge (see Zaneveld et al. 2011 for review); omics experiments and in situ observations (Orphan 2009; Zelezniak et al. 2015); or metabolic networks and diversity graphs (Tzamali et al. 2011). These recent modeling approaches are all complementary and reinforce the emergence of the new subdiscipline called systems ecology (Klitgord and Segrè 2011) that aims to tackle complex ecological questions by merging heterogeneous data with new computational techniques. However, applications of these techniques remain difficult when communities are experimentally out of reach, which is particularly the case for studying cross-feedings between extremophile species.

This study overcomes this issue by proposing a modeling framework for metagenomic consortia analysis that not only considers the presence/absence of several bacterial species, but rather their precise genome composition and corresponding metabolic networks. Complementary to previous integrative methods (Segata et al. 2013; Zelezniak et al. 2015), our framework emphasizes putative functional species interactions in the metagenomic consortium through the integration of genome-wide genomic and metabolic knowledge. For its application, we considered a chemoautotrophic microbial community related to the copper bioleaching process, one of the most extensive and complex biohydrometallurgic processes in which a series of chemical and biological reactions facilitate the oxidation of insoluble sulfide ores, releasing soluble metals such as copper. Beside their economic interest, different theories postulate that these mineral complexes were the principal source of energy used by ancient bacteria consortia (Chemoautotrophic Iron-Sulfur World theory) (Drobner et al. 1990; Wächtershäuser 2000). This particular mining microbial ecosystem is characterized by high concentrations of metals and elevated acidity which advocates for the use of these living microbial species to study adaptation under extreme conditions. Furthermore, the community exhibits a remarkably simple habitat for understanding how ancient microbial communities work (Baker and Banfield 2003). Finally, this microbial ecosystem could be resumed by a limited number of bacterial strains (Yin et al. 2008). This reductionist advantage is of particular interest to benchmark bioinformatics methods on a simple community system while maintaining realistic ecological features.

In terms of current knowledge, several bacteria participating in bioleaching of copper have been isolated and sequenced (Barreto et al. 2003; Mi et al. 2011; Valdés et al. 2011; Travisany et al. 2012). Even though the study of these strains significantly improved the bioleaching knowledge, studying single organisms did not allow to understand bioleaching as a whole. For instance, little is known about the relative functional importance of each bacterium in the bioleaching.

To decipher the respective genome-wide role of each bacterium within this mining ecosystem, this work integrates, at community scale, genomic, and metabolic knowledge. Genomic and metabolic features integration relies on a parsimonious assumption that considers the different omics knowledge connected linked by intrinsic and direct properties connections, as adapted from a previous single-cell systems biology studies (Boyer et al. 2005; Bordron et al. 2011). In practice, our approach connects genomic and metabolic knowledge by considering the genome organization and the biochemical reactions catalyzed by enzymes encoded by its genes. The underline parsimonious principle assumes that genes that must be jointly regulated to activate a metabolic reaction cascade, herein a bioleaching pathway, and should be close enough in the genome organization (for illustration dashed lines in Fig. 1). The corresponding sets of genes satisfying the above mentioned constraints are defined by *Sets from Genome Segments* (SGS) (respectively, pink and blue segments in bacterium 1, and the red segment in bacterium 2 in Fig. 1). SGS represents a segment of consecutive genes in a bacterial genome with a maximum number of genes that participates in a given metabolic pathway. Through this selection, SGS decipher putative sets of genes that (1) take an active part in metabolic pathways while being closely connected via metabolic networks and (2) are consecutive on each of the genomes involved.

We applied this approach to a reduced but exhaustive bacterial community composed of five different copper biomining microbial strains which are simultaneously growing in an industrial bioreactor, fully operational for large-scale cultures (CODELCO, Radomiro Tomic Division, Antofagasta, Chile): *Acidithiobacillus ferrooxidans ATCC 23270*, *Acidiphilium cryptum JF-5*, *Acidithiobacillus*

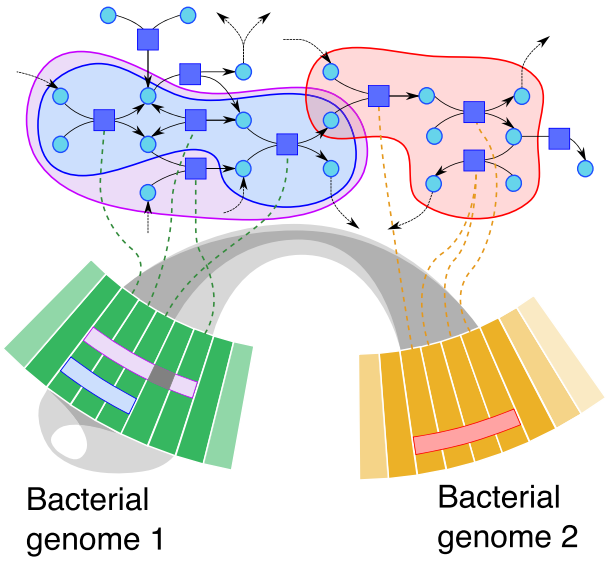


Figure 1. Illustration of *sets from genome segments* (SGS) when applied on a toy microbial community. The upper part of the figure illustrates a metabolic network, where circles are metabolic compounds, and squares are reactions that happen between them. The two bands at the lower part represent parts of two bacterial genomes as sequences of genes. The catalytic function of a gene, via its enzyme, is represented by a dashed line between the gene and the reactions it catalyzes. A SGS appears on the genome as set of genes contained into a segment. The segment containing a SGS can contain genes that do not participate into the SGS (e.g., without catalytic function). The projection of a SGS on the metabolic network defines a set of reactions. Two SGS are linked together by a gray ribbon if they can be chained through the metabolic network.

thiooxidans ATCC 19377, *Leptospirillum ferriphilum ML-04,* and *Sulfobacillus thermosulfidooxidans DSM 9293* strains. The selected bacterial strains; beside being the main five mining genomes (sequenced, assembled, and annotated) available today; also cover the four dominant bacterial taxonomic groups present in copper mines, that represents more than the 50% of the total bacterial genera present in mine ecosystems (Yin et al. 2008).

After an exhaustive description of the SGS method and omics data used for its application (e.g., genomes and metabolic networks), this study provides an integrative view of bioleaching at both metabolic pathway and microbial genome levels. In particular, SGS will be enumerated and their relative distribution on all the five genomes of bacteria further detailed. At the metabolic level, SGS modeling paradigm pinpoints evident complementarities of bacterial strains to promote selected bioleaching pathways. In addition, beyond the simple mapping of bacterial strain metabolism on these pathways of interest, SGS depict functional units, similar to operons, that must be combined to genetically monitor the whole bacterial participation in the bioleaching process, as well as specific transporters that should be further investigated to understand putative metabolic collaborative processes at the community level.

Materials and Methods

Genomes

Five distinct genomes from bacteria involved in copper bioleaching were considered: *A. cryptum* JF-5 (NCBI - plasmid & chromosome GenBank ID: from CP000689 to CP000697) (Magnuson et al. 2010), *At. ferrooxidans* ATCC 23270 (NCBI - GenBank ID: CP001219) (Valdes, 2008), *At. thiooxidans* ATCC 19377 (NCBI - GenBank ID: AFOH00000000) (Valdés et al. 2011), *L. ferriphilum* ML-04 (NCBI - GenBank ID: CP002919) (Mi et al. 2011), and *Sb. thermosulfidooxidans* DSM 9293 (JGI database - Project ID: 97948). The three annotated genomes of *A. cryptum*, *At. Ferrooxidans,* and *L. ferriphilum* are, respectively, composed of 3691 (3158 genes from main chromosome and 533 from plasmids), 3304, and 2527 genes, whereas 2884 and 3602 potential genes were predicted, respectively, for the two nonannotated genomes *At. thiooxidans* and *Sb. thermosulfidooxidans*.

Metabolic networks

Metabolic networks of *A. cryptum*, *At. Ferrooxidans,* and *L. ferriphilum* were downloaded from Metacyc database v17.5 (Caspi et al. 2012). *Sb. thermosulfidooxidans* and *At. thiooxidans* metabolic networks were reconstruct following a standard procedure: both genome sequences were annotated using a local GenDB platform (Meyer et al. 2003). Gene predictions were made using Glimmer and functional gene annotations were processed by using state-of-the-art methods such as SignalP and TMHMM, and by performing local BLAST searches against databases NCBI nr, Swissprot, Omniome, PDB, KEGG, COG, and TCDB. After gene annotation, a GenBank format file was constructed for each genome and used as a general input for metabolic reconstruction using Pathway Tools software v16.0 (Karp et al. 2010). To reconstruct a metabolic network, we considered a metabolic reaction present when a gene encoding for an enzyme associated to this reaction was identified within the genome. Reactions have then been connected together if a product of a reaction was the substrate of another. To build a metabolic network of the microbial community, the union of the five different metabolic networks was considered.

Complementary, and for the sake of illustration, a list of 13 pathways considered as related to the copper bioleaching process was emphasize (Quatrini et al. 2009): (1) Fe(II) oxidation (Metacyc ID: PWY-6692); (2) heme biosynthesis: (a) Superpathway of heme biosynthesis from uroporphyrinogen-III (Metacyc ID: PWY0-1415), (b) heme

biosynthesis from uroporphyrinogen-III I (Metacyc ID: HEME-BIOSYNTHESIS-II), c) heme biosynthesis from uroporphyrinogen-III II (Metacyc ID: HEMESYN2-PWY), d) Superpathway of heme biosynthesis from glutamate (Metacyc ID: PWY-5918), and e) Superpathway of heme biosynthesis from glycine (Metacyc ID: PWY-5920); (3) iron-oxidizing/O2-reducing supercomplex (Metacyc ID: CPLX-8218); (4) NAD biosynthesis: a) NAD biosynthesis I (from aspartate) (Metacyc ID: PYRIDNUCSYN-PWY), b) NAD biosynthesis II (from tryptophan) (Metacyc ID: NADSYN-PWY), and c) NAD biosynthesis III (Metacyc ID: NAD-BIOSYNTHESIS-III); (5) Superpathway of sulfate assimilation and cysteine biosynthesis (Metacyc ID: SULFATE-CYS-PWY); (6) [2Fe-2S] iron-sulfur cluster biosynthesis ([Fe-S] cluster biosynthesis) (Metacyc ID: PWY-7250); (7) glutathione biosynthesis (Metacyc ID: GLUTATHIONESYN-PWY).

Sets from Genome Segments modeling framework and associated methods

The metabolism of a given bacterial system is defined by a set of compounds and the related metabolic reactions. A metabolic network was represented as a directed graph, also called metabolic graph, in which nodes represent reactions and an edge between two reactions exists when a product of the input reaction is a substrate of the targeted one. Notice herein that for the sake of modeling and following recommendations in Croes et al. (2005), we have deliberately neglected common and highly connected compounds that are present in the environment (i.e., compounds like water, ATP, NADP, etc., that are present in more than 40 reactions in this paper) and cofactors listed in Christian et al. (2009) to avoid artificial interconnections between reactions (see Ravasz et al. 2002; Guimera and Amaral 2005 for similar assumptions). Complementary, each bacterial chromosome were represented as an ordered list of genes, called *sequence*. Some of these genes encode for enzymes known to catalyze metabolic reactions. We called them metabolic genes. Notice that a given reaction may be catalyzed by several enzymes, implying that a reaction can be associated to one or several metabolic genes.

Assuming a direct causal link between a metabolic gene and a reaction via its encoded enzyme, we designed the SGS paradigm to integrate both knowledge. A SGS is a set of *metabolic genes*, contained in a segment (i.e. a *subsequence*) of the genome, that connects a predetermined initial reaction (and related gene) to a predetermined ending reaction (and related gene) within the metabolic network. Among the whole combination of SGS, our technique seeks to decipher SGS that satisfy the following parsimonious properties (see appendix S1 for a formal definitions and S2 for a parameter sensitivity analysis):

1. For all five bacterial chromosomes, the search was restricted to SGS with at most 20 consecutive metabolic genes (no more SGS were found for more than 20 genes). All pairs of initial and end reactions were considered.
2. From the whole set of SGS, those with a genomic density greater than or equal to 0.6 were selected (no more than ~⅓ of genes are missing): the genomic density is the ratio of the number of genes involved in the metabolic pathway covered by the SGS by the total number of genes within the SGS (see appendix S1 for formal definition). For example, in Figure 1, the purple SGS has a genomic density equal to 0.8, whereas blue and red SGS have a genomic density of 1. When a SGS has a genomic density equal to 1, all genes in the SGS are involved in the targeted metabolic pathway. Conversely, a genomic density close to zero implies that few genes in the SGS are involved in the metabolic pathways.
3. To avoid a "Russian doll effect", among dense SGS, we selected the *dominant* ones, that are, those that are not included into larger SGS.

Transcriptomic enrichment

In order to support the SGS prediction, we compared our results to a set of microarray experiments performed on cultures of *At. ferrooxidans* strain Wenelen (DSM 16786 ; Latorre et al. 2015). Wenelen was grown in 62 mmol/L $FeSO_4$-$7H_2O$ containing modified 9 K medium ($(NH_4)_2SO_4$: 0.4 g/L; K_2HPO_4: 0.4 g/L; $MgSO_4$-$7H_2O$: 0.4 g/L) adjusted to pH 1.8 with concentrated sulfuric acid in batch conditions until early exponential phase of the culture (24 h at 30°C). Afterward, four minerals (quartz concentrate (20% p/v), sample of pyrite concentrate (10% p/v), chalcopyrite concentrate (5% p/v) and elemental sulfur powder (5% p/v)) were added to the bacterial cultures (except to the control), which were grown at 30°C without shaking but with forced air supply. After 16 h of grown the cells were collected to RNA extraction, cDNA synthesis and microarrays hybridization. From a transcriptomic viewpoint, the expression for common genes between ATCC23270 and Wenelen strains is identical (Levican et al. 2008; see ortholog list between both strains in File S2).

Results

Overlap of metabolic reactions within the community

Acidiphilium cryptum, *Acidithiobacillus ferrooxidans*, *Acidithiobacillus thiooxidans*, *Leptospirillum ferriphilum*,

and *Sulfobacillus thermosulfidooxidans* have, respectively, 176, 75, 61, 131, and 263 specific metabolic reactions (from 1470, 1190, 1182, 1194, and 1418 total reactions). Merging the five metabolic networks generates a network composed of 2311 reactions, where 30% of them (706 reactions) are exclusive to single strains (Fig. 2A). Conversely, around 70% (1605 reactions) are conserved in the five genomes, describing a core of common pathways (Fig. 2B). Conserved reactions are related to the generation of precursor metabolites, energy, and basal metabolism. Specific reactions are mostly related to degradation of complex sugars, organic acids and subproducts of protein synthesis, indicating that major metabolic specificities are related to secondary metabolism processes. Noteworthy, these metabolic specificities distinguish two major groups: (1) reactions associated with energy source requirements (mainly iron oxidation) present in autotrophs (*At. ferrooxidans* and *L. ferriphilium*) and chemo-heterotroph (*Sb. thermosulfidooxidans*), and (2) reactions related to organic degradation compounds of organoheterotrophs (*A. cryptum*) corroborating previous descriptions of specific proteins involved in bioleaching (Baker and Banfield 2003).

SGS are highly specific to each bacterial strain and represent operons

Overall, SGS involve 707 distinct reactions (30.6 % of the meta-metabolism). Figure 2A and C show that among them, 362 distinct reactions (51%) are *single strain SGS specific*. This proportion is surprisingly large considering that only 105 of 707 reactions (~15%) are *monospecific* (Fig. 2A and B). In addition, when considered as sets of reactions, most of SGS are *single strain specific* (183 or ~60 % – see third column in Table 1) and almost unique in their strain (4–11% of SGS from a strain share the same set of reactions with another SGS of the same strain).

In parallel, sets of genes emphasized by SGS were compared to predicted operons (Pathway Tools (Karp

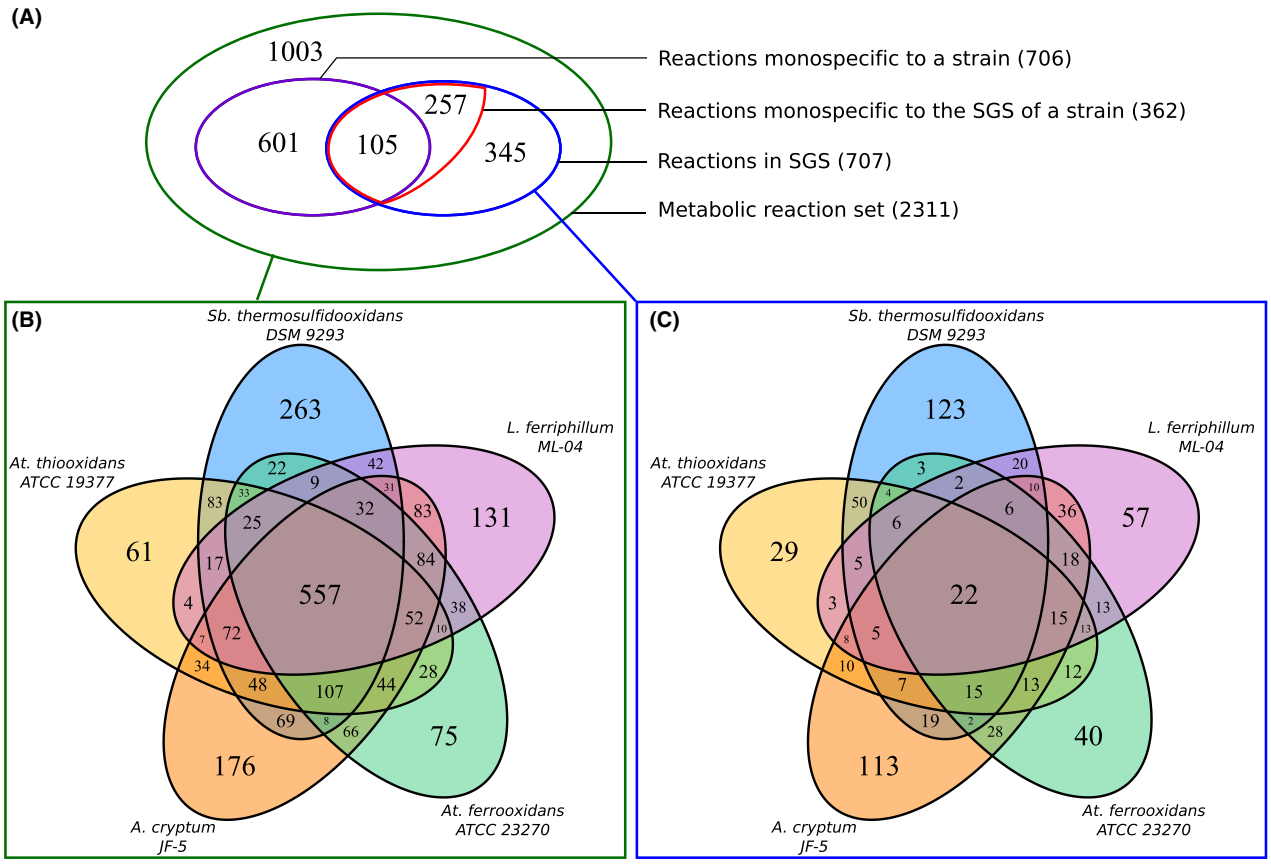


Figure 2. Reactions distribution of the five biomining bacteria. Diagram (A) illustrates how the set of reactions composing the meta-metabolism is monospecific or multispecific, and also which part of a reaction is involved in *sets from genome segments* (SGS). The Venn diagram (B) illustrates how the set of reactions composing the meta-metabolism is distributed among bacteria. The Venn diagram (C) illustrates how the set of reactions involved into the SGS is distributed among the bacteria.

Table 1. Number of *sets from genome segments* (SGS) and the related number of sets of reactions obtained when each SGS is projected onto the metabolic network of the corresponding bacterial strain and the microbial consortium meta-metabolism. Due to the existence of common reactions to several organisms, the number of reactions within SGS of the consortium is not the sum of the number of reactions of each strain. The comparison of reaction sets was done only by considering the frontier of the SGS: two reaction sets are considered as similar if they have at least one start reaction and one end reaction from their SGS in common. The specific ones are those from distinct organisms that are not similar.

Bacteria	Number of distinct SGS	Number of distinct reaction sets	Number of specific reaction sets according to SGS boundaries
A. cryptum JF-5	98	92	45
At. ferrooxidans ATCC 23270	61	59	22
At. thiooxidans ATCC 19377	67	61	28
L. ferriphilum ML-04	78	72	38
Sb. thermosulfidooxidans DSM 9293	92	83	50
Community	396	308	183

A. cryptum, Acidiphilium cryptum; At. ferrooxidans, Acidithiobacillus ferrooxidans; At. thiooxidans, Acidithiobacillus thiooxidans; L. ferriphilum; Leptosprillum ferriphilum; Sb. thermosulfidooxidans, Sulfobacillus thermosulfidooxidans.

et al. 2010)) or known operons as stored in ProOpDB (Taboada et al. 2012) and DOOR2 (Mao et al. 2009) databases for *A. cryptum* and *At. ferrooxidans*. A set of SGS gene was considered as an operon when its Jaccard measure is greater than or equal to 0.6, which represents 65.4% of all sets of SGS genes (highlighted as pink segments in Fig. 4). Complementary, and for the sake of support the SGS prediction, SGS genes expression was analyzed. These genes are significantly differentially expressed for others set of genes located in the vicinity; which confirms the SGS functional interest in accordance to available experimental stresses (see File S3 and Appendix S3).

SGS are complementary to promote metabolic pathways and show putative cooperations

Figure 3 and Figure S3 show the projection of SGS on metabolic pathways of interest (resp. on Superpathway

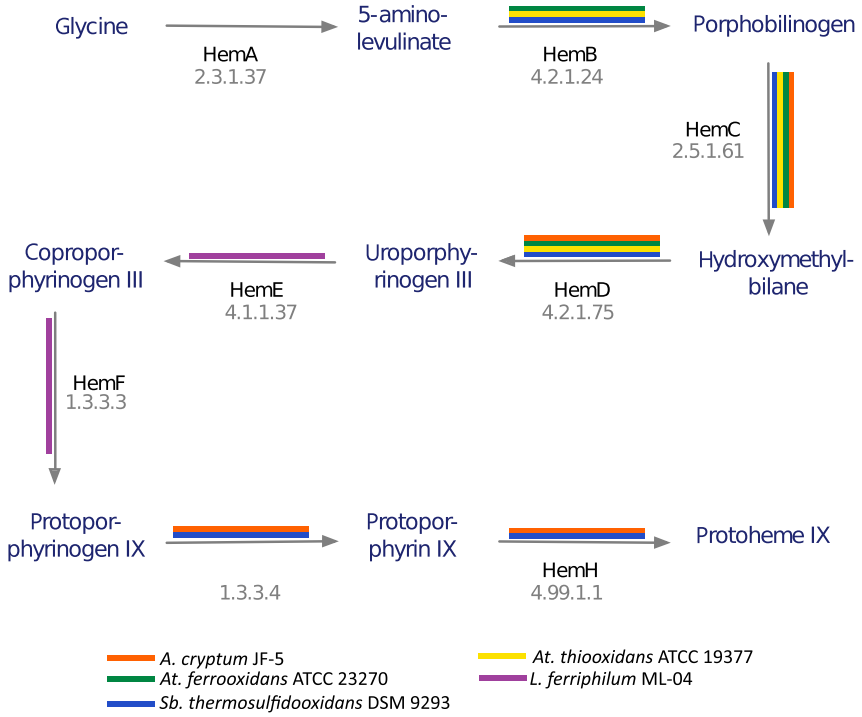


Figure 3. *Superpathway of heme biosynthesis from glycine* from Metacyc (PWY-5920). Each color edge represents a reaction involved into a *sets from genome segments* (SGS). The orange one is for *Acidithiobacillus cryptum*, the purple one for *Leptospirillum ferriphilum*, the blue one for *Sulfobacillus thermosulfidooxidans*, the yellow one for *Acidithiobacillus thiooxidans,* and the green one for *At. ferrooxidans*.

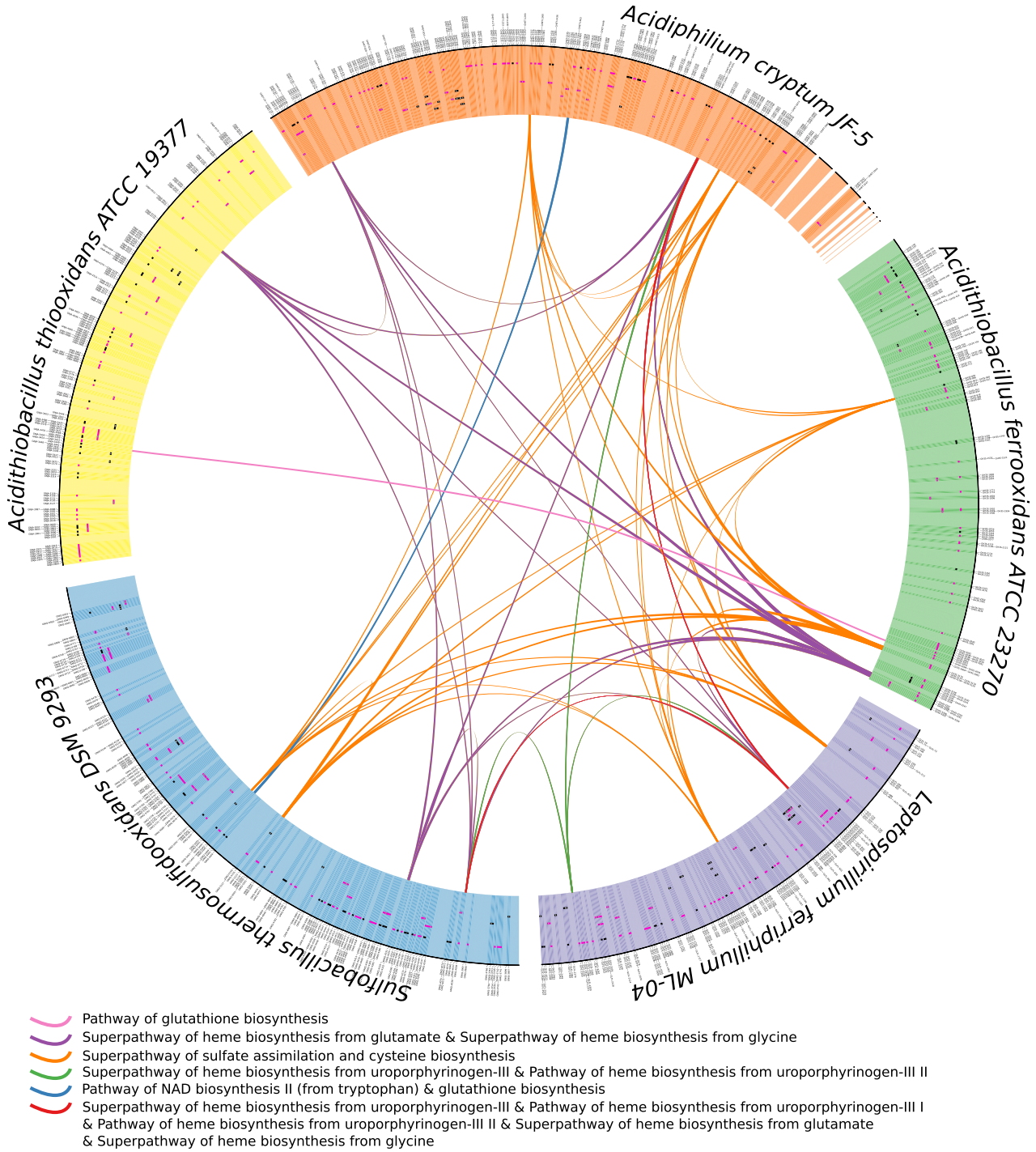


Figure 4. Metabolic copper bioleaching relationships between *sets from genome segments* (SGS) at the metagenomic scale. The outside bands represent the five bacterial genomes. The blue, purple, green, orange and yellow bands refer, respectively, to *Sulfobacillus thermosulfidooxidans*, *Leptospirillum ferriphilum*, *Acidithiobacillus ferrooxidans*, *Acidiphilium cryptum,* and *Acidithiobacillus thiooxidans* genomes. The black and pink segments over the genomes illustrate the SGS, where the pink ones are SGS similar to operons whereas the black ones are those that are not similar to operons. Gray parts of the segments indicate genes that do not participate in the meta-metabolic scale. A link connecting two SGS indicates that those two SGS participate in the same pathway. The color of the link is specific to a set of pathways.

of heme biosynthesis from glycine and the whole community metabolism) with a particular emphasis on monospecific SGS. While SGS are widely spread over the community metabolic network, it is worth noticing that they remain mostly complementary. Indeed, whereas four SGS are totally disconnected from others, mostly all SGS sets of reactions are separated by at most two reactions (98% of SGS sets of reactions are separated by a gap of one reaction only). In particular, bioleaching pathways are very well covered by distinct SGS (see supplementary materials for details): Superpathway of sulfate assimilation and cysteine biosynthesis (nine SGS), NAD biosynthesis I (from aspartate) (2 SGS), NAD biosynthesis II (from tryptophan) (2 SGS), Superpathway of heme biosynthesis from uroporphyrinogen-III (4 SGS), heme biosynthesis from uroporphyrinogen-III I (3 SGS), heme biosynthesis from uroporphyrinogen-III II (4 SGS), Superpathway of heme biosynthesis from glutamate (8 SGS), Superpathway of heme biosynthesis from glycine (8 SGS), and glutathione biosynthesis (2 SGS).

Interestingly, no single bacterial strain is able, alone, to execute a whole pathway related to bioleaching, but rather a cooperation of SGS between different strains is necessary. For illustration, Figure 4 pinpoints putative collaborations of each bacteria to bioleaching pathways via their respective SGS. Each ribbon connects SGS of the five bacteria that participate in the same pathway (one ribbon color per group of bioleaching pathways). Precisely, the bacterial participation is as follows: *A. cryptum* has six linked SGS, whereas *At. ferrooxidans* and *Sb. thermosulfidooxidans* have five each, *L. ferriphilum* has four and *At. thiooxidans* has two. Notably these SGS are homologous (see Fig. S10 and the SGS list). Similarly, Figure 4 (blue) shows a putative homologous collaboration between one SGS from *A. cryptum* and one from *Sb. thermosulfidooxidans* in the NAD biosynthesis I (from aspartate) pathway (see Fig. S5). Contrarily, the red link presents a chaining of two SGS (one from *A. cryptum* and the other one from *L. ferriphilum*) into the five pathways of the heme biosynthesis (the orange and purple segments in Fig. S6–10). The combination of duplicated SGS across bacteria and chaining of SGS is also observed. The purple links depict the putative collaborative participation of five SGS into the Superpathway of heme biosynthesis from glutamate but also the Superpathway of heme biosynthesis from glycine. These superpathways are variants but they share the same reactions. In this pathway, the SGS from *A. cryptum* and *Sb. thermosulfidooxidans*, but also those from *At. ferrooxidans* and *At. Thiooxidans,* are similar and participate into a chain with the SGS of *L. ferriphilum* (see Fig. 3 and Fig. S10).

Transporters and common goods as consequence of SGS combination

Assuming that SGS from different species decipher putative intricate collaborations of microbial strains at the community level, metabolites surrounded by two SGS from different species represent a potential common good that must be, respectively, imported and exported from bacterial cytoplasm via transporters. Following this assumption, APS (adenosine phosphosulfate), PAPS (phosphoadenosine phosphosulfate), protoporphyrinogen, uroporphyrinogen-III, and hydrogen peroxide are theoretical common goods, because they connect SGS reactions for bioleaching pathways. Except hydrogen peroxide these compounds are metabolic intermediates that participate mainly in heme biosynthesis and amino acid synthesis processes (Valdés et al. 2003). While, there is no currently experimental evidence for proteins involved in the direct mobilization of these metabolites inside or outside bacterial organisms, interestingly we identified classical transporters like ATP Binding Cassette (ABC), Resistance Nodulation Division (RND), and Major Facilitator Superfamily (MFS) systems encoded next to SGS involved in sulfate assimilation, serine biosynthesis, and mainly heme pathways. These results pinpoint the already known central role of heme to control the whole bioleaching process (Valdés et al. 2003), whereas APS, PAPS, protoporphyrinogen, and uroporphyrinogen-III may play an indirect role via their modulation of concentrations that potentially impact heme biosynthesis. Interestingly, one must notice that these common good metabolites connect SGS from *L. ferriphilum* on one hand to the SGS from all four other strains on the other hand.

Discussion

Generic paradigm for genome-wide integrative study for a bacterial community

This study integrates genomic and metabolic knowledge to study a reduced microbial community. By using a simple parsimony assumption on topological knowledge (i.e., genome organization and metabolic network), the SGS paradigm proposes a genome-wide description of functional units that are necessary to achieve a given function hold by a microbial community. This approach could be considered as a promising alternative to microbial cross-feeding analysis when quantitative experiments are not feasible. In particular, one considers SGS technique as preliminary to recent Flux Balance like techniques that need to be considered as an objective function or quantitative features for each strain.

These limitations are particularly true for extremophiles strains for which biological knowledge is sparse. Indeed,

whereas genome sequencing and assembly techniques are trustworthy for extremophiles, it remains difficult to consider an exhaustive definition of metabolic networks, that are mostly incomplete (McCloskey et al. 2013) or misannotated (Liberal and Pinney 2013). To overcome this weakness, our SGS technique tolerates uncertainties, via a flexible constraints-based implementation. For instance, one considers genes within SGS that are not directly related to the metabolic pathways of interest or even not depicted as catalytic genes (c.f. gray genes in Fig. 1).

To further confirm this methodological advantage, we advocate for complementary application studies. In particular, SGS applications are of potential interest for analyzing metagenomic or meta-transcriptomic results that investigate, as observed in nature, more complex microbial natural communities. Thus, beyond a further validation of SGS, such a extensive application could be of potential use to give insights about microbial richness and its relation to metabolic pathways use and/or biogeochemical processes of interest.

Compartments are the consequence of genome organization

Interestingly, the bioleaching community shares most of the reactions involved in the analyzed pathways (Fig. 2B), which supports the idea of a metabolic redundancy within the community that usually promotes the use of a "single-cell assumption" to investigate meta-genome experiments. All these shared reactions are part of the core of conserved bacterial pathways present in most strains described to date (Lapierre and Gogarten 2009), indicating that conservation of the pathways is a general feature of bacterial organisms and not a particular property of the bioleaching community. However, despite this high redundancy of metabolic reactions, SGS are not evenly distributed between bacteria of interest (Fig. 2C). SGS mostly covered functional units known as operons. As accepted, the conservation of a gene in an operon is a strong parameter of functional neighborhood inferences, where essential genes are more clustered than the average genes (Nuñez, 2013). Few SGS or functional units are replicated between all bacterial strains, even between pairs of bacteria, which implies that, overall, functional units at the community level are specialized to each strain. When projected on the whole bioleaching metabolic pathways, SGS are mostly complementary. This point shows, at a metagenomic level, evidences that genome organization could play a role to explain cross-feedings between microbial strains at the community level. Furthermore, the bacterial strain-specific SGS distribution characterizes a *functional compartmentalization* that is, again, a direct consequence of a simple parsimony assumption on genome and metabolic knowledge integration. Complementary, no single bacterium is able alone, to monitor the whole copper bioleaching network (Fig. 3 and Fig. S3), which clearly indicates a specialization of the strains. From an evolutionary point of view, this compartmentalization reflects the structured evolution of the five genomes and their corresponding metabolisms. Compartments could be also considered as a way to contain all toxicities that are enhanced by the whole copper bioleaching. In particular, the mining community is composed of extremophiles species that are already known to handle severe and distinct environmental stresses. By promoting bacterial diversity rather than the presence of a single ubiquitous bacterial strain that handles the whole bioleaching, one can assume that the community systems share all stresses based on respective strain capacity while maintaining an overall oxidation process that is chemically and energetically optimal for the whole. From a systems biology point of view, this emergence of functional compartments represents another insight that promotes the concept of modularity for improving the stability of the whole system as previously observed at the single-cell metabolic scale (Kitano 2004; Larhlimi et al. 2011), but herein at the community level. Finally, from a biomining viewpoint, this clearly advocates for the need for considering bacterial compartments and their interactions and confirms, with no a priori, the lack of interest in studying a single bacterial specie like previously *At. ferrooxidans* to embrace the whole bioleaching process.

Putative community common goods

When following a given metabolic pathway, one emphasizes connections between different SGS that belong to different bacterial strains (see Fig. 1 and Fig. 3), supporting the concept that functional interactions between members are crucial (Baker and Banfield 2003). Interestingly, the six bioleaching superpathways are all functionally interconnected via SGS. More precisely, SGS are connected either because they are functionally redundant (e.g., superposed colors in Fig. 3 and Figure S3) – or complementary on metabolic pathways (e.g., a succession of colors in Fig. 3 and Figure S3). In particular, *Sb. thermosulfidooxidans* and *A. cryptum* share most of the SGS. Interestingly, besides its elevated connectivity, *A. cryptum* shows the largest number of SGS related with NAD(H) biosynthesis metabolic pathway, and this despite not considering NAD(H) molecule compound for connectivity purpose (see Method section). Such connections between all five bacterial strains by SGS related to NAD(H) superpathways confirms that bioleaching requires the reduction potential given by the NAD(H) molecule, denoting its multispecificity participation in bioleaching, as well as an elevated robustness to external factors (Yus et al. 2009).

Previous studies also correlate the capability of the community to bioleach copper with the ability to generate biomass, a process that requires NAD potential for its realization (Valdés et al. 2010). Because most of NAD(H) SGS are monitored by *A. cryptum*, this strain plays an unexpected but major role in structuring the bioleaching community, indicating that the collaboration inside the community lies principally in its ability to complement different metabolic functions, as spread between the five strains.

Finally, such collaborations highlight putative transporters between bacterial strains, as well as the common good metabolites shared by two species in the bioleaching context. The analysis of these metabolites surrounded by two SGS indicates potential transporters to seek within genomes, but also confirms the main interest of heme for monitoring bioleaching at the community scale. Indeed SGS analysis advocates that APS, PAPS, protoporphyrinogen and uroporphyrinogen-III are potential regulators of heme pathway. Other transporters, localized next to some SGS, were identified related to serine production and sulfate assimilation. Interestingly, it has been described that in *At. ferrooxidans* these two processes are directly involved in cysteine production and Fe-S cluster formation, two crucial molecules highly required during the bioleaching of copper (Valdés et al. 2003). The genomic proximity between SGS and transporters strongly suggests that strain-specific SGS can interact with other organisms of the consortia through the biosynthesis and transport of common metabolic goods, in this case mainly related with heme and sulfur assimilation pathways. Beyond the bioleaching application and speculative interpretations, SGS results provide functional enrichment to metagenomic knowledge as well as guidelines for future molecular investigations at the community scale, in particular in the search and identification of putative transporters necessary for a cooperative metabolic behavior at the metabolic scale, especially when cross-feeding experiments are not feasible.

Acknowledgments

This work was supported by grants Fondap 15090007, Basal program PFB-03 CMM, IntegrativeBioChile INRIA Assoc. Team, ANR-10-BLANC-0218 and CIRIC-INRIA Chile (line Natural Resources).

Data Archiving

All data and programs are available in: http://philippe.bordron.net/sgs-at-the-community-scale.html. Program available via a web application: http://mobyle.inria.cl/. Supplementary information is available at Journal's website.

Conflict of Interest

None declared.

References

Baker, B. J., and J. F. Banfield. 2003. Microbial communities in acid mine drainage. FEMS Microbiol. Ecol. 44:139–152.

Barreto, M., R. Quatrini, S. Bueno, S. Arriagada, and J. Valdés. 2003. Aspects of the predicted physiology of *Acidithiobacillus ferrooxidans* deduced from an analysis of its partial genome sequence. Hydrometallurgy 71:97–105.

Bordron, P., D. Eveillard, and I. Rusu. 2011. Integrated analysis of the gene neighbouring impact on bacterial metabolic networks. IET Syst. Biol. 5:261–268.

Borenstein, E., M. Kupiec, M. W. Feldman, and E. Ruppin. 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc. Natl Acad. Sci. USA 105:14482–14487.

Boyer, F., A. Morgat, L. Labarre, J. Pothier, and A. Viari. 2005. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. Bioinformatics 21:4209–4215.

Brown, M. V., G. K. Philip, J. A. Bunge, M. C. Smith, A. Bissett, F. M. Lauro, et al. 2009. Microbial community structure in the North Pacific ocean. ISME J. 3:1374–1386.

Caspi, R., T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti, I. M. Keseler, et al. 2012. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res. 40:D742–D753.

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

Christian, N., P. May, S. Kempa, T. Handorf, and O. Ebenhoeh. 2009. An integrative approach towards completing genome-scale metabolic networks. Mol. BioSyst. 5:1889–1903.

Croes, D., F. Couche, S. J. Wodak, and J. van Helden. 2005. Metabolic pathfinding: inferring relevant pathways in biochemical networks. Nucleic Acids Res. 33:W326–W330.

DeLong, E. F. 2005. Microbial community genomics in the ocean. Nat. Rev. Microbiol. 3:459–469.

Drobner, E., H. Huber, G. Wächtershäuser, D. Rose, and K. O. Stetter. 1990. Pyrite formation linked with hydrogen evolution under anaerobic conditions. Nature 346:742–744.

Faust, K., and J. Raes. 2012. Microbial interactions: from networks to models. Nat. Rev. Microbiol. 10:538–550.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, et al. 2012. Microbial co-occurrence relationships in the human microbiome. PLoS Comput. Biol. 8:e1002606.

Guimera, R., and L. A. N. Amaral. 2005. Functional cartography of complex metabolic networks. Nature 433:895–900.

Hingamp, P., N. Grimsley, S. G. Acinas, C. Clerissi, L. Subirana, J. Poulain, et al. 2013. Exploring nucleo-cytoplasmic large DNA viruses in tara oceans microbial metagenomes. ISME J. 7:1678–1695.

Karp, P. D., S. M. Paley, M. Krummenacker, M. Latendresse, J. M. Dale, T. J. Lee, et al. 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform. 11:40–79.

Karsenti, E., S. G. Acinas, P. Bork, C. Bowler, C. de Vargas, J. Raes, et al. 2011. A holistic approach to marine eco-systems biology. PLoS Biol. 9:e1001177.

Kitano, H. 2004. Biological robustness. Nat. Rev. Genet. 5:826–837.

Klitgord, N., and D. Segrè. 2011. Ecosystems biology of microbial metabolism. Curr. Opin. Biotechnol. 22:541–546.

Lapierre, P., and J. P. Gogarten. 2009. Estimating the size of the bacterial pan-genome. Trends Genet. 25:107–110.

Larhlimi, A., S. Blachon, J. Selbig, and Z. Nikoloski. 2011. Robustness of metabolic networks: a review of existing definitions. BioSystems 106:1–8.

Latorre, M., N. Ehrenfeld, M. P. Cortés, D. Travisany, M. Budinich, A. Aravena, et al. 2015. Global transcriptional responses of *Acidithiobacillus ferrooxidans* Wenelen under different sulfide minerals. Bioresour. Technol. 200:29–34.

Levican, G., J. A. Ugalde, N. Ehrenfeld, A. Maass, and P. Parada. 2008. Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations. BMC Genom. 9:581.

Liberal, R., and J. W. Pinney. 2013. Simple topological properties predict functional misannotations in a metabolic network. Bioinformatics 29:i154–i161.

Magnuson, T. S., M. W. Swenson, A. J. Paszczynski, L. A. Deobald, D. Kerk, and D. E. Cummings. 2010. Proteogenomic and functional analysis of chromate reduction in *Acidiphilium cryptum* JF-5, an Fe(III)-respiring acidophile. Biometals 23:1129–1138.

Mao, F., P. Dam, J. Chou, V. Olman, and Y. Xu. 2009. DOOR: a database for prokaryotic operons. Nucleic Acids Res. 37:D459–D463.

McCloskey, D., B. Ø. Palsson, and A. M. Feist. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol. Syst. Biol. 9:661.

Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, et al. 2003. GenDB - an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31:2187–2195.

Mi, S., J. Song, J. Lin, Y. Che, H. Zheng, and J. Lin. 2011. Complete genome of *Leptospirillum ferriphilum* ML-04 provides insight into its physiology and environmental adaptation. J. Microbiol. 49:890–901.

Nuñez, P.A., H. Romero, M.D. Farber, and E.P.C. Rocha. 2013. Natural selection for operons depends on genome size. Genome Biol Evol, 5:2242–2254.

Orphan, V. J. 2009. Methods for unveiling cryptic microbial partnerships in nature. Curr. Opin. Microbiol. 12:231–237.

Patel, P. V., T. A. Gianoulis, R. D. Bjornson, K. Y. Yip, D. M. Engelman, and M. B. Gerstein. 2010. Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. Genome Res. 20:960–971.

Quatrini, R., C. Appia-Ayme, Y. Denis, E. Jedlicki, D. S. Holmes, and V. Bonnefoy. 2009. Extending the models for iron and sulfur oxidation in the extreme acidophile *Acidithiobacillus ferrooxidans*. BMC Genom. 10:394.

Raes, J., I. Letunic, T. Yamada, L. J. Jensen, and P. Bork. 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol. Syst. Biol. 7:473.

Ranjard, L., S. Dequiedt, N. Chemidlin Prévost-Bouré, J. Thioulouse, N. P. A. Saby, M. Lelievre, et al. 2013. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. Nat. Commun. 4:1434.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. 2002. Hierarchical organization of modularity in metabolic networks. Science 297:1551–1555.

Roesch, L. F. W., R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, et al. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. 1:283–290.

Ruan, Q., D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun. 2006. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. Bioinformatics 22:2532–2538.

Ruan, J., A. K. Dean, and W. Zhang. 2010. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst. Biol. 4:8.

Segata, N., D. Boernigen, T. L. Tickle, X. C. Morgan, W. S. Garrett, and C. Huttenhower. 2013. Computational meta'omics for microbial community studies. Mol. Syst. Biol. 9:666.

Taboada, B., R. Ciria, C. E. Martinez-Guerrero, and E. Merino. 2012. ProOpDB: Prokaryotic Operon DataBase. Nucleic Acids Res. 40:D627–D631.

Travisany, D., A. Di Genova, A. Sepúlveda, R. A. Bobadilla-Fazzini, P. Parada, and A. Maass. 2012. Draft genome sequence of the *Sulfobacillus thermosulfidooxidans* Cutipay strain, an indigenous bacterium isolated from a naturally extreme mining environment in Northern Chile. J. Bacteriol. 194:6327–6328.

Tzamali, E., P. Poirazi, I. G. Tollis, and M. Reczko. 2011. A computational exploration of bacterial metabolic diversity identifying metabolic interactions and growth-efficient strain communities. BMC Syst. Biol. 5:167.

Valdés, J., F. Veloso, E. Jedlicki, and D. Holmes. 2003. Metabolic reconstruction of sulfur assimilation in the extremophile *Acidithiobacillus ferrooxidans* based on genome analysis. BMC Genom. 4:51.

Valdés, J., I. Pedroso, R. Quatrini, R.J. Dodson, H. Tettelin, R. Blake, et al. 2008. *Acidithiobacillus ferrooxidans* metabolism: from genome sequence to industrial applications. BMC Genom. 9:597.

Valdés, J., J. P. Cárdenas, R. Quatrini, M. Esparza, H. Osorio, F. Duarte, et al. 2010. Comparative genomics begins to unravel the ecophysiology of bioleaching. Hydrometallurgy 104:471–476.

Valdés, J., F. Ossandon, F. Quatrini, M. Dopson, and D. S. Holmes. 2011. Draft genome sequence of the extremely acidophilic biomining bacterium *Acidithiobacillus thiooxidans* ATCC 19377 provides insights into the evolution of the Acidithiobacillus genus. J. Bacteriol. 193:7003–7004.

de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605.

Wächtershäuser, G. 2000. Origin of life. Life as we don't know it. Science 289:1307–1308.

Yin, H., L. Cao, M. Xie, Q. Chen, G. Qiu, and J. Zhou. 2008. Bacterial diversity based on 16S rRNA and gyrB genes at Yinshan mine, China. Syst. Appl. Microbiol. 31:302–311.

Yus, E., T. Maier, K. Michalodimitrakis, V. van Noort, T. Yamada, W. H. Chen, et al. 2009. Impact of genome reduction on bacterial metabolism and its regulation. Science 326:1263–1268.

Zaneveld, J. R. R., L. W. Parfrey, W. Van Treuren, C. Lozupone, J. C. Clemente, D. Knights, et al. 2011. Combined phylogenetic and genomic approaches for the high-throughput study of microbial habitat adaptation. Trends Microbiol. 19:472–482.

Zelezniak, A., S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil. 2015. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proc. Natl Acad. Sci. USA. doi:10.1073/pnas.1421834112.

Zengler, K., and B. Ø. Palsson. 2012. A road map for the development of community systems (CoSy) biology. Nat. Rev. Microbiol. 10:366–372.

Zomorrodi, A. R., M. M. Islam, and C. D. Maranas. 2014. d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. ACS Synth. Biol. 3:247–257.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1. Definition.

Appendix S2. Choosing SGS parameters.

Appendix S3. Transcriptomic insights.

Appendix S4. Coexpression in SGS.

Figure S3. Projection of reaction sets corresponding to SGS on the meta-metabolic network.

Figure S4. Superpathway of sulfate assimilation and cysteine biosynthesis from Metacyc (SULFATE-CYS-PWY).

Figure S5. Pathway of NAD biosynthesis I (from aspartate) from Metacyc (PYRIDNUCSYN-PWY).

Figure S6. Pathway of NAD biosynthesis II (from tryptophan) from Metacyc (NADSYN-PWY).

Figure S7. Superpathway of heme biosynthesis from uroporphyrinogen-III from Metacyc (PWY0-1415).

Figure S8. Pathway of heme biosynthesis from uroporphyrinogen-III I from Metacyc (HEME-BIOSYNTHESIS-II)

Figure S9. Pathway of heme biosynthesis from uroporphyrinogen-III II from Metacyc (HEMESYN2-PWY).

Figure S10. Superpathway of heme biosynthesis from glutamate from Metacyc (PWY-5918).

Figure S11. Pathway of glutathione biosynthesis from Metacyc (GLUTATHIONESYN-PWY).

Data S1. The file SpplementarymaterialsS1.xls contains the list of orthologs between the *At. ferrooxidans* ATCC23270 and *At. ferrooxidans* Wenelen strains.

Data S2. The file SpplementarymaterialsS2.xls contains the complete list of SGS.

Data S3. The file SpplementarymaterialsS3.xls contains the list of SGS involved in the bioleaching pathways.

3.4 Discussion

These results show the interest of using a weighted graph to integrate different biological knowledge. Although the choice of this computational abstraction is consistent with those proposed in Chapter 2, one must discuss here two complementary viewpoints. First, contrary to the undirected graph used to depict co-occurrence networks, the integration of different knowledge necessitates the use of directed graphs. Such a direction is the result of considering additional constraints, such as those enclosed in a public metabolic database and available metabolic networks. Here again, the interest in computer sciences consists in modeling the knowledge integration by identifying the best abstraction in which one can solve a well-defined biological problem. Second, and complementary to the sole description of the biological abstraction (i.e., directed weighted network), we have shown herein the need for an optimization process that computes the result of a parsimonious assumption on the integrated graph. The choice of the optimization must be coherent with the biological question but also with the chosen abstraction. Performing such optimization will allow focussing on graph local properties. As a result, such a property indicates either functional units or putative metabolic cross-feeding between microbial strains in a more global ecosystem framework.

Although performing such a protocol for investigating a cellular system remains feasible, the same question is computationally challenging when dealing with large ecological networks. As a first attempt, we proposed herein the use of Answer Set Programming (ASP) to enumerate local properties. Again, ASP is a declarative problem-solving paradigm from the field of logic programming and knowledge representation, that offers a rich modeling language along with highly efficient inference engines [80] based on Boolean constraint solving technology. Following encouraging preliminary results of ASP applied on integrated networks, one proposes to generalize the use of ASP to other systems ecology questions, for instance, to reduce the size of ecological networks while conserving local properties (i.e., network delineation). Indeed some edges represent false positives that could be removed from the graph with no degradation of global knowledge such as the clusters obtained using topology cluster analysis. To manage the combinatorial explosion of this removal, ASP will allow a flexible encoding that can be easily adjusted to test different pairwise metrics while still being computationally efficient. Once the graph reduced, we will be able to apply the previously mentioned graph decomposition techniques [46]. These techniques will emphasize patterns of species or genes as well as cliques of species and their associations with environmental parameters. (i.e., bacteria consortium or set of genes that are related to the nitrate concentration or carbon export). As natural following, one will investigate these identified biological compounds by studying their dynamical properties via

dedicated modelings and simulations, as depicted in Chapter 4.

Synthesis of the contributions to Integrative Biology

- Stanislas Thiriet-Rupert, Gregory Carrier, Camille Trottier, Damien Eveillard, Benoit Schoefs, Gael Bougaran, Jean-Paul Cadoret, Benoit Chénais, and Bruno Saint-Jean. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of Tisochrysis lutea. *Algal Research*, 30:59–72, March 2018

- Julie Laniau, Clémence Frioux, Jacques Nicolas, Caroline Baroukh, Maria Paz Cortés, Jeanne Got, Camille Trottier, Damien Eveillard, and Anne Siegel. Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, 5:e3860, 2017

- François Thomas, Philippe Bordron, Damien Eveillard, and Gurvan Michel. Gene Expression Analysis of Zobellia galactanivorans during the Degradation of Algal Polysaccharides Reveals both Substrate-Specific and Shared Transcriptome-Wide Responses. *Frontiers in microbiology*, 8:1808, 2017

- Sylvain Prigent, Clémence Frioux, Simon M Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, Frédéric Plewniak, Thierry Tonon, and Anne Siegel. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. *PLoS computational biology*, 13(1):e1005276, January 2017

- Vicente Acuña, Andrés Aravena, Carito Guziolowski, Damien Eveillard, Anne Siegel, and Alejandro Maass. Deciphering transcriptional regulations coordinating the response to environmental changes. *BMC bioinformatics*, 17(1):35, January 2016

- Philippe Bordron, Mauricio Latorre, Maria Paz Cortés, Mauricio González, Sven Thiele, Anne Siegel, Alejandro Maass, and Damien Eveillard. Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen*, 5(1):106–117, February 2016

- Thierry Tonon and Damien Eveillard. Marine systems biology. *Frontiers in genetics*, 6(20):181, 2015

- Simon M Dittami, Damien Eveillard, and Thierry Tonon. A metabolic approach to study algal-bacterial interactions in changing environments. *Molecular ecology*, 23(7):1656–1660, April 2014

- Guillaume Collet, Damien Eveillard, Martin Gebser, Sylvain Prigent, Torsten Schaub, Anne Siegel, and Sven Thiele. Extending the Metabolic Network of Ectocarpus Siliculosus using Answer Set Programming. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, LNAI8148:245–256, September 2013

- Philippe Bordron, Damien Eveillard, Alejandro Maass, Anne Siegel, and Sven Thiele. An ASP application in integrative biology: identification of functional gene units. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, 8148:206–218, 2013

- Thierry Tonon, Damien Eveillard, Sylvain Prigent, Jérémie Bourdon, Philippe Potin, Catherine Boyen, and Anne Siegel. Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment. *Omics : a journal of integrative biology*, 15(12):883–892, December 2011

- Philippe Bordron, Damien Eveillard, and Irena Rusu. SIPPER: A flexible method to integrate heterogeneous data into a metabolic network. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pages 40–45. IEEE, February 2011

- Philippe Bordron, Damien Eveillard, and Irena Rusu. Integrated analysis of the gene neighbouring impact on bacterial metabolic networks. *IET systems biology*, 5(4):261–268, July 2011

Chapter 4

Dynamical and quantitative modelings of biological systems

A whole, which is not a heap, is not identical to the elements, into which it is divided. The whole is (contains) something else besides the elements. The something else is not an element. The something else does not consist of a single element. The something else does not consist of many elements. The something else is a principle (cause, substance, nature). Suppressed conclusion: The something else is the form of the whole.

Aristotle

4.1 Introduction & Context

Above mentioned advances in molecular biology and computational biology have transformed approaches to qualitatively characterize biological systems. However, among the most significant challenges in biology is the ability to quantitatively predict a phenotype, by combining molecular data and quantitative physicochemical data. Since the seminal work of Jacques Monod [150], simple biological modeling has been prominent in microbiology. Because of their experimental tractability and purported simplicity, microbial experimental systems have fostered the rise

of several cross-scale modeling approaches from the gene to the population level, which have been extended to test eco-evolutionary hypotheses. These modeling approaches proposed and addressed foundational hypotheses that developed into new biological paradigms such as growth rate or identification of functional units. Reductionist assumptions were driving the first microbial models (e.g., intracellular quota combined with kinetics mimicking biochemistry rules) yet demonstrated remarkable predictive power for portraying the growth of microbes in simple systems such as chemostats [150]. The Monod model takes the following form:

$$\begin{aligned} \frac{dS}{dt} &= D.S_{in} - \frac{1}{Y}.\mu_{max}\frac{S}{S+k_S}, \\ \frac{dX}{dt} &= \mu_{max}\frac{S}{S+k_S} - D.X, \end{aligned}$$

where bacterial biomass (X) is subject to variation following its Michaelis-Menten growth (μ_{max} and k_S being respectively maximum growth rate and half life constants) and dilution (D); and uptake efficiency (Y); and quantity of substrat (S) is subject to variation due to the dilution of initial quantity of substrat (S_{in}) and the efficiency of the substrate uptake from the microbial strain. Similar quota assumptions were used to model phytoplankton physiology [52, 12] and later for modeling simplified global ocean ecosystems [67]. Reductionist modeling approaches have generally been parameterized from data gleaned from laborious bench experiments. Thus, initial biological modeling efforts were inspired by models of physical systems and formalized using nonlinear ordinary differential equations representing dynamic behaviors of gene activity or molecular concentrations:

$$\frac{dx_i}{dt} = f_i(\mathbf{x}), 1 \leq i \leq n,$$

where $\mathbf{x} = [x_1, ..., x_n]' \geq 0$ represents a vector of concentrations of proteins or quantity of species biomass or any quantity of biological compounds (with respect to the units) and subject to biological transformations represented by a function f_i. Reaction rates associated with particular mechanistic behaviors such as Michaelis-Menten or Hill functions drive interactions within these models and express the rate of synthesis of i as dependent upon the concentration of $\mathbf{x}$. As summarized in [43], to consider negative feedback loops, one must consider more general equations in the form:

$$\begin{aligned} \frac{dx_1}{dt} &= \kappa_{1n} r(x_n) - \gamma_1 x_1, \\ \frac{dx_i}{dt} &= \kappa_{i,i-1} x_{i-1} - \gamma_i x_i, \quad 1 < i \leq n, \end{aligned}$$

where $\kappa_{1n}, \kappa_{2,1}, ...\kappa_{n,n-1}$ are production or consumption constants when positive or negative resp., and $\gamma_1, ..., \gamma_n > 0$ degradation constants; and for x_1, $r(x_n)$ describes a non linear regulation function.

When applied to communities, similar mathematical systems are used to describe specific population growth rates as a function of densities of other populations. Among others, the Lotka-Volterra model is the most known. Complementary types of nonlinear differential equation modeling include spatial representation, which is useful to represent phenomena such as biofilm formation. However, as the complexity of the biological system increases, it is often not possible to model a population behavior through simple equations. Individual-Based Models generally overcome this problem. In this modeling, one considers sets of cells. Each cell describes a space by its coordinates, and its behavior depends on an algorithm that makes decisions based on the population enclosed in it [160]. Iterative simulations are then run to mimic the population in space and extracting emerging rules or global properties from complex local behaviors [89].

To model biogeochemical cycles within ecosystems, the objective is to describe the dynamics of one or more elements (carbon, nitrogen, phosphate, sulfur, and other nutrients) for a given ecosystem. Usually, different organisms are thus classified into functional groups, and one considers flows of chemical elements between them by following quantities in state variables. Environmental parameters are then introduced as forced constraints. In this context, these so-called Trait-Based Models link specified traits, i.e., properties at an individual scale (e.g., size and concentration) to ecological function, such as energy or matter flux, primary production, acid production. 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 always available and need to be inferred from extensive experimental data and validated by experts. This limitation illustrates the general restriction of nonlinear differential equation modelings that necessitate more effort during the parametrization step than in the modeling per se. Furthermore, worth noticing, parameter values obtained from in-vitro experiments can differ from in-vivo conditions. Moreover, for technical reasons, these parameters are even less reachable for cellular systems.

4.2 Qualitative modeling to simulate cellular systems

The need to use formal methods to analyze such cellular models has been discussed extensively in the last 15 years (see De Jong [43] for the state-of-the-art review or Fisher and Henzinger [66]). The application of formal methods mainly results in the discretization of all interactions, such that the model becomes qualitative. An advantage in this context is that the resulting qualitative models are computationally scalable and do not need to incorporate a large number of parameters that

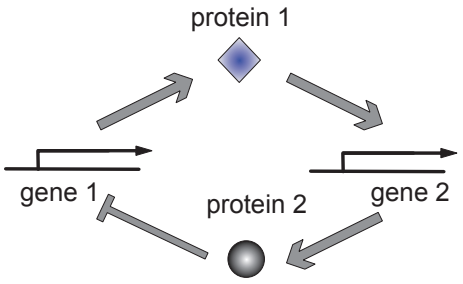


Figure 4.1: Schema representing the two major rules considered in a Gene Regulatory Network. The transcription and translation of gene 1 activates the transcription of genes, which could be formalized by a signed edge $g_1 \xrightarrow{+} g_2$. Reversely, the transcription and translation of gene 2 represses the transcription of gene 1, which could be formalized by the signed edge $g_2 \xrightarrow{-} g_1$

are mostly out of experimental reach, including parameters that show evident sensitivity to experimental conditions [54]. Thus such modelings consider the gene interaction as the cornerstone to represent a biological behavior. It summarizes a protein production that activates or represses the target gene. From a biological viewpoint, this feature is complementary to the metabolic network illustrated above but consider the feedback of the protein production to the gene activity, a biological feature that was neglected when the sole metabolic network is considered. From a computational viewpoint, these modeling approaches exploit the structure of signed directed graphs (e.g., interlocked feedback loops) rather than the numerical values of biological compound concentrations

Among the qualitative modeling techniques, most known approaches are based on Piecewise-Affine Differential Equations (PADEs) [44] or on the René Thomas's formalism and extended works [197, 196]. They all gave astonishing results when applied on real biological systems, in particular to decipher putative marker genes of disorders [37]. Thus, above dynamical systems in PADEs could be resumed by:

$$\frac{dx_i}{dt} = g_i(\mathbf{x}), \quad 1 \leq i \leq n,$$

The function g_i is defined as

$$g_i(\mathbf{x}) = \sum_{l \in L} \kappa_{i,l} b_{i,l}(\mathbf{x})$$

where $b_{i,l}$ is a Boolean function that indicates the conditions under which the gene is expressed at a rate $\kappa_{i,l}$, and L a set of vertices pointed to the gene i. Following the

description from [43], the Boolean conditions are specified by two step functions s^+ and s^- such as

$$s^+(x_j, \theta_j) = \begin{cases} 1 & \text{if} \quad x_j > \theta_j \\ 0 & \text{if} \quad x_j < \theta_j \end{cases} \text{ and } \quad s^-(x_j, \theta_j) = 1 - s^+(x_j, \theta_j)$$

where θ_j represents the threshold above which the gene x_j is activated. Considering such a qualitative abstraction, qualitative states (combination of genes that are activated or inactivated) depict the behavior of the system at one given time and whereas transitions between these qualitative states resume the dynamical biological behaviors. All transitions between qualitatives states describe the state graph (directed graph where nodes are qualitative states and edges are transitions between them) that one investigate via automatic methods for the sake of model qualitative validation. As shown in [10, 169], these techniques correspond to a class of hybrid systems [81] for which we can use existing powerful methods for the verification and the control of these hybrid systems. In particular, they permit an automatic investigation of qualitative properties of the genetic regulatory networks [11].

Nonetheless, even if such models are sufficient to represent microbial gene regulatory networks, they are generally not enough for modeling quantitative biological behaviors as needed in the context of simulation of microbial populations and communities, and subsequently, biogeochemical processes. Nevertheless, one can mention here an extension of these qualitative models that propose to consider temporal properties into account. It consists of a new class of hybrid systems [60], dedicated to biological system modeling with the time delay. Note that such a setting was often neglected before, despite documented variations of specific products over time. The time delay represents a unique opportunity to refine existing qualitative models by showing qualitative properties that verify experimental temporal constraints. Conversely, it emphasizes a need for modeling that includes both qualitative properties, arisen from the biological network structure, and delays associated with the dynamics of genes or gene products [184]. To a lesser extent, in [73], we also propose a new hybrid modeling technique that uses the qualitative and partial temporal experimental data directly. Such modeling does not claim to substitute for existing qualitative modelings but remains a preliminary approach for investigating the complex biological system. As a significant feature, it abstracts the structure of the network, i.e., positive and negative feedback loops, by focusing on the variation of signs of the gene products following given qualitative behaviors (see Figure 4.2 for illustration). This discretization will be further used in Section 4.4. In this qualitative abstraction, as described in [73], we add some constraints on delays for a natural refinement of the qualitative response.

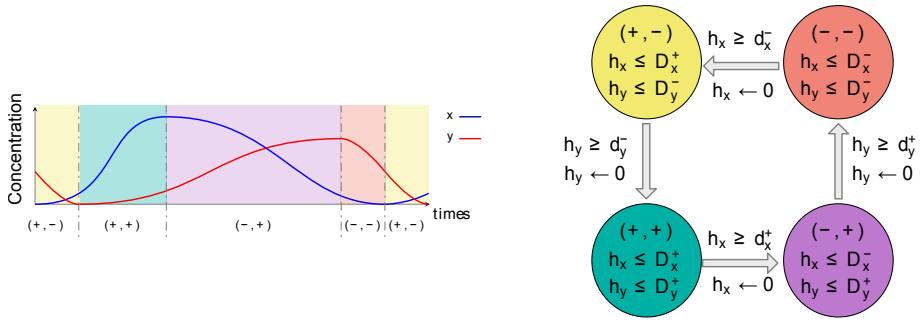


Figure 4.2: Concentration variations over time are discretized by considering the sign of the derivative. On the left panel, the whole simulation of two variables x and y from the Figure **??** describes a qualitative cycle depicted in the right panel. Such a cycle describe the structure of an hybride automaton where nodes are qualitative states (like (+,+)) and transitions describe how to reach one qualitative state from another. On each qualitative state, one considers an additional constraint constraint called invariant ($h_x \leq D_x^+$ and $h_y \leq D_y^+$), that represent the clock and the delay in which one remains in the given qualitative state (time in the increase of x and y concentrations). For each transition (e.g., (+,+) to (-,+)), there is both a reset of a clock ($h_x \leftarrow 0$) and a constraint called guard that forbid to take the transition before a given delay ($h_x \leq d_x^+$, that describes the time for which x must increase).

4.3 Quantitative modeling of biological behaviors at steady states

Quantifying the biological response of living systems is a question that still eludes us. Because of their physical nature, biological objects are subject to quantitative measurements that mostly represent their physiologies. We advocate herein that these motivations were the fundamental assumption for modeling living systems via ordinary differential equation systems. In the above Section 4.1, we present the interest of qualitative modeling to describe biological behaviors over time formally, without considering kinetic parameters. However, these modelings did not perceive the gain of a description as proposed by the omics knowledge. For instance, in his seminal model, Jacques Monod mimics the physiology of a heterotrophic bacteria with two variables. In contrast, the current genome-scale metabolic network description of *Eschericha coli* (strain K12, MG1655) implies the use of 3011 metabolites linked by reactions driven by 1402 associated enzymes and 387 associated transporters [178]. Considering this high number of variables, a standard

dynamical model is out of reach. So, for practical reasons, several new formalisms have been proposed to model metabolism and, by extension, a better understanding of physiological responses [208]. Thus, the use of *Genome-Scale Models* has gained much interest recently to describe organism physiology. In this framework, complementary to metabolic network description that defines chemical species and their interactions (see Section 3.1), one must consider exchanges of an individual organism with the media (for example, maximal uptake rate) as well as its growth rate [112, 158] Genome-Scale Models must then consider full genomic descriptions of microorganisms and but also perform quantitative simulations.

4.3.1 Quantitative modeling at steady states of organisms described at genome-scale

Modeling a metabolic network implies to describe how metabolites are exchanged and transformed within a network. For instance, a general description can take the form of a chemical equation:

$$a_1 S_1 + a_2 S_2 + ... + a_p S_p \longrightarrow ... + a_{k-1} P_{k-1} + a_k P_k$$

where S_i and P_k depicts one metabolite among the k that is respectively a substrate or a product of the reaction, and a_i its corresponding stoichiometry to satisfy the mass balance law. By definition, substrates are on the left side of the equation, whereas product one the right one. Considering such a reaction, one formulate its reaction rate by the rate of degradation of substrate (i.e., negative by convention); or the rate of production of product (i.e., positive by convention).

$$v_i = -\frac{1}{a_1}\frac{dS_1}{dt} = -\frac{1}{a_2}\frac{dS_2}{dt} = ... = \frac{1}{a_k}\frac{dP_k}{dt} = \frac{1}{a_p}\frac{dM_p}{dt}$$

with i corresponding to a given reaction, and considering that a metabolite M_i is either product or consumed, its variation rate is express by:

$$\frac{dM_i}{dt} = a_{i1} v_1 + a_{i2} v_2 + ... + a_{in} v_n = S_i.\mathbf{v} = \sum_{j=1...n} a_{ij} v_j$$

Using a vector notation, the above equation could be written:

$$\frac{d\mathbf{M}}{dt} = \mathbf{S}\mathbf{v},$$

where $\mathbf{M}$ is a vector that enclosed all metabolite concentrations, $\mathbf{v}$ is flux vector that summarizes all fluxes v_i, and $\mathbf{S}$ the matrix that stores stoichiometric coefficients. Worth the notice here, the definition of this dynamical system is related to

the one proposed by Jacques Monod but with linear rates. Notice also that the effect of the temperature does not hold here. Complementary to this definition of the metabolic network, one must consider metabolites that belong to the surrounding environment. By convention, one considers a *synthetic exchange reaction* such as:

$$M_{iex} \longrightarrow M_i, \text{or}$$
$$M_i \longrightarrow M_{iex}$$

for uptake and secretion respectively.

Considering that most of the bacteria and eukaryotic cells are homeostatic, they keep their internal metabolite concentrations as constant as possible, which occurs very fast modifications, one assumes the system at quasi-steady-states [205], where

$$\frac{d\mathbf{M}}{dt} = \mathbf{S}\mathbf{v} = 0$$

In addition to this system of linear constraints, one can also consider the bounds on fluxes to model thermodynamical laws associated to each reaction. By convention, a flux v_i associated to an irreversible reaction must satisfy the following inequality:

$$0 < v_i \leq ub_i,$$

where ub_i represents the upper bound of the flux v_i, and the lower bound of v_i is equal to 0. Following the similar convention, fluxes of reversible reactions must satisfy another inequality:

$$lb_i \leq v_i \leq ub_i,$$

where lb_i represent the lower bound of the flux v_i and $lb_i < 0$.

Steady-state flux space All solutions of $\mathbf{v}$ that satisfies these constraints are biologically accurate. The set of solutions describes a steady-state flux space F defined as:

$$F := \{\mathbf{v} \in \mathbb{R}, \mathbf{S}\mathbf{v} = 0, \quad \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub}\}$$

where $\mathbf{lb}$ and $\mathbf{ub}$ are all lower and upper bounds of fluxes stored in $\mathbf{v}$. Because F encloses all biologically feasible solutions, several studies focused on its formal description. Among others, one could mention several generators of F that are of biological interest: the *Elementary modes* [182, 181, 76, 42] that compute the vertices of the polyhedron decsribed by F; the *Extreme pathways* [162, 214] that compute extreme rays of the pointed polyhedron (i.e., reversible reactions are modeled by two irreversible reactions of opposite direction - which states the extreme pathways as subsets of elementary modes [115]); or the *Minimal Metabolic Behaviors* [126] that considers the polyhedron rooted in an affine space and then describes the generators of F via polyhedron face descriptions rather than its vertices for above approaches.

Flux balance analysis Because the flux space is challenging to interpret, *Flux Balance Analysis* (FBA) proposes to select a subset of fluxes that is of particular interest [205]. The selection relies on Linear Programming (LP). It consists of adding a function $\mathbf{c}$ to maximize (or minimize). Such an objective function usually model a biological objective; i.e., maximizing given biological components of interest that constitute the biomass of the system. The formulation of the constraints is then as follow:

$$\begin{aligned} \text{maximize} \quad & z = \mathbf{c}^T\mathbf{v} \\ \text{subject to} \quad & \\ & \mathbf{Sv} = 0 \\ & \mathbf{lb} \leq \mathbf{v} \leq \mathbf{ub} \end{aligned}$$

However, it is important to notice here that if an optimal value z exists, this is unique, which unfortunately cannot be guaranteed for the corresponding values of $\mathbf{v}$. Indeed, a priori, many flux distributions could satisfy unique optimal objective function values, which could bring quantitative uncertainties that are inherent to the biological modeling (see [167] for general overview of FBA limitations).

Flux variability analysis To analyze the multiplicity of flux solutions that satisfy an optimal value of z, [139] proposed an automatic exploration of these multiple optimal flux distribution. Considering z_{obj} the optimal objective value as computed by above FBA, one can formalize both complementary LP problems for each reaction i:

Case 1

$$\begin{aligned} \text{maximize} \quad & v_i \\ \text{subject to} \quad & \\ & \mathbf{c}^T\mathbf{v} = z_{\text{obj}} \\ & \mathbf{Sv} = 0 \\ & lb_i \leq v_i \leq ub_i, \quad i = 1, ..., n \end{aligned}$$

Case 2

$$\begin{aligned} \text{minimize} \quad & v_i \\ \text{subject to} \quad & \\ & \mathbf{c}^T\mathbf{v} = z_{\text{obj}} \\ & \mathbf{Sv} = 0 \\ & lb_i \leq v_i \leq ub_i, \quad i = 1, ..., n \end{aligned}$$

The range of values for each flux for the given optimal objective value are very informative. First, the size of the range is an indicator of the theoretical variability that could be associated to a given reaction (e.g., eventually of its corresponding gene expression). Such a property could be relied on the biological robustness and quantitative uncertainties as taken from experiments [125]. In particular, recently, Basler et al. [8] have shown that those boundaries that reflect uncertainties are necessary to understand intracellular metabolic flux distribution better. Second,

considering the respective maximal and minimal values of v_i (resp. $v_i^{\max}$ and $v_i^{\min}$), one could extract that the qualitative property of the associated reaction.

- $0 \notin [v_i^{\min} \mathrm{v}_i^{\max}]$ means that the reaction i is *obligatory*. A flux must always occurs in this reaction such as the system could optimize its objective (i.e., its biomass).

- $0 \in [v_i^{\min} \mathrm{v}_i^{\max}]$ means that the reaction is *alternative*. A null flux may occur in this reaction while still promoting the optimal solution of the system objective.

- $v_i^{\max} = 0 \quad$ and $\quad v_i^{\min} = 0$ means that the reaction is *blocked*. A no flux must occur in this reaction to reach the optimal solution of the system objective.

Altogether these techniques belong to Constraints-Based modeling. The interest of this modeling paradigm consists in putting more effort into formalizing the problem than in solving it. Moreover, the solutions are not deterministic in the way that one defines all the constraints that occur, and all solutions belong to the solution space. Then, as a natural follow up, one can learn properties from the analysis of this space. A massive number of modeling techniques are following this schema with several applications in metabolic engineering (see [163] for the state-of-art review of the domain). Among them, one could mention the dynamic extension of FBA that overcomes the quasi-steady-state hypothesis. This method divides the simulation into time intervals and computes FBA for each of them. The results computed for one interval will be used as an external input for the computing FBA in the following interval.

4.3.2 Quantitative modeling of genome-scale microbial communities at steady states

Following the rise of omics knowledge and the metabolic description of ecosystems in Section3.1, modeling the metabolism of the microbial ecosystem saw a recent gain of interest [117]. In particular, two publications by [13] and [158], had reviewed the use of metabolic modeling in communities. In both works, they promote the use of Constraint-Based Modeling as previously done on single organisms, but mention three modeling extensions: Lumped or "Soup," compartmentalization and bi-level optimization.

Lumped or "Soup" approach is perhaps the most straightforward approach. It consists of ignoring the boundaries between species and stores all detected genes (and corresponding reactions) into a single virtual entity, assuming a generalized biomass objective function that could represent the whole community. Here, the

metabolic capabilities of all organisms present are the focus. However, one must notice here that this approach changes the fundamental properties of the network and the accuracy of flux values carried out by specific organisms (see [116]).

On the contrary, the compartmentalization approach considers each species as a compartment of the ecosystem, and compartments share metabolites via an extra compartment, which represents the extracellular environment. Thus a general matrix that depicts the ecosystem is composed of several stoichiometric matrices that depict individual metabolic models. Each metabolite is defined not by its chemical definition but also by the compartment it is involved (i.e., metabolite M_1 will be defined as $M_{1\,A}$ to describe the appurtenance of M_1 to the organism A). The exchange reactions between each species and the shared extracellular space allow capturing mechanistic interactions such as mutualism or competition [188, 192, 117, 111]. In these modelings, the sum of the biomass of all organisms describes the objective function of the ecosystem, that one will optimize.

Finally, the OptCom framework constitutes the bi-level formulation of an optimization problem for modeling an ecosystem [223, 222]. OptCom addresses the optimization of several problems: (i) inter-organisms constraints to model a given mechanistic scenario and (ii) intra-organism constraints as already formulated by standard FBA.

As an alternative, we proposed another description of a microbial ecosystem in:

Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE*, 12(2):e0171744, 2017

In this study, we consider (i) the use of multiple compartments to model the interplay of several organisms; (ii) a specific objective function is associated with each compartment (but not the extracellular compartment); and (iii) all these objective functions must be maximized concurrently. This new formulation calls for a change of modeling paradigm: Multi-Objective Optimization. The ecosystem solution space will then be described as a Pareto front that results when multiple objectives are satisfied.

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

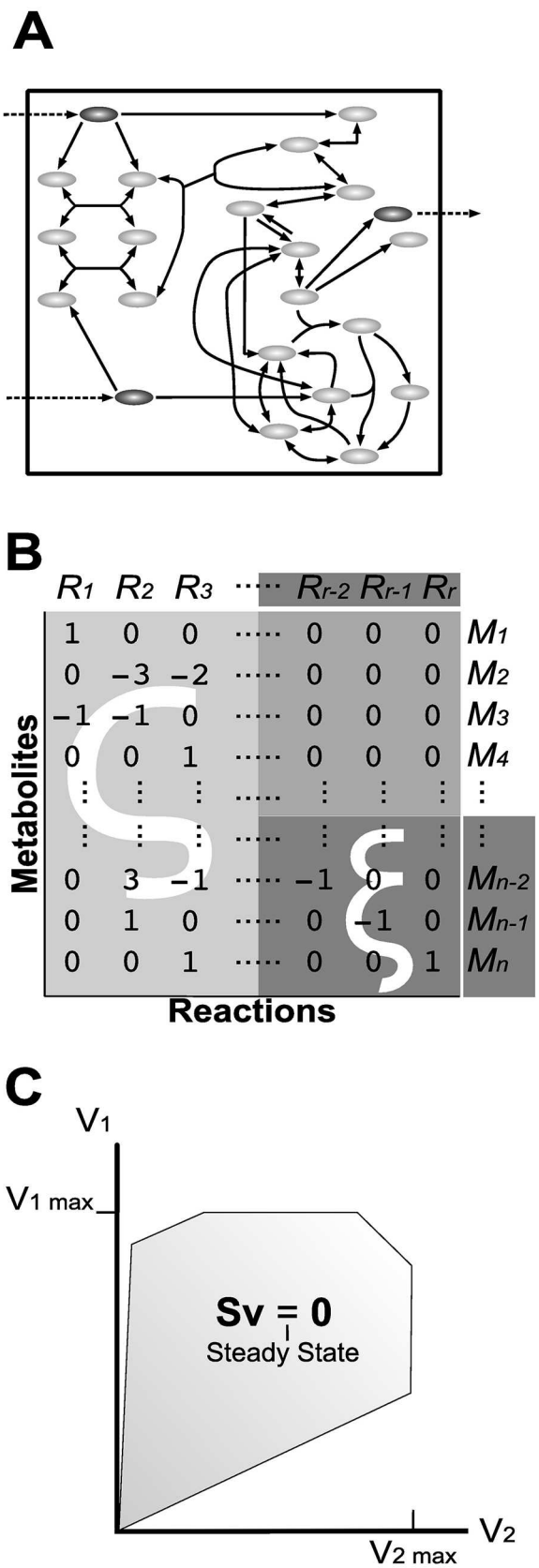

A
B
R1 R2 R3 ····· Rr-2 Rr-1 Rr
1 0 0 ····· 0 0 0 M1
0 −3 −2 ····· 0 0 0 M2
−1 −1 0 ····· 0 0 0 M3
0 0 1 ····· 0 0 0 M4
0 3 −1 ····· −1 0 0 Mn-2
0 1 0 ····· 0 −1 0 Mn-1
0 0 1 ····· 0 0 1 Mn
S
ξ
Metabolites
Reactions
C
V1
V1 max
Sv = 0
Steady State
V2
V2 max

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):

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

where $\mathbf{c}^{\mathrm{T}}\mathbf{v}$ 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

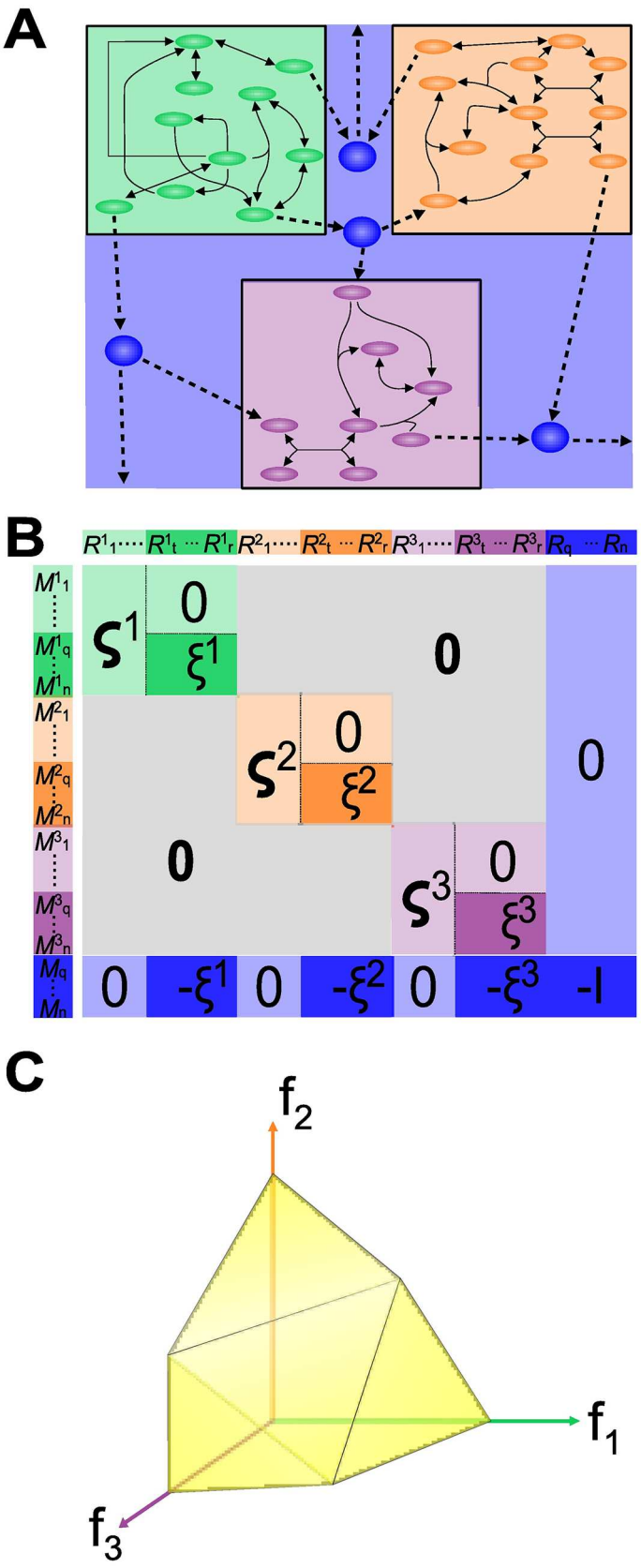

A
B
C
f2
f1
f3

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:

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

$$\text{subject to}$$

$$\mathbf{S}^{\sigma}\mathbf{v} = \mathbf{0}$$

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

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 $\mathbf{x}^w \in \mathcal{X}$ is a **weak efficient** solution if there is no $\mathbf{x}^* \in \mathcal{X}$ 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 $\mathbf{x}^* \in \mathcal{X}$.

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 $\mathbf{x} \in \mathcal{X}$ 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 $\mathbf{x} \in \mathcal{X}$. 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, $\mathcal{Y}_E := \{\mathbf{C}\mathbf{x} \mid \mathbf{x} \text{ is efficient}\}$, referred to as "outer approximation" or "Benson type" algorithms, have been described [48–51]. Generally speaking, these type of algorithms calculate $\mathcal{Y}$ 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

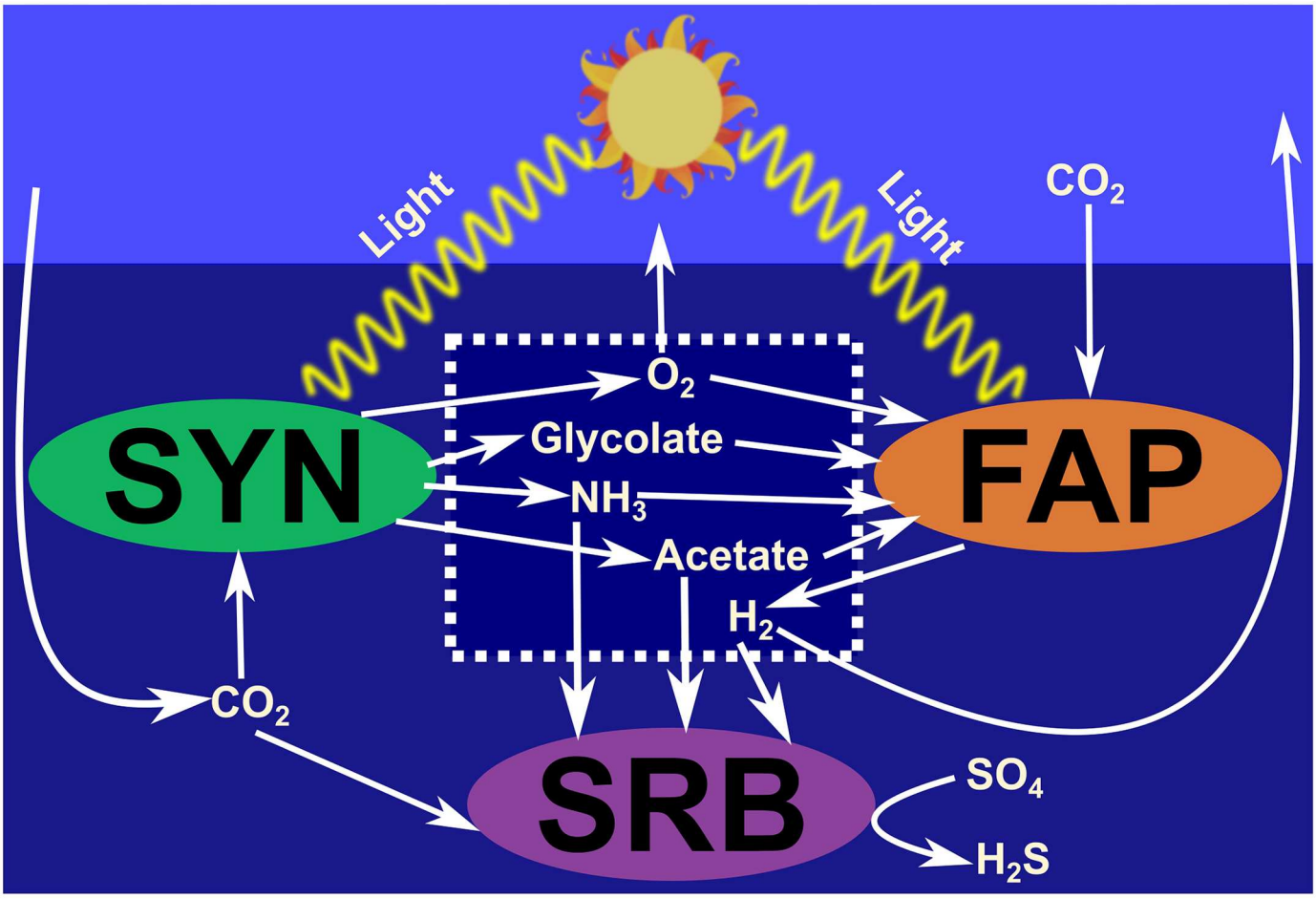


Fig 3. Day Model of the Hot Spring Mat Community. The model comprises three guilds of microorganisms of the SYN, FAP, and SRB phenotypes. Organics acids produced by SYN may be utilized by FAP and SRB. FAP is capable of fixing carbon by anoxygenic photosynthesis. Under anoxygenic fermentation conditions, 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 $\mu_i = \mu_i^0 \pm \mathrm{dg}$, where $\mathrm{dg} = RTln(10^3) \approx 20$ (kJ.mol^{-1}) for T = 75°Celsius. For water and *hv*, concentrations were considered as fixed at 1 M, implying $\mathrm{dg}_{H_2O} = \mathrm{dg}_{hv} = 0$.

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

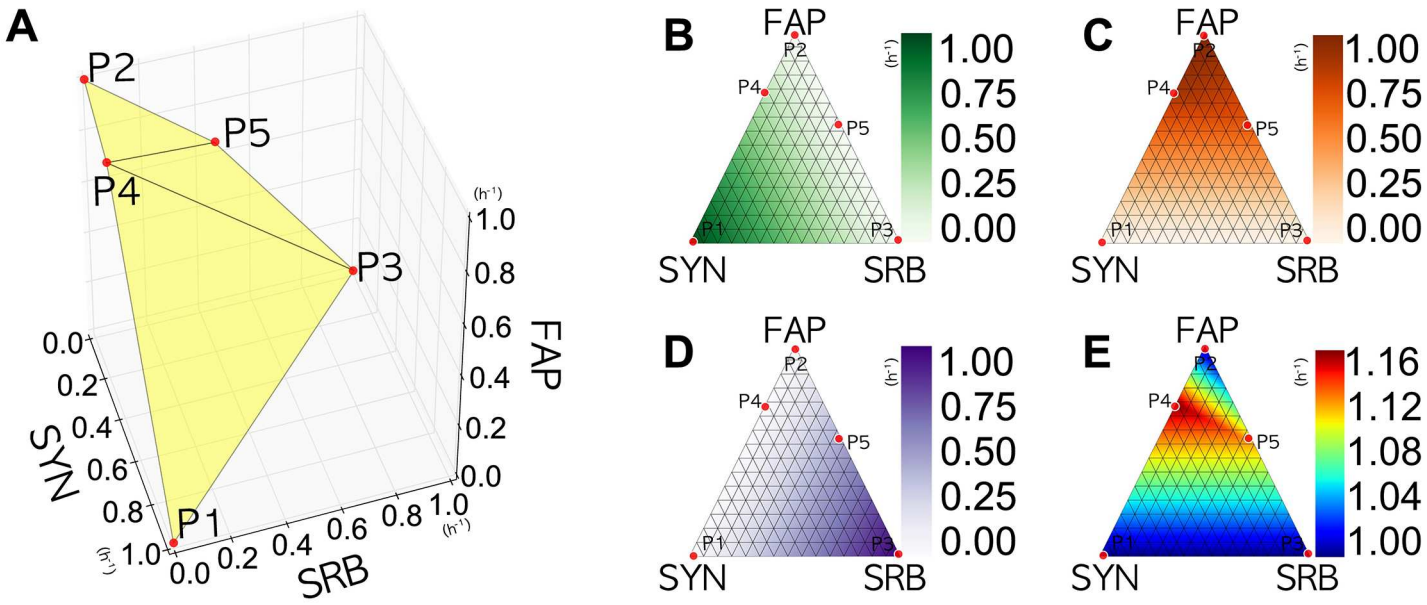


Fig 4. 3D and 2D Projections of Pareto Front. (A) shows a 3D Pareto front, in yellow, describing the maximal growth rates of SYN, FAP, and SRB (in terms of units per hour, h^{-1}), when considered as a system. It is evident that a decrease in the growth rate of one organism results in an increase in 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

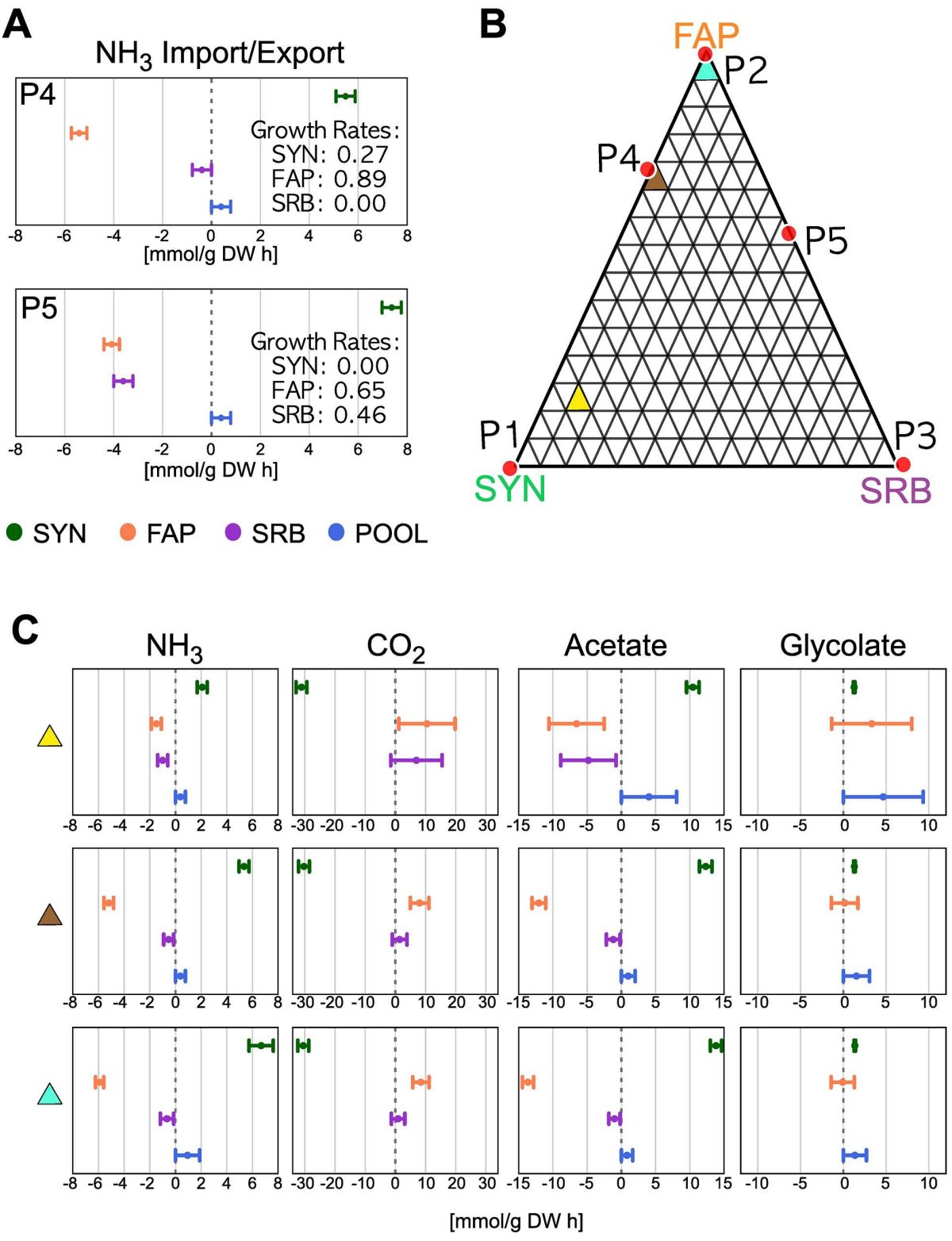


Fig 5. Multi Objective FVA. (A) shows NH_3 maximal and minimal fluxes for SYN, FAP, SRB, and pool compartments (green, yellow, purple, and blue respectively) for extreme points P4 and P5. The export of NH_3 by SYN is correlated with a drop in their growth rate; similarly, increases in NH_3 intake are 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.

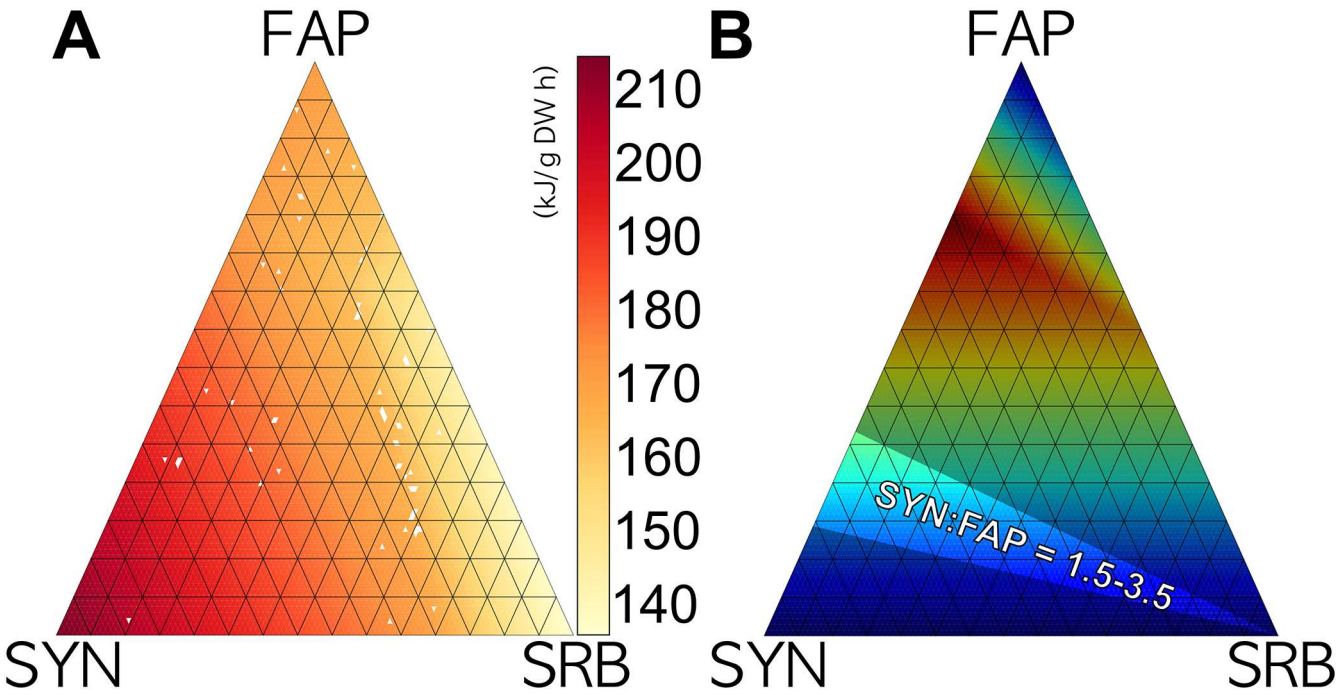


Fig 6. Thermodynamics in the Pareto front. (A) Description of the chemical motive force (kJ.gr^{-1}.DW^{-1}.h^{-1}) for each point of the Pareto front; red regions indicate thermodynamically favored growth rates, while the points where the solver does not reach the optimal criteria are shown in white. 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

4.4 Quantitative modeling of dynamical biological behaviors

As illustrated above, not only contemporary next-generation sequencing (NGS) approaches provide an unprecedented characterization of the diversity of microbial communities, but also provide a significant amount of data that represent one experiment alongside its associated uncertainties about quantitative measurements. Thus, worth noting that increasing the data set provides a concomitant increase in the number of uncertainties that must be considered, uncertainties that are also amplified when one considers time-series. Uncertainties are accounted for in the dynamical modeling process (e.g., via averaging) but are hidden in the model itself. Moreover, despite the tremendous predictive power of such modelings, the resulting models are not necessarily biologically meaningful. In particular, once parameterized, a model could be overfitted to a data set without reflecting emergent properties and precluding or reducing knowledge discovery. Considering that a single microbial system could produce several distinct data sets through several experimental approaches, several different models are built accordingly [185, 179]. All corresponding models must be then investigated via automatic learning and verification techniques to take into account their common properties rather than considering each model in isolation. Within biological models, these uncertainties can be accounted for by machine learning techniques (see Libbrecht and Noble [132] for a review) that produce automatically predictive (deterministic) models from experimental measurements. Despite the challenges mentioned above, uncertainties must be accounted for and integrated into microbial modeling approaches. Such an issue remains a general problem across biology, and even in ecology despite a long tradition of dealing with quantitative uncertainties [153]. In [47], we advocate granting these uncertainties per se, in particular within quantitative biological modelings, by promoting a computational convergence on uncertainties rather than simple simulations.

4.4.1 Modeling the evolution of protein concentrations with a microbial cell-based on genetic activity

Among the techniques that integrate uncertainties, the Bayesian network is a probabilistic graph model that represents the biological compound interactions via a directed acyclic graph [71]. However, the Bayesian network is not able to take into account the feedback loops necessary to represent gene regulatory network dynamics, as well as the accumulation of quantities (e.g., the abundance of microorganisms or concentrations) over time, such as is necessary to depict general biological dynamical behaviors. For this purpose, it would be preferable to use an extension

of Bayesian networks: dynamical Bayesian networks. These dynamic networks consist of the repetition of an elementary Bayesian network, as previously defined, linked together in order to abstract dynamical effects, including feedback loops. Nevertheless, despite being of practical interest, such a combination of networks drastically increases model complexity. Such an extension is not always appropriate to model deterministic behaviors. By proposing Probabilistic Boolean Networks (PBN), [183] overcame this limitation. PBNs combine the expressibility of Boolean networks to describe deterministic dynamical behaviors and uncertainty via the use of probability (see [131] for a complete comparison between PBNs and dynamical Bayesian networks in the context of gene regulatory circuit modeling). Overall, PBN represents a general probabilistic modeling framework that combines deterministic modeling and uncertainties. PBN offers plenty of applications in the context of biological networks, with a strong emphasis on qualitative modelings. Nevertheless, PBN does not permit a more quantitative modeling approach required to depict general biological properties, including the dynamical behavior of biological properties attributable to continuous variables.

For this purpose, we proposed a novel approach called Event Transition Graph (ETG) that combines Boolean modeling and probabilistic approaches but integrates descriptive mechanistic measurements alongside more quantitative knowledge of a given system.

Jérémie Bourdon, Damien Eveillard, and Anne Siegel. Integrating quantitative knowledge into a qualitative gene regulatory network. *PLoS computational biology*, 7(9):e1002157, September 2011

ETG was originally developed to model multi-scale systems, and we used it to determine the impact of *E. coli* gene regulatory networks on intracellular protein concentrations under diverse growth conditions [174]. Unlike traditional biological modeling techniques (e.g., ordinary differential equation approaches where all processes are equivalent), ETG classifies the order of biological events, such as gene transcription, and transitions from one state to another via a set of probabilities such that the succession of states accurately reproduces experimental observations. Such a classification of biological events, being controlled only by probabilities, avoids the need for kinetic parameterization, which is usually unknown for microbial {eco}systems, but rather advocates for the addition of uncertainties to a deterministic schema.

Unlike other probabilistic modelings, ETG takes both a qualitative description and quantitative biological data as inputs. This combination of knowledge makes our probabilistic modeling sensitive to mechanistic descriptions but, on the other hand, drastically increases its computational complexity compared to other state-of-the-art probabilistic modelings. However, contrary to less complex modelings,

the main contribution of ETG does not stand on simulations only, but rather on the reasoning that one performs during the learning step. In particular, such reasoning emphasizes the genes that are the most sensitive to parameter value variations, which makes them candidates for marker genes to monitor physiological response to perturbations.

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Integrating Quantitative Knowledge into a Qualitative Gene Regulatory Network

Jérémie Bourdon[1]*, Damien Eveillard[1], Anne Siegel[2]

1 Computational Biology (ComBi) Group, LINA UMR 6241, Université de Nantes, Ecole des Mines de Nantes & CNRS, Nantes, France, **2** IRISA, Symbiose, INRIA Rennes-Bretagne-Atlantique, Rennes, France

Abstract

Despite recent improvements in molecular techniques, biological knowledge remains incomplete. Any theorizing about living systems is therefore necessarily based on the use of heterogeneous and partial information. Much current research has focused successfully on the qualitative behaviors of macromolecular networks. Nonetheless, it is not capable of taking into account available quantitative information such as time-series protein concentration variations. The present work proposes a probabilistic modeling framework that integrates both kinds of information. Average case analysis methods are used in combination with Markov chains to link qualitative information about transcriptional regulations to quantitative information about protein concentrations. The approach is illustrated by modeling the carbon starvation response in *Escherichia coli*. It accurately predicts the quantitative time-series evolution of several protein concentrations using only knowledge of discrete gene interactions and a small number of quantitative observations on a single protein concentration. From this, the modeling technique also derives a ranking of interactions with respect to their importance during the experiment considered. Such a classification is confirmed by the literature. Therefore, our method is principally novel in that it allows (i) a hybrid model that integrates both qualitative discrete model and quantities to be built, even using a small amount of quantitative information, (ii) new quantitative predictions to be derived, (iii) the robustness and relevance of interactions with respect to phenotypic criteria to be precisely quantified, and (iv) the key features of the model to be extracted that can be used as a guidance to design future experiments.

Citation: Bourdon J, Eveillard D, Siegel A (2011) Integrating Quantitative Knowledge into a Qualitative Gene Regulatory Network. PLoS Comput Biol 7(9): e1002157. doi:10.1371/journal.pcbi.1002157

Editor: Richard Bonneau, New York University, United States of America

Received January 13, 2011; **Accepted** June 28, 2011; **Published** September 15, 2011

Funding: This work is partially supported by the French National Agency for Research (ANR-10-BLANC-0218 BioTempo project) and by PEPS QuantOursins CNRS grant, AtlanSTIC CNRS FR2819, LINA CNRS UMR 6241 collaborative projects and INRIA Rennes-Bretagne-Atlantique. 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.

* E-mail: Jeremie.Bourdon@univ-nantes.fr

Introduction

There have been a number of success stories in macromolecular network modeling during the last decade. Special attention has been paid to dynamical modeling approaches. Among a broad spectrum of strategies, qualitative models and their associated methods have played a central role, allowing modelers to investigate the full space of possible discrete behaviors of several regulatory networks. To that end, a variety of methods for qualitative modeling, analysis and simulation of genetic regulatory networks (GRN) have been proposed since the seminal works of Kauffman [1] and Thomas [2,3] (see [4] for a review). As they rely on discrete representations of both time and variables, these methods share two main advantages: first, the space of possible states is finite (although possibly large), making it possible to hypothesize about the dynamics of biological regulatory systems despite the lack of kinetic information at transcriptional level. Second, regulatory networks can be built from local experimental observations or knowledge-based information (gene-gene or gene-protein interactions).

Although these approaches provide high-level insights into the functioning of gene networks, they often do not accurately reflect the real dynamics of GRN. Indeed, transitions between states in a GRN may exhibit a stochastic component as observed in [5]. This stochastic signal is closely related to population average behaviors [6]. Consequently, the dynamics of GRNs have a stochastic component which is difficult to observe in real time and to capture in discrete models. This has emphasized the need for probabilistic models and methods for analyzing and simulating GRN. Such probabilistic representations of gene networks are now widespread to complement discrete approaches. The Probabilistic Boolean Network (PBN) approach [7,8] is one of these. Due to its flexibility and the fact that it can be inferred directly from data, it has been extensively studied over the last decade. In [9], finite state Markov chains are also proven to be useful in dealing with microarray data. It was established that the automatically reconstructed Markov chain gave rise to steady state distributions in accordance with some phenotypic biological observations. This suggests that Markov chain models are capable of mimicking biological behavior. More generally, Markov chain models are usually applied in the following way. First, a model that fits a given set of data is inferred [10,11]. Then, steady state distributions are computed, giving access to biological information, as they reflect some expected phenotypes [8,12]. In a final step, important product nodes are exhibited, as they control the steady-state distribution and the phenotype [5,13,14]. This latter task gives insights useful in designing new biological experiments, allowing both a better validation of the model and suggesting some therapeutic targets. Although those approaches are very efficient, they mainly rely on the quality of the

Author Summary

Understanding the response of a biological system to a stress is of great interest in biology. This issue is usually tackled by integrating information arising from different experiments into mathematical models. In particular, continuous models take quantitative information into account after a parameter estimation step whereas much recent research has focused on the qualitative behaviors of macromolecular networks. However, both modeling approaches fail to handle the true nature of biological information, including heterogeneity, incompleteness and multi-scale features, as emphasized by recent advances in molecular techniques. The principle novelty of our method lies in the use of probabilities and average-case analysis to overcome this weakness and to fill the gap between qualitative and quantitative models. Our framework is applied to study the response of *Escherichia coli* to a carbon starvation stress. We combine a small amount of quantitative information on protein concentrations with a qualitative model of transcriptional regulations. We derive quantitative predictions about proteins, quantify the robustness and relevance of transcriptional interactions, and automatically extract the key features of the model. The main biological novelty is therefore the presentation of new knowledge derived from the combination of quantitative and qualitative multi-scale information in a single approach.

network reconstruction process, that yields a two sides issue: inferring the ''structure'' of the gene regulatory network and computing transition probabilities that are consistent with the available data. In concrete terms, the lack of accurate observation datasets on the result of transition in a GRN usually makes the inference of the structure more accurate than the computation of the probabilities [5].

In a quite complementary way, [15,16] have proven that adding a probabilistic aspect to already qualitatively validated discrete models may help in determining parameters of the qualitative model. To do so, the authors add a probabilistic dimension to a discrete piecewise affine model. They introduce unknown transition probabilities between two states as the ratio of volumes defined by the qualitative parameters of the system. The main novelty of their approach is that they compute the whole set of transition probability matrices leading to given qualitative attractors of the system, instead of selecting a precise matrix as the above-mentioned approach does. This approach allows them to exhibit relations between transition probabilities and important coefficients of the system such as synthesis rates. However, as they use an analytic description of the set of accurate probability matrices, their method is limited to small networks composed of two or three genes.

In the present work, we advance the idea of studying discrete knowledge-based transcriptional ''intracellular'' regulatory information given by qualitative models within a global probabilistic approach. The main novelty of our approach is that we compute the full set of probability transition matrices that correspond to quantitative ''population scale'' observations provided by protein time-series measurements. We rely on methods inspired by average-case analysis of algorithms theory [17,18], making use of Markov chains coupled with transition costs to study statistical properties of pattern matching issues. We design a probabilistic framework allowing population scale observations to be integrated into a qualitative gene expression network assumed to be shared by several individual cells. Our approach should therefore be considered as a bridge between purely discrete modeling approaches and probabilistic simulations. We introduce three main novel features: first, we rely on a strong asymptotic property of Markov chains to fully describe the set of all possible weighted probabilistic networks matching with protein time-series observations. Second, we overcome computational problems as we drastically reduce the size of the model by focusing on slope changes (switch from a variable increase to a variable decrease, for instance) instead of changes in product levels. Third, we develop numerical methods to incorporate a set of suitable Markov chains – all those matching the numerical observations – rather than a single Markov chain that cannot be uniquely determined from the few quantitative observations we have at hand. These three novelties allow us to increase the robustness of our approach while reducing the set of data required to perform the analysis. Concretely, our approach involves first computing a discrete (non-deterministic) description of possible succession of slope variations. This can be deduced from knowledge-based transcriptional information, *i.e.*, either a logical graph or a qualitative event succession like those observed in novel generations of microarrays [19]. This provides us with a graph of transcriptional event transitions. The transcriptional events, arising on the scale of an individual cell, affect the protein concentrations, observed on a population scale. These two scales are related by adding an *impact cost* for each transition over a given protein concentration. This cost is easily deduced by fixing an arbitrary ''natural'' degradation rate and by applying an equilibrium principle as follows. Intuitively, in the absence of any information – when all the transition probabilities are chosen to be uniform – the expected protein concentrations will be constant. The next step consists of numerically determining the set of transition probability matrices that fit a global quantitative observed outcome. As an example, we expect the model to fit the time-series quantitative observations of the mean concentration of a single protein over a cell population - in this paper we focused on carbon starvation response in *Escherichia coli*. We have combined theoretical properties of Markov chains - inspired by symbolic dynamics - with reverse-engineering methods (local inference methods) to describe the full space of weighted Markov chains having the appropriate topological structure and whose global mean outcome fits the time-series curve. Then we investigate the geometric structure of the space of Markov chains to derive biological properties of the system: we derive a ranking of gene interactions with respect to their importance in achieving the considered protein variations. Such a classification is confirmed by the literature. We also accurately predict the quantitative time-series evolution of several non-observed population-cell protein concentrations using only knowledge of discrete gene interactions and very few quantitative observations on a single protein concentration. According to our modeling framework, variations in protein quantities appear to be driven by the dynamical behaviors, qualitatively described, that occur underneath at the gene regulatory scale.

Method

Main features

As a major modeling contribution, and in the light of the above assumptions, this paper establishes a relationship between the concentration time series (*i.e.*, quantitative knowledge) and the qualitative behaviors of the biological system, as modeled by genetic regulatory networks. To that end, two matrices are considered (see Figure 1). Note herein that an exhaustive illustration of following features is proposed in the end of the Method section. The first matrix describes an *event transition Markov chain* which constitutes the core of the model. It depicts the probabilities (latent variables of the model) that the system will switch from one qualitative ''basic behavior'' to another, where a

qualitative basic behavior means a constant slope for the variation of a product. The structure of the matrix is determined by the current extent of our knowledge of what regulates the system. Its numerical coefficients stand for the mean ratio of trajectories of the system that may cross a given transition. Our reverse engineering method aims at computing these numerical non-zero coefficients. As a companion matrix to this event description, a family of *impact matrices* is built for each protein involved in the system. Given a protein P, the corresponding impact matrix will describe the global outcome of each transition between two events – corresponding to an arrow of the Markov chain – over the concentration of the protein P. By way of example, if we assume that the system goes through a transition that activates the mRNA production of a gene g, the effect (or "impact") of this event may be modeled by an increase in the production rate of the protein G encoded by g, say 20%. Additionally, the effect of this event on all other proteins in the system may be modeled by a decrease in the production rate, a free parameter that we fix to 5%, since they undergo a natural degradation process and are not affected by the event transition. As detailed hereafter, the exact rates that are used are computed so that active and passive degradation have the same average impact during a random process. With these two matrices at hand, average-case analysis properties of Markov chains reveal a relationship between the event transition matrix, the impact matrices and the quantitative evolution of a protein concentration under given stimuli, allowing to establish some relations between observable variables of the model (the observed growth ratio of given proteins) and the latent variables of the model. Roughly, the time-series concentrations of a given protein make it possible to recover the main eigenvalue of the event transition matrix, which can be reformulated to infer times-series concentrations of other proteins, as well as global properties of the system.

Average impact of a Markov chain over an accumulation rule

A *Markov chain* is a random process for which the next state depends on the current state only. It is described by a graph over the set of nodes V, and edges labeled with probabilities in (0,1). Likewise, the random process can be described by a *transition matrix* $T=(T_{u,v})_{(u,v)\in V\times V}$. The Markov chain is described as *minimal* when this matrix is aperiodic and irreducible meaning that for sufficiently large n and all vertices $v\in V$, there exists an n-length cycle including v. A *stationary state* of the Markov chain represents a numerical distribution of the nodes that does not evolve anymore, which corresponds to the eigenvector of the matrix T.

The main goal is to estimate the quantitative asymptotic impact $Q(t)$ of the Markov chain on a biological product quantity or a generic yield. Biologically, such a quantity is any of the phenotypic measurements that is impacted by the gene regulatory network, *i.e.*, any experimental bio-product concentration that might be inferred from either a cell growth rate or a protein concentration encoded by a gene that belongs to the system. To this end, an *impact matrix* $C^{(Q)}$ is linked to the transition matrix T of the Markov chain. The impact matrix is the same size as T. Zero-coefficients in T yield zero-

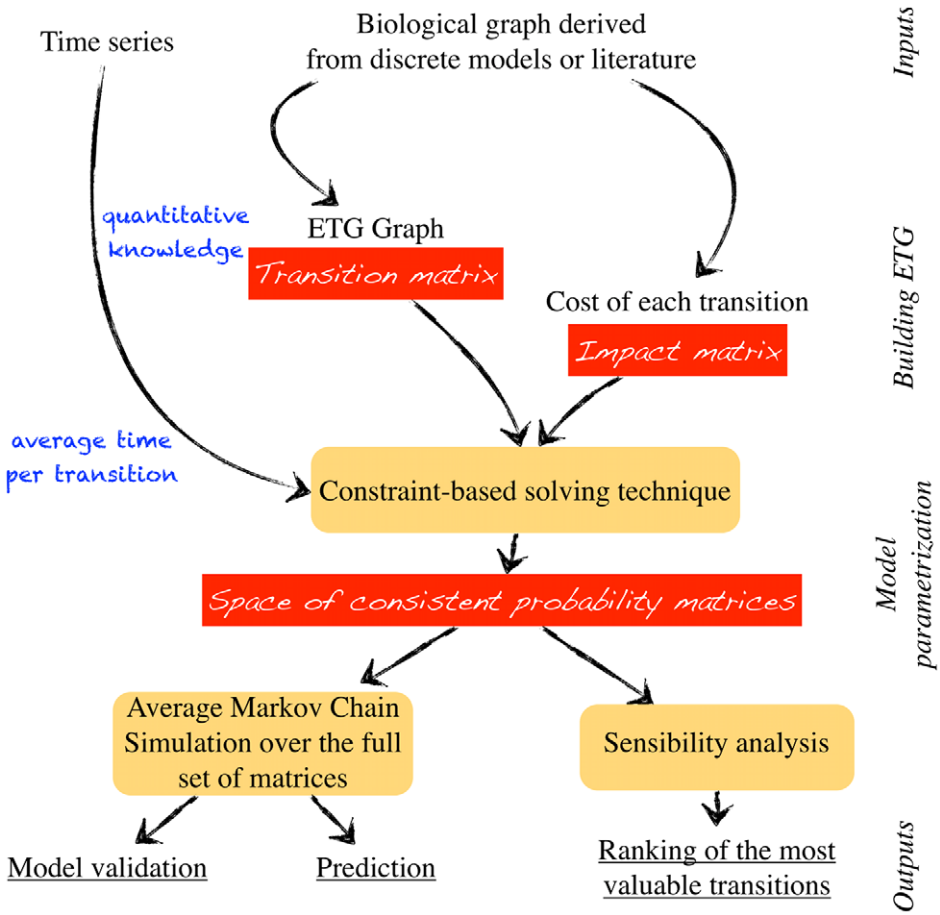


Figure 1. Flowchart of the Event Transition Markov chain modeling protocol.
doi:10.1371/journal.pcbi.1002157.g001

coefficients in the impact matrix. Coefficients of the impact matrix are positive real values that describe the estimated cost of a transition on the change in the phenotypic quantity.

Impact matrices simulate the effect of a Markov process over the global quantity Q as follows. Let A, B be two nodes of the Markov chain connected by an edge $A \rightarrow B$. Let $T_{A,B}$ denotes the probability of this transition and $C^{(Q)}_{A,B}$ its impact. The *elementary cost* of the transition $A \rightarrow B$ over the quantity Q is defined as $T_{A,B}\, C^{(Q)}_{A,B}$. The induced *elementary cost matrix* is denoted by $T_* C^{(Q)}$. The quantity $Q(n)$ is then said to evolve following a *multiplicative accumulation rule from an initial distribution* μ. Its mean value at time n – *that is, after* n *iterations of the Markov process* – (*i.e.*, the average of the costs of all trajectories of length n) is strongly related to powers of elementary cost matrix, that is $Q(n) = (1, \ldots, 1)\,[T_* C^{(Q)}]^n\, \mu$. In other words, to compute the mean value of the quantity at step n, the elementary cost is multiplied along all paths of length n – therefore introducing $[T_* C^{(Q)}]^n$. Each path is weighted with the probability of starting from its initial node – information given by μ. The final impact is given by the sum of all these quantities – therefore multiplying by $(1, \ldots, 1)$. In particular, as detailed below, such a multiplicative accumulation rule is useful to model the burst effect of a gene regulatory network on a metabolic scale, in which a single mRNA stochastically transcribed produces a burst of protein copy numbers [20–23].

When a Markov chain is fully determined and when an impact matrix is given, simple linear algebraic computations allow to calculate the growth rate of the corresponding quantity. The added value of a multiplicative law over a Markov chain relies on its asymptotic behavior, that is proved to be exponential, as stated in Theorem 1. More precisely, a multiplicative accumulation rule follows an explicit log-normal law with explicit mean, variances and estimation of error terms. All these characteristics, such as the growth rate δ of the exponential, are related to dominant eigenvalues of the elementary cost impact matrix $T_* C^{(Q)}$. It should be noted that when the Markov chain reaches a stationary state, the accumulation law itself enters a *permanent regime*, where its exponential rate is fixed. The error term is also exponential, but with a much smaller growth rate, ensuring that the stationary state of the Markov chain is quickly reached.

Theorem 1

(Average case analysis theory for accumulation rules) Let E be a minimal Markov chain with transition matrix T. A multiplicative accumulation rule $Q(t)$ with impact matrix C asymptotically satisfies a $\log -$*normal law with mean and variance*

$$\mathrm{E}[Q(n)] = \beta\, \exp(\delta\, t) + o(\Lambda_1^n) \qquad \mathrm{Var}[Q(n)] = \alpha\, \gamma^n + o(\Lambda_2^n),$$

where e^{δ} is the dominant eigenvalue of the elementary cost matrix $T_* C$. The other quantities express by means of a generation of the elementary cost matrix, $\mathbb{A}(u)$ defined by $\mathbb{A}_{i,j}(u) = T_{i,j} u^{\ln C_{i,j}}$. More precisely, $\gamma = \max(\lambda(e^2), \lambda(e)^2)$ express by means of the dominant eigenvalue $\lambda(u)$ of $\mathbb{A}(u)$, β and α are constants corresponding to the dominant eigenvectors of $\mathbb{A}(e)$ and $\mathbb{A}(e^2)$. There exists $\eta < 1$ such that the error terms Λ_1 and Λ_2 verify $\Lambda_1/\delta \leq \eta$ and $\Lambda_2/\gamma \leq \eta$.

Here, the minimality assumption restricts applications to a biological process such that (*i*) its underlying Markov chain is aperiodic and irreducible; and (ii) for every considered cost matrix, there exists at most one aperiodic trajectory (meaning that the cost evolution is aperiodic through times for this trajectory). Note that in the present work, these assumptions are those that will most restrict the biological referential. For instance, biological systems that display oscillatory behavior are outside the natural range of the approach. Nonetheless, one may overcome this weakness by modeling an input with oscillatory behavior and modeling the steps of the dynamics with independent Markov chains. This modeling device is particularly useful when one aims at modeling the circadian system. For a better illustration, please see below how to build such a Markov chain that describes the behaviors of a gene regulatory network.

Reverse engineering of a transition matrix from impact accumulation rules and growth rates

Given a set of impact rules and assuming that they all follow accumulation rules, optimization techniques were used to infer a Markov chain fitting all available experimental results – the growth rate of several biological quantities. The identification process was divided into two optimization problems. First, in the exact case, a Markov chain is computed which minimizes the euclidean distance between the growth rates δ and β – see Theorem 1 above – of every impact rule associated with the Markov chain and the objective numerical values provided by the experimental results at hand. Local search algorithms are well suited to such an inference task (see [24] for a review). Here, it is necessary to develop an ad-hoc local search algorithm capable of handling eigenvalues that have only an implicit definition.

In order to take experimental errors into account, we considered a second optimization problem, in which the objective values were defined by an interval of validity. Our goal was to infer a Markov chain such that the growth rate of every impact rule belongs to its objective numerical interval, allowing some sets of valid Markov chains to be defined. These sets were approximated by using a polyhedra, defined as follows. First the local search algorithm was used to find a Markov chain whose growth rates were close to the middle of every objective intervals. This Markov chain defines a point, hereafter called the source point in the sequel, inside the solution set. Some points on the boundary of the solution set were then identified by setting a random direction and using a dichotomy method to find the intersection between the boundary and the line, starting from the source point with the expected random direction. As shown in the results section, the volume provides particularly meaningful information. In both cases, sensitivity analysis was performed by considering the following definition. The function $E(T,g)$ was introduced, standing for the Euclidean distance between the growth rate of all impact rules and their objective numerical values. The *sensitivity of a transition* $T_{i,j}$ is then defined by the $E(T,g)$ modification, in percent, when $T_{i,j}$ is modified by 1%. Note that it is closely related to the partial derivative according to variable $T_{i,j}$ of the function $E(T,g)$. The higher is the sensitivity of a transition, the more sensitive is the overall score to small variations of this variable.

Event transition Markov chain associated with a gene regulatory network

The previous theoretical framework can easily be adapted to the biological regulatory networks that display discrete dynamics [25]. Products of the system are gathered in a set P and a relevant Markov chain summarizes the dynamics of the system. In order to handle computational issues of reverse engineering, the focus is on shapes of trajectories instead of graph states, formalized as follows.

The main component of the modeling operation are transcriptomic events, *i.e.*, elements of $P \times \{+, -\}$. They describe the possible slopes in the variation of a bioproduct during a time unit (i.e. increasing or decreasing). For instance, $(fis, +)$, also denoted by fis_+, stands for the increase in the transcriptional activity, or mRNA production, of the gene *fis*. The two events occurring over a product g are denoted by g_+ and g_-. It is sometimes useful to add some supplementary biological events such as a complex

formation, when the information is available. This increases the accuracy of the model. The *Event Transition Graph* (ETG) encodes the possible successions of events. Its nodes are given by the set of events. An event $g_s^{(1)}$ targets $g_t^{(2)}$ if, in at least one trajectory of the system, $g^{(2)}$ varies with the slope s and then $g^{(2)}$'s slope changes to the sign t. This graph may be derived easily from a state transition graph such as those produced by logical asynchronous multivalued Thomas mode piecewise linear models [26].

An *Event transition Markov chain* is an event transition graph endowed with a matrix probability T. Biologically, considering a Markov chain means considering an average behavior of the system over a set of different cells. Since the focus is on events only (i.e. successions of changes in the slope variations of products) instead of states, the stationary states of the Markov chain correspond to cell populations where the proportion of cells with increasing/decreasing transcripts is fixed. Therefore, the stationary states of Markov chains do not correspond to stationary states of the biological system (where all transcripts have a stable concentration). In order to avoid misunderstandings, a stationary state of an event transition Markov chain is called a *permanent regime*.

The *Initial state* of the Event transition Markov chain depends on the biological process that is studied. Assuming that the cells within a population are not synchronized suggests that the initial distribution of events in the system is uniform. If the cells are forced to be synchronized at an early stage of the experiments, a dedicated initial state describing the forced condition must be taken into account.

Multiplicative impact matrix of the Markov chain over the production of each protein

It was pointed out that the evolution of one – or several – protein concentrations resumes a multiplicative phenotypic impact of the gene regulatory network [21,23]. The multiplicative assumption was considered as relevant since the protein concentrations in a single cell follow standard evolution laws which are of exponential nature, similarly to the behaviors of systems governed by multiplicative laws [27]. Let g be a gene in the system at hand and P its encoded protein. The impact matrix $C^{(P)}$ describes the impact of the event transition Markov chain on the protein production. To define this matrix, an *active impact scale* p and a *passive impact scale* d must be introduced. If a given transition impacts a given gene via its mRNA production, we assume that its encoded protein production increases or decreases by the scale p. Otherwise the protein rate is assumed to decrease via its natural degradation by the scale d. Formally, let $\cdot \to g_s$ be an edge in the Markov chain (g can be any product and s is either $+$ or $-$). Reaching state g_s means that the activity of gene g changes leading to an *active production or degradation* of its associated protein P. During all other transitions $\cdot \to h_s$, where h_s does not encode the protein P, the system undergoes a natural degradation of protein P. The production and degradation rate values are chosen as follows. The *passive effect* d is set as equal to 0.95 (*i.e.*, a natural degradation of 5%). The *active degradation coefficient* is defined according to the following equilibrium rule. Let D_- (resp. D_+) be the set of all events associated to an active degradation (resp. production) of the given protein. We first fix all the transitions to be uniform (*i.e.*, all the probabilities of leaving a given state are equal), and denotes by π the steady-state distribution of the associated Markov chain. Protein P concentration is stable if

$$p\,\pi_- + 1/p\,\pi_+ + d\,(1-\pi_- -\pi_+) = 1,$$

where $\pi_- = \sum_{s\in D_-}\pi_s$ and $\pi_+ = \sum_{s\in D_+}\pi_s$. This defines a degree two equation. Simple arguments prove that this equation has only one solution smaller than 1 that is assigned to p. The *active production coefficient* is then defined as $1/p$, the inverse of the active degradation coefficient. Eventually, the impact matrix associated to the protein P is fulfilled thanks to the passive effect rate and the passive and active degradation rates.

Inferring growth rates from protein observations

As the approach is dedicated to prokaryotic systems, a linear relationship between gene activities and their protein concentrations is assumed. This induced a standard evolution law to describe the quantitative evolution of the protein concentrations in the system in accordance with the qualitative events as described by the event transition Markov. More precisely, it was assumed that, as with other modeling studies [23,27], a protein concentration evolves according to a succession of exponential laws $(\varphi_1(t),\ldots,\varphi_k(t))$, with $\varphi_i(t) = B_i exp(D_i(t-t_i)) + C_i$. The cutting points $t_1,\ldots,t_k$ are obtained using the available experimental data. The meaning of this succession is that the protein concentration at time t is $\varphi_i(t)$ if $t\in[t_i,t_{i+1}]$. Then, for each i, δ_i, β_i and γ_i expresses by

$$D_i = \frac{\log[(\varphi_i(t_{i+1})-C_i)/(\varphi_i(t_i)-C_i)]}{t_{i+1}-t_i}, \quad B_i = \varphi_i(t_i) + C_i.$$

It can be noted here that the concentration of a protein that is only degraded tends to C_i, which is its basal concentration. Assuming it to be null leads to simpler formulas for D_i and B_i.

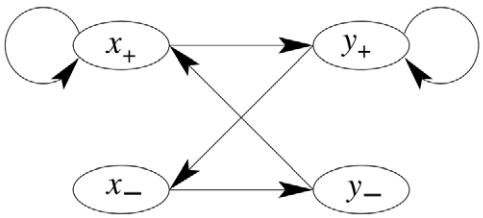
According to the hypotheses discussed below, we assume that the protein concentration φ_i follows a multiplicative accumulation rule Q_i in each time interval $[t_i,t_{i+1}]$. Let τ be the mean duration of a transition. In the permanent regime of φ_i, which is reached very quickly, the relation $Q_i(n) \simeq \varphi_i(n\tau)$ holds. According to Theorem 1, this equation implies that the product τD_i is nothing but the dominant eigenvalue δ_i of the elementary cost matrix of Q_i. Additionally, B_i introduced below equals the constant β_i introduced in Theorem 1.

Taking all into account, the growth rates δ_i and β_i required to apply our reverse-engineering methods described below, can be calculated from the protein concentration shape as soon as the mean duration time τ of a translation has been estimated. To that end, it is assumed that the duration is independent from the studied dynamics, allowing it to be computed from experimental knowledge on passive degradation. We introduce τ_0 the shortest half-life of amino-acids of the protein of interest – usually available in the literature. According to the N-end rule, as depicted in [28], fixing a passive degradation rate of 5% entails that $\tau_0 = \log(0.5)/\log(0.95)\tau$, which allows an explicit computation of τ and completes the inference of growth rates.

Illustration of the method on a two gene network

For the sake of clarity, we propose to illustrate now the modeling method when applied on a simplistic Event Transition Graph (core model). It is composed of two genes that monitor four events as depicted in Figure 2. The graph is also depicted using a transition matrix in which one adds two unknowns (latent variables) for describing a Markov chain: $v_1 = p_{x_+\to x_+}$ and $v_2 = p_{y_+\to y_+}$. To solve the problem in a biological context, one then considers the two following complementary informations:

- *Costs per transition (free parameters)*: Assuming a passive degradation rate (free parameter) of 5% and applying the above equilibrium rule, the active degradation rate for both protein X and protein Y equals 0.882 (−12.8%) while the active



	x_+	y_+	x_-	y_-
x_+	v_1	$1-v_1$	0	0
y_+	0	v_2	$1-v_2$	0
x_-	0	0	0	1
y_-	1	0	0	0

Figure 2. Event Transition Graph composed of 2 genes (left) and its corresponding probability transition matrix (right), that includes two unknowns v_1 and v_2.
doi:10.1371/journal.pcbi.1002157.g002

production rate equals 1.134 (+13.4%). See Supplementary Text S1 for the matricial description. Here we assume that the time unit is one iteration of the Markov chain. In some more general cases, the definition of time units is tricker as mentioned above.

- *Fictive experimental knowledge (observable variables)*: For illustration and as tutorial, one considers that the protein X relative quantity or concentration, is multiplied by 100 in 100 iterations or time units (*i.e.*, two measures points are thus (1,1) and (100,100), which defines an asymptotic growth rate equals to $\exp(\log(100/1)/100) = 1.0471$).

These informations are then used to infer v_1 and v_2 and relative probabilities. The inference task is performed by an adhoc MATLAB script (The complete package and its corresponding tutorial are available in http://pogg.genouest.org). As a general result, several combinations of probabilities satisfy the given constraints. They are depicted in Figure 3. Emphasizing a unique set of probabilities is therefore not possible. Unlike other Markov-like techniques, the Event Transition Markov chain models the impact of the Markov chain behaviors over the production of each protein of the system. We are thus able, for each combination of probabilities that satisfies the constraints, to estimate the protein growth rates in the permanent regime. Indeed, one can describe the distribution of Y protein growth rates for 10,000 probability combinations that satisfy the constraints (Figure 4(A)). This distribution is obviously sensitive to the probabilities. For illustration, the distribution of the protein Y growth rate for 10 000 probability combinations picked randomly is different, as attested when one depicts the difference of random and constrained distributions of Y protein growth rates in Figure 4(B), illustrating the close relations between protein X and Y concentration evolutions. Computing the distribution is not an easy task when one considers more than 3 genes or 6 events. In practice, we then overcome this problem by estimating the mean of each growth rate (*i.e.*, 1.0152 (prediction) in the case of the Y protein growth rate as presented above), instead of each growth rate distribution. This provides some accurate predictions of protein concentration evolutions.

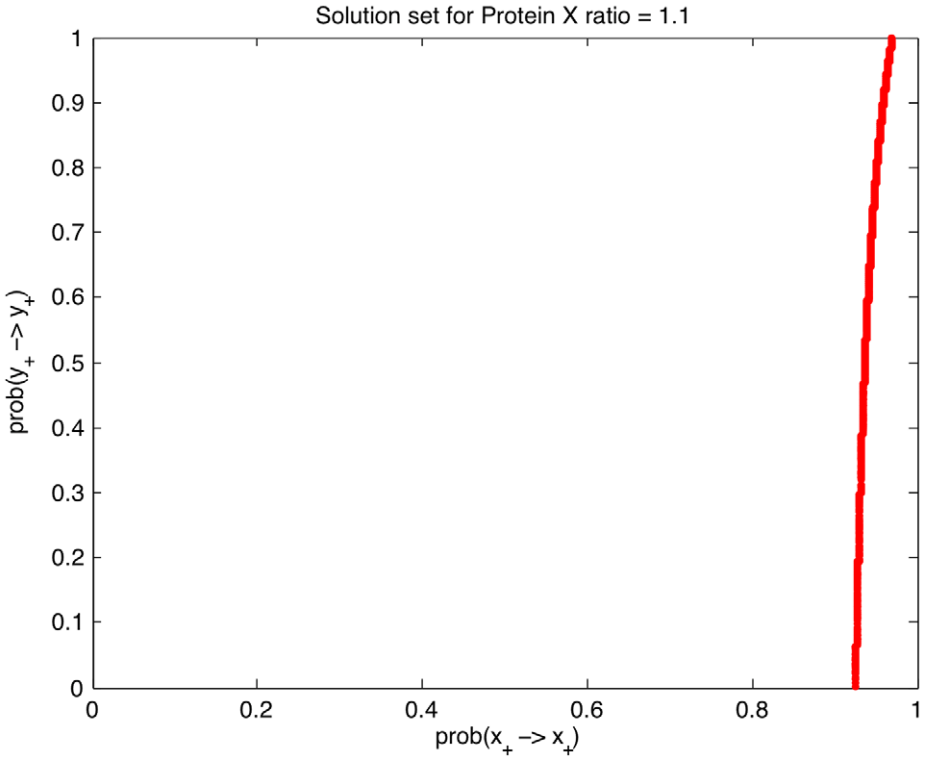


Figure 3. Set of probabilities that satisfy the constraints for the Event Transition Graph depicted in Figure 2.
doi:10.1371/journal.pcbi.1002157.g003

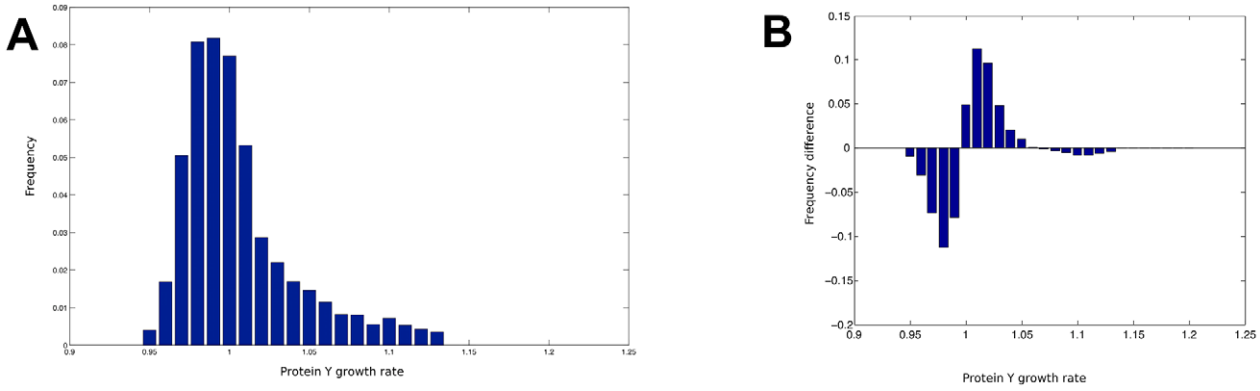


Figure 4. Comparison of the protein Y growth ratio in two different situations. (A) Distribution of the Y protein growth rate estimated from probabilities randomly picked; (B) Difference of the distribution described in (A), and the distribution of protein Y growth rate estimated from 10 000 combinations of probabilities that satisfies the constraints of the ETG model that depict the interactions of two genes.
doi:10.1371/journal.pcbi.1002157.g004

Results

To illustrate the accuracy of the use of Event Transition Markov chains in a biological context, we propose now to focus the Event Transition Markov chain approach on predicting the behavior of protein concentrations during a period of bacterial stress. D. Ropers and collaborators model the growth phase transition of *Escherichia coli* after a period of nutritional stress [29]. In particular, their model shows the move from an exponential growth state to stationary growth during a carbon starvation stage. This elegant *"switch"* is evidenced at gene regulatory level with implications at phenotypic level. This model is based on the qualitative results available in both the literature and gene regulatory experiments as performed by the authors (see Figure 5). Furthermore, the proteins encoded by the genes that interact within the model have been well researched by independent studies [30,31]. This provides partial quantitative information that may be introduced into the qualitative model.

Event transition graph

The original model [29] is given as a system of piecewise affine differential equations. It contains 6 genes and 37 constraints over inequalities and thresholds. This yields a state transition graph of 912 qualitative domains. The corresponding Event Transition Graph was automatically computed by applying the definition introduced in the method section and detailed in Supplementary Text S2. The resulting graph, composed of 22 edges and 11 nodes, is depicted in Figure 6. Note that for the sake of clarity, we manually introduced a component named "complex" that summarizes the effect of cAMP metabolite as depicted in [32]. This node, in accordance to the original model [29], stands for a complexation of the Crp and Cya proteins and the carbon starvation signal. Following our formalization, this component is thus a natural product of cya_+, crp_+ and the signal component. Although the event transition graph roughly summarizes the behaviors of the original qualitative model, it still highlights the major biological properties by its reading. For illustration, the repression of the *crp* gene by the Fis protein [33] is depicted by an active effect of fis_+ on crp_-. However, the information about *crp* controlled by two distinct promoters is lost.

Event transition Markov chain: Impact and transition matrices

As detailed above in the method section, we computed the impact matrices based on bacterial protein production growth rates. This setting appears to be suitable since *E. coli* can be viewed as a multi-scale system. Indeed, the change in protein concentration shall be considered as a protein scale amplification of events that occurs at the transcriptomic scale that are depicted as protein burst by experiments [20–22]. By way of illustration and following the equilibrium rule defined above, in the impact matrix over the Fis protein, the concentration of Fis, denoted by q_{Fis}, undergoes a 46% increase for each transition targeting fis_+. It suffers from a 32% decrease for all transitions targeting fis_-. Finally, it goes through a 5% decrease for all other transitions, reflecting a natural degradation for Fis (see Supplementary Text S2 for a complete

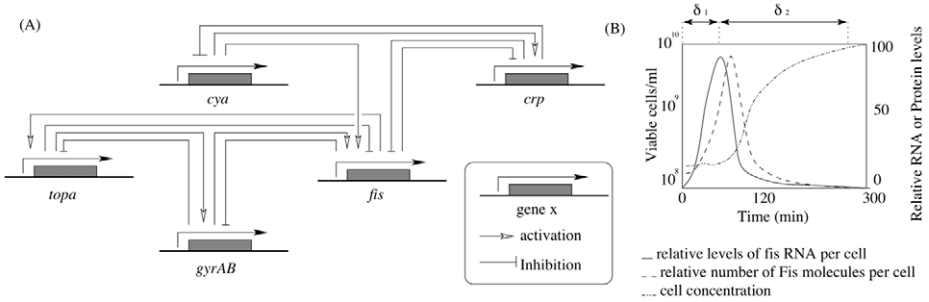


Figure 5. Biological information concerning *Escherichia coli* carbon starvation system. (A) represents interactions between genes involved in the regulatory network (adapted from [29]). (B) shows quantitative variations of macromolecules of interest (based on [30]). Note the linear relationship between *fis* mRNA and Fis protein productions that allows to infer protein product behaviors based on the gene regulatory network.
doi:10.1371/journal.pcbi.1002157.g005

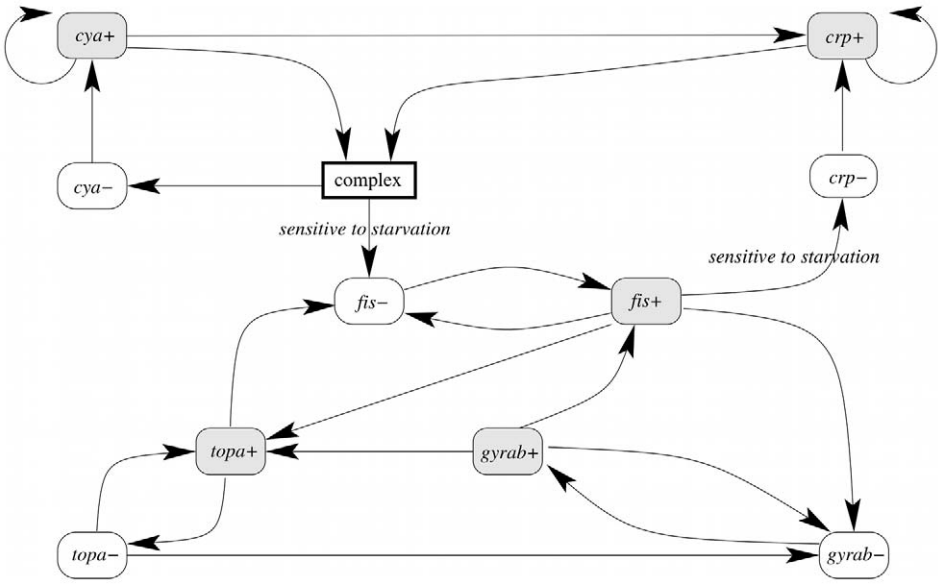


Figure 6. Even transition graph of the genes regulatory network of carbon starvation response in E. coli. Each component represents an active event that concerns a gene product (x), either its increase (x_+) or its decrease (x_-). Arrows between events depict the active effect of one event on another. Two transitions are absent when the system is subject to carbon starvation.
doi:10.1371/journal.pcbi.1002157.g006

description of the impact matrix). This depicts the Event Transition Markov chain.

We used quantitative information about changes in Fis protein concentration to reverse-engineer the transition matrix. Experimental evidence [30] shows that the Fis concentration multiplies by 10 in 80 minutes, during the stationary growth phase (i.e. carbon starvation conditions) and then decreases in the exponential phase (see Figure 7 and Supplementary Text S2 for details). Therefore, the protein concentration curve was approximated by two successive steps φ_1 (stationary phase, from $t_1=2min$ with $\varphi_1(t_1)=10$ until $t_2=80min$ with $\varphi_1(t_2)=100$) and φ_2 (exponential phase, from $t_2=80min$ with $\varphi_2(t_2)=100$ until $t_2=130min$ with $\varphi_2(t_3)=10$). The shortest half-life of amino-acids of the protein of interest is estimated as $\tau_0=2min$ by the literature [28], leading to a mean transition duration of $\tau=0.148min$. Applying our inference growth rate procedure – see method section – resulted in the computation of the growth rates for both the accumulation rules corresponding to the stationary phase ($B_1=10$, $D_1=0.0295$, *i.e.*, $\varphi_1(t)=10\exp(0.0295(t-2))$) and the exponential phase ($B_2=100$, $D_2=-0.0461$, *i.e.*, $\varphi_2(t)=100\exp(-0.0461(t-80))$). Then, the

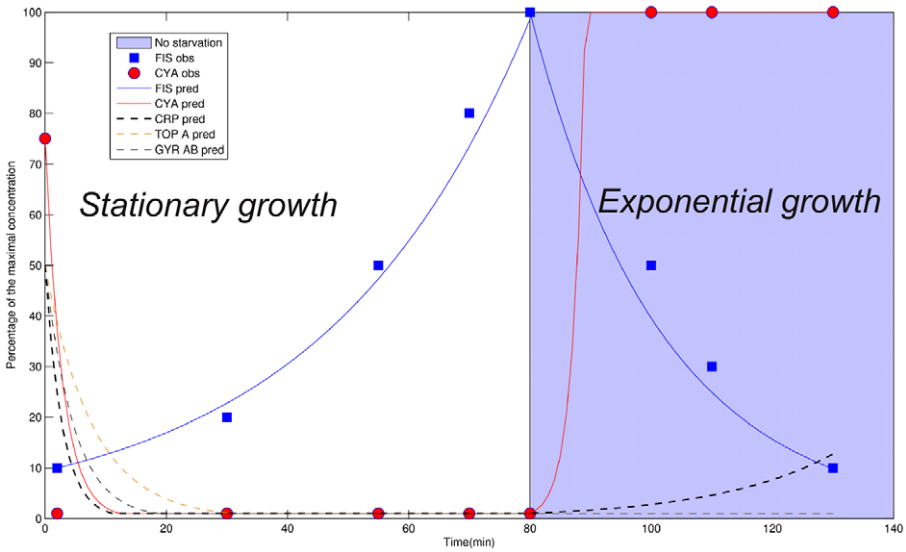


Figure 7. Simulations of changes in bacterial protein concentration during both stationary and exponential growth phases. The corresponding probability matrix is estimated in the stationary growth condition based on three experimental data for the protein Fis. After 80 minutes, the signal of carbon starvation manually switches from 1 to 0, emphasizing a switch from starvation to non-starvation conditions, which leads respectively to a stationary and an exponential growth phase of the bacterial population. Experimental data are marked with dashed lines, whereas computation results are depicted using plain lines for the five proteins of interest (Fis, Cya, Topa, GyrAB and Crp).
doi:10.1371/journal.pcbi.1002157.g007

reverse-engineering approach using $\beta_1 = 10$, $\delta_1 = 0.0044$, $\beta_2 = 100$, $\delta_2 = -0.0068$ (see Method section) produced a probability transition matrix T that fits the protein growth rates in both stationary and exponential growth phases. By repeating several times this procedure, one obtains a sampling of the set of all probability matrices that fits the given experimental protein growth rates.

Asymptotic behavior of the system

Using the transition matrix of the Event Transition Markov chain, we perform several simulations on protein concentrations, as impacted by the gene regulation network. First, the transition matrix was coupled with impact matrices on proteins Fis and Cya to simulate their permanent regimes during the stationary phase. Then, after 80 minutes, it is assumed that the exponential phase is initiated, inducing a change in the structure of the gene regulatory network. This change takes place by adding 2 transitions from the "signal" box on the Event transition Markov Chain which activates crp_+ and the "complex" compound. Because of the given initial conditions during the exponential growth phase, these transitions were neglected, but not in stationary phase conditions. Then, based on the same matrices (impact and probability transition), new simulations are performed on the evolution of Fis and Cya concentrations. Figure 7 depicts the predicted variations of the Cya and Fis proteins during both phases.

Compared to the available independent experimental results [30,31], the simulations and experiments are overall significantly similar according to a Pearson correlation test. The transition matrix allows us to compute the quantitative behavior of Cya in both stationary and exponential phases. Based on sparse information about Fis only, the predicted Cya behavior is consistent with the experimentally observed behavior ($R^2 = 0.9599$, $p\text{-}value = 10^{-5}$) [31], which is a quantitative validation of our model. Notice herein that we also predict the complete time series of Fis ($R^2 = 0.9937$, $p\text{-}value = 6.5\ 10^{-8}$), which confirms the exponential growth rate assumption. As a complementary result, the system remains for only a short time in the transient regime (*i.e.*, the error made herein when one computes the mean is significantly lower than 1% after 7 minutes, or 20 iterations of the Markov Chain), which backs up our assumption of studying this microbial system in permanent regime in both growth conditions. This confirms the usefulness of our modeling approach for this specific biological system.

Automatic classification of key gene interactions

In addition to the prediction feature, properties of the Markov chain provide insights into biological system behavior. According to the inference process, the proteins Cya and Crp have the same predicted behavior, as *a posteriori* confirmed by [34]. Furthermore, the sensitivities associated with the transitions of the Markov chain also represent an appreciation of the impact of a given biological compound. In particular, this demonstrates that, in stationary growth phase, $fis_+ \rightarrow crp_-$ transition is highly constrained. Interestingly, this transition implicitly monitors the CAMP-CRP complex that controls the metabolism of alternative carbon sources [33]. It is closely related to ability to the bacterial system to switch between both growth phases in function of the carbon starvation. Furthermore, Schneider and co-workers [35] suggest that *fis* is involved in a fine tuning of the homeostatic control of DNA supercoiling. A small change in the supercoiling drastically affects the expression of the gene *fis*, which is in total agreement with the constraints extracted from the Event Transition Markov chain. We performed a similar analysis over the whole system (*i.e.*, in both stationary and exponential growth conditions). The most sensitive transitions are reported in Table 1, in which we detail the biological meanings of such interactions. Not surprisingly, *fis* regulation is one of the corner stone genes of the system, but it might be a natural consequence of the inferring process in our modeling approach. However, with no specific transition matrix inference, *gyrAB* also emerges as one of the most, if not the most, important gene of the microbial system. Implicitly, this confirms the usefulness of the DNA topology for *E. coli* under carbon starvation conditions.

Discussion

Our purpose was to illustrate the strength of coupling Markov models together with accumulation rules to study the dynamics of a gene regulatory network, by focusing on its effects at a larger scale – the quantitative protein scale. We assumed that the production of a protein by a gene that belong to a regulatory network, follows a multiplicative accumulation rule. This implies that a permanent distribution of the protein system will be reached in a very short time. In such a regime, each protein concentration follows an exponential dynamic. The permanent regime may be modified by external events, inducing a short transition to another permanent regime. This paper details why observing such a permanent distribution – possibly several – at the protein level allows us to recover the main probabilistic law that governs the gene regulatory network. The law is thus described by a Markov chain over the succession of transitions at the transcriptomic scale. Very general properties of this Markov chain – average case analysis (see Theorem 1) – allow us to infer the Markov chain from a variety of heterogeneous information, such as qualitative behaviors based on existing models and partial quantitative data. We proposed an efficient algorithm based on this average case analysis to infer the Markov chain. In this method, it must be emphasized that the fundamental interest is to focus on transitions between biological events (slope variations of products during a time unit) instead of

Table 1. Summary of the most important transitions of the system according to their sensibility measure.

Transition in ETG	Sensitivity	Biological significance	Ref.
$fis_+ \rightarrow crp_-$	15.5%	control of CAMP-CRP complex	[33]
$gyrab_+ \rightarrow fis_+$	11.6%	*fis* regulation controlled by the DNA supercoiling level	[37]
$gyrab_+ \rightarrow topa_+$	8.1%	Topoisomerase I regulation by the DNA supercoiling	[38]
$fis_+ \rightarrow topa_+$	7.1%	Homeostatic control of DNA topology	[35,39]
$fis_+ \rightarrow gyrab_-$	5.5%	Homeostatic control of DNA topology	[35,39]
$gyrab_+ \rightarrow gyrab_-$	4.8%	*gyrAB* expression regulation by the DNA supercoiling	[35]

doi:10.1371/journal.pcbi.1002157.t001

state variation as proposed by other state-of-the-art methods. Indeed, this abstraction of the system is required to reduce the size of the Markov chain in order to achieve the inference process.

Having determined this Markov chain allows us to study the main asymptotic properties of the dynamic system: identifying the main transitions implied in the permanent regime and sorting the relevance of transition patterns. All these predictions may be quite easily checked with additional experimentation. Conversely, experimentation allows refinement of the Markov chain inference process. Taken together, mixing the properties of a Markov chain with accumulation rules, provides a tool to determine the quantitative and asymptotic properties of a dynamic system.

For illustration and validation purposes, we computed a Markov chain for the event transitions of the *Escherichia coli* system in the carbon starvation. The computations were performed by using a gene regulatory network of this process and quantitative data about protein Fis production during the stationary phase. Our predictions of the behavior of Fis during the exponential phase and of Cya protein changes were confirmed by independent experimental observations, which emphasizes the ability of our approach to spread partial quantitative information through an Event Markov chain built from qualitative models. Moreover, our results produce various emerging properties such as (i) the sensitivity of a specific transition within the Markov chain or (ii) the quantitative prediction of gene products that are not directly optimized during the simulation. All these features reinforce our interpretation of the global quantitative behaviors of the natural system as modeled.

From a technical viewpoint, the main interest of this approach is as follows: it is not necessary to build quantitative differential dynamic systems that need accurate and complex parameter estimations. Our method uses the results of several available observations to recover the main characteristics of the dynamics (its exponential ratio) and to export several dynamic and biological features. Such probabilistic-like reasoning shall be considered as complementary to formal verification techniques used for validating the qualitative properties of a system [29].

Other recent methods also use probabilistic techniques for studying gene regulatory networks [7,9,36]. However, their main purpose is to embed a deterministic model with probabilities. Their main analyses therefore focus on estimating impacts of variation. Probability matrices are computed to represent experiments accurately. Finally, transition probability matrices are used to compute permanent distributions. We argue that our approach is complementary since our average case analysis theory allows us to emphasize emerging properties of the system. Relations between the two scales of observations allow us to exhibit constraints between the gene regulatory network and protein observations. Eventually, this process elucidates transition probabilities that did not come to light with other available methods.

A weakness of our approach relies on the fact that the Markov Chain inference process is based on knowledge of a full qualitative gene regulatory network [4]. This shortens the range of application of our method since, nowadays, relatively few biological systems are described at this level of abstraction. However, this flaw will be moderated by the fact that the gene regulatory network is used only in order to build a global frame of the event transition Markov chain, which is much more abstracted and smaller that the gene regulatory dynamics description. It is reinforced by our main approach which is to build the Markov chain automatically from biological assumptions – either from the literature or experiments such as microarrays.

Another weakness lies in the assumption of a linear relationship between gene activity and the production of the corresponding protein (relevant for a microbial system only). To avoid such a restriction, one must build novel accumulation rules based on other biological abstractions – metabolic and environmental phenotypes are the most natural candidates here. Extending the construction of event transition Markov chain to the models containing reactions instead of qualitative regulations – for instance, signaling networks – is also under study to extend the range of application of our approach. A final range of future works relies on extracting more precise properties from the Markov chain description of a given dynamic system. Such studies shall initially focus on the interpretation of the concentration joint law, standing as a correlation coefficient between time-series observations. They will also investigate the use of these Markov chains to isolate experimental noise from the noise inherent to the chaotic properties of the system. This would provide an estimation of measurement qualities. Finally, average case analysis can be performed on a class of probabilistic models that is much larger than Markov chains. This would allow us to deal with Markov chains that may handle slight variations over the course of times, eventually studying the adaptation of the model behaviors under given environmental variations.

Supporting Information

Text S1 Application of the method on a simple two genes example.
(PDF)

Text S2 A complete description of *Escherichia coli* model, information used during the inference task and results.
(PDF)

Acknowledgments

The authors wish to thank Alejandro Maas for its constructive comments.

Author Contributions

Analyzed the data: JB DE. Contributed reagents/materials/analysis tools: JB DE AS. Wrote the paper: JB DE AS. Designed the software used in analysis: JB DE.

References

1. Kauffman SA (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol 22: 437–467.
2. Thomas R (1991) Regulatory networks seen as asynchronous automata: A logical description. J Theor Biol 153: 1–23.
3. Thomas R, Thieffry D, Kaufman M (1995) Dynamical behaviour of biological regulatory networks–i. biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull Math Biol 57: 247–76.
4. de Jong H (2002) Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol 9: 67–103.
5. Chen PCY, Chen JW (2007) A Markovian approach to the control of genetic regulatory networks. Biosystems 90: 535–545.
6. Paulsson J (2004) Summing up the noise in gene networks. Nature 427: 415–418.
7. Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18: 261–74.
8. Shmulevich I, Gluhovsky I, Hashimoto RF, Dougherty ER, Zhang W (2003) Steady-state analysis of genetic regulatory networks modelled by probabilistic boolean networks. Comp Funct Genom 4: 601–608.
9. Kim S, Li H, Dougherty ER, Cao N, Chen Y, et al. (2003) Can Markov Chain Models Mimic Biological Regulation? J Biol Syst 10: 337–357.
10. Marshall S, Yu L, Xiao Y, Dougherty ER (2007) Inference of a probabilistic Boolean network from a single observed temporal sequence. EURASIP J Bioinf Syst Biol: 32454 p.
11. Dougherty E (2006) Design of probabilistic Boolean networks under the requirement of contextual data consistency. IEEE T Signal Proces 54: 3603–3613.

12. Zhang SQ, Ching WK, Ng MK, Akutsu T (2007) Simulation study in Probabilistic Boolean Network models for genetic regulatory networks. Int J Data Min Bioin 1: 217–240.
13. Datta A, Choudhary A, Bittner ML, Dougherty ER (2004) External control in Markovian genetic regulatory networks: the imperfect information case. Bioinformatics 20: 924–930.
14. Li H, Zhan M (2006) Systematic intervention of transcription for identifying network response to disease and cellular phenotypes. Bioinformatics 22: 96–102.
15. Chaves M, Gouzé JL (2010) Piecewise Affine Models of Regulatory Genetic Networks: Review and Probabilistic Interpretation. In: Lévine J, Müllhaupt P, eds. Advances in the Theory of Control, Signals and Systems with Physical Modeling, vol 407. pp 241–253.
16. Chaves M, Farcot E, Gouzé JL (2010) Transition probabilities for piecewise affine models of genetic networks. In: Proc. Int. Symp. Mathematical Theory of Networks and Systems (MTNS 10), Budapest, Hungary, Jul. 2010. pp 1–8.
17. Flajolet P, Sedgewick R (2009) Analytic Combinatorics. New York, NY, USA: Cambridge University Press.
18. Bourdon J, Vallée B (2006) Pattern matching statistics on correlated sources. In: Proc. of LATIN'06. Number 3887 in Lect. Notes Comput. Sci.. pp 224–237.
19. Gibson G (2008) The environmental contribution to gene expression profiles. Nat Rev Genet 9: 575–81.
20. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. Nature 440: 358–362.
21. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533–538.
22. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. Science 311: 1600–1603.
23. Elgart V, Jia T, Fenley AT, Kulkarni R (2011) Connecting protein and mRNA burst distributions for stochastic models of gene expression. Phys Biol 8: 046001.
24. Hoos HH, Stützle T (2004) Stochastic Local Search: Foundations & Applications (The Morgan Kaufmann Series in Artificial Intelligence). Morgan Kaufmann, 1 edition.
25. Balleza E, Alvarez-Buylla ER, Chaos A, Kauffman S, Shmulevich I, et al. (2008) Critical dynamics in genetic regulatory networks: examples from four kingdoms. PLoS One 3: e2456.
26. de Jong H, Gouzé JL, Hernandez C, Page M, Sari T, et al. (2004) Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull Math Biol 66: 301–40.
27. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. Science 309: 2075–8.
28. Varshavsky A (1997) The N-end rule pathway of protein degradation. Genes Cells 2: 13–28.
29. Ropers D, de Jong H, Page M, Schneider D, Geiselmann J (2006) Qualitative simulation of the carbon starvation response in *Escherichia coli*. Biosystems 84: 124–52.
30. Ball CA, Osuna R, Ferguson KC, Johnson RC (1992) Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. J Bacteriol 174: 8043–56.
31. Notley-McRobb L, Death A, Ferenci T (1997) The relationship between external glucose concentration and cAMP levels inside *Escherichia coli*: implications for models of phosphotransferasemediated regulation of adenylate cyclase. Microbiology+ 143: 1909–1918.
32. Harman JG (2001) Allosteric regulation of the cAMP receptor protein. Biochim Biophys Acta 1547: 1–17.
33. González-Gil G, Kahmann R, Muskhelishvili G (1998) Regulation of crp transcription by oscillation between distinct nucleoprotein complexes. EMBO J 17: 2877–85.
34. Ishizuka H, Hanamura A, Inada T, Aiba H (1994) Mechanism of the down-regulation of cAMP receptor protein by glucose in *Escherichia coli*: role of autoregulation of the crp gene. EMBO J 13: 3077–82.
35. Schneider R, Travers A, Muskhelishvili G (2000) The expression of the *Escherichia coli* fis gene is strongly dependent on the superhelical density of DNA. Mol Microbiol 38: 167–75.
36. Zhou X, Wang X, Pal R, Ivanov I, Bittner M, et al. (2004) A bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. Bioinformatics 20: 2918–2927.
37. Snoep JL, van der Weijden CC, Andersen HW, Westerhoff HV, Jensen PR (2002) DNA supercoiling in *Escherichia coli* is under tight and subtle homeostatic control, involving gene-expression and metabolic regulation of both topoisomerase I and DNA gyrase. Eur J Biochem 269: 1662–9.
38. Weinstein-Fischer D, Elgrably-Weiss M, Altuvia S (2000) *Escherichia coli* response to hydrogen peroxide: a role for DNA supercoiling, topoisomerase I and Fis. Mol Microbiol 35: 1413–20.
39. Travers A, Schneider R, Muskhelishvili G (2001) DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. Biochimie 83: 213–7.

4.4.2 Probabilistic modeling of microbial networks for integrating partial quantitative knowledge within the nitrogen cycle

Microbial systems ecology tackles complex ecological questions by coupling observational (*e.g.*, molecular and geochemical) data with new computational techniques [165, 117, 218]. We saw above that recent computational advances allowed to qualitatively describe microbial communities by emphasizing *"who is there and who is not"* [166]. However, as for other biological systems, among the most significant challenges in microbial systems ecology is the ability to quantitatively predict microbial community composition and function, by combining molecular data and quantitative physicochemical data. Theoretically, this challenge necessitates the consideration of both measurements (*e.g.*, community composition, or associated geochemistry) alongside an uncertainty analysis associated with these measurements. However, such a coupling is still elusive in predictive modeling (see [153, 47] for review, or [128] for a similar question in the broad context of Computer Sciences). Previous applications in ecology (*e.g.*, [98]; [143]), promote the use of advanced computational approaches to integrate statistical analysis into a mechanistic modeling framework, but both concepts of determinism and randomness are still usually considered as independent [3]. Following the previous probabilistic modeling, we advocate herein that one can transpose the same modeling framework to ecological questions, such as the modeling of biogeochemical cycles carrying out by myriads of microbial metabolisms. In other words, required inputs for ETG modeling are (i) the chronological descriptions of biological events (*i.e.*, metabolic reactions) and their potential connections (*e.g.*, auxotrophy), and (ii) a quantitative behavior to reproduce (*e.g.*, the trajectory of functional groups under fluctuating environmental conditions, or time series of quantities as presented in Figure 4.3). As a result, ETG will learn parameters from quantity variations while considering uncertainties. In this purpose, ETG weighs the transitions between discrete events by probabilities that reproduce, on average, the quantitative behaviors observed in nature. The main insights gleaned from this approach can bring further understanding and prediction of the temporal succession of community assemblages [75, 22]. In particular, this approach can relate key microbial functional guilds to changes in the metabolites consumed or produced across gradients in co-occurring and interacting environmental variables.

For the sake of application and consistency with studies shown in Section 2.3, we focus here on the nitrogen cycle. Beyond the intrinsic importance of nitrogen for biological systems, its cycling results from versatile redox chemical reactions. Combined, these reactions promote complex biogeochemical transformations and structure microbial communities. From a modeling viewpoint, the nitrogen cycle presents three features that make it a promising candidate for new quantitative mod-

elings. First, and despite recent studies uncovering new reactions and pathways [121], nitrogen metabolic pathways are well understood and therefore constitute a metabolic map that provides a stable and mechanistic description of the biological processes involved [104]. This map represents a set of biological events that can be quantitatively described. Second, because of recent technological advances, especially in biogeochemistry and isotopic studies, the main processes involved in nitrogen transformation (e.g., nitrogen fixation, nitrification, denitrification) can also be depicted through quantitative rate measurements, which provide an overall ecosystem behavior. These rates are ETG goals to be reproduced by the trained model. Finally, high-throughput sequencing technologies provide greater insight into the ecology of the microbial functional guilds playing an essential role in the nitrogen cycle, in particular, the organisms responsible for different redox reactions and their putative interactions (see [101] for an illustration).

Event Transition Graph modeling: data and biological knowledge formatting

ETG requires expert biological knowledge to be formalized as a graph. Experimental knowledge will then be incorporated into the model via a learning procedure that weights the edges of this graph.

Network or graph of interactions The first input into ETG modeling is a list of biological events as well as the consequences of these events. For the sake of illustration, when representing the nitrogen cycle, the events are reactions (e.g., nitrification, denitrification), and their consequences are the respective production and consumption of metabolites (e.g., NH_4^+ NO_3^-). This knowledge is necessary to estimate the *"cost"*, or effect when one event occurs over another. Here we derive a nitrogen network composed of a hypothetical series of reactions (*i.e.*, fixation, nitrification, denitrification, and anammox), as laid out in the KEGG database [104], without assigning taxonomy to the microorganisms that mediate these reactions. After removing duplicated reactions, this set of reactions, called sequential biological events, consists of 14 reactions (see Supp. Material for technical details and required format).

Concomitantly, as an additional modeling input, interactions between events take the form of a graph that links reactions (*i.e.* nodes of the graph) when the product of one reaction becomes the substrate for another reaction (directed edge). Thus, the above 14 reactions result in a graph of 14 nodes and 32 edges, as illustrated in Figure. 4.3A. Notice herein that KEGG IDs for *amoA* and *nir* have been replaced below by their reaction names for the sake of clarity, whereas the central reaction points towards reactions linked to the nitrogen cycle but involved in other metabolic pathways, such as carbon or phosphate.

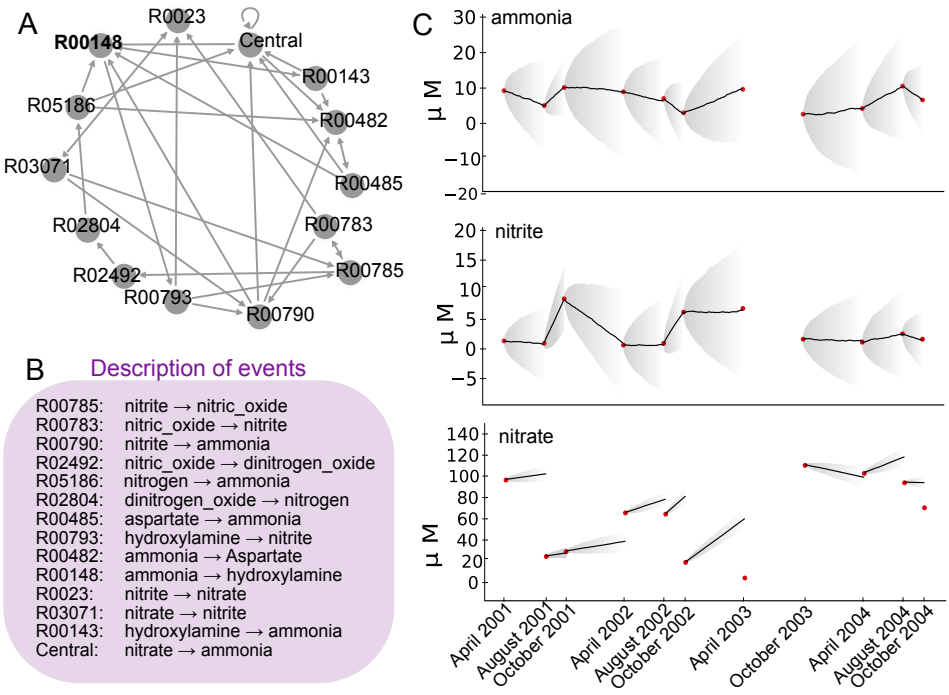


Figure 4.3: Network of the nitrogen cycle and its probabilistic simulation. A represents the nitrogen cycle where nodes are reactions as described in KEGG, and edges putative transitions between reactions when a product of a reaction is a substrate of another. B depicts eleven time point nutrient concentrations as described in [22] station CB100 in Chesapeake Bay, as well as a probabilistic simulation of the ETG model trained on ammonia and nitrite concentrations between 2001 and 2004.

Initial costs In addition to the overall definition of an event (*i.e.*, reactions and product/substrate definition) and description of the interactions within events (through the construction of a graph), the cost of considering one event over another must also be defined. Each event consumes and produces compounds. For instance, each reaction within the nitrogen cycle can be described by its stoichiometry (*i.e.*, -1 for a metabolite consumption and + 1 for a metabolite production). However, when randomly crossed, the graph could promote an artificial increase or decrease of a given compound, solely due to the graph topology. Such a result would not represent a correct output of the modeling approach, but rather a prospective flaw. To avoid this, one must compute the cost (denoted initial cost) for all compounds

Table 4.1: Dissolved inorganic nitrogen concentrations (μM) over the time-course of dataset from sampling station CB100 surface as presented in [22]

Time course Samples	Ammonia (μM)	Nitrite (μM)	Nitrate (μM)
April 2001	8.4	0.8	88.7
August 2001	4.2	0.4	19.9
October 2001	9.3	7.9	24.2
April 2002	8.1	0.1	59.1
August 2002	6.2	0.4	11
October 2002	2.2	5.7	19.3
April 2003	6.3	0.5	76.8
October 2003	1.8	1.1	101.9
April 2004	3.4	0.6	94.7
August 2004	9.7	2	86.2
October 2004	5.8	1.1	63.7

for each event. This cost is necessary to maintain every compound at a stationary amount when every transition is equiprobable (*i.e.*, steady states). For each compound, this initial cost will be assigned to events that do not mention them explicitly. For instance, for all reactions that do not consider ammonia, nitrite, or nitrate as metabolites, one must compute a cost for these metabolites. Thus, -1.5, -1.00, and -0.25 are the costs related to these metabolites (resp. ammonia, nitrite, or nitrate) when not explicitly mentioned in their stoichiometry.

Quantitative data or training dataset ETG modeling estimates probabilities associated with interactions between events (herein reactions) such that the succession of events reproduces quantitative experimental data. For illustration, we use physicochemical variables from [22], which describes a time series of ammonia, nitrite, and nitrate (see Table 4.1). In order to fit such quantitative experimental data with ETG, one must transform quantitative variations as rates, which necessitates the assignment of a time-step. For instance, when considering a time-step of two hours, a variation from 8.4 μM to 4.2 μM of ammonia between April and August 2001 requires 1476 time-steps (123 days x 12), representing an overall variation rate of:

$$\mathrm{rate}_{\mathrm{NH_3}} = \frac{4.2 - 8.4}{1476} \approx -0.0028455 \tag{4.1}$$

Experimental variation in rates for each season (from April to August, from August to October and from October to April) for the years 2001, 2002, 2003, and 2004, and for each nutrient was thus estimated from Table 4.1. These rates are the

training data and represent the quantitative variations that must be reproduced by the probabilistic modeling once parameterized.

Probability estimation and probabilistic simulations Once the ETG model considers (i) a set of events and their putative interactions (Sec. 4.4.2); (ii) a cost for each event (Sec. 4.4.2) and (iii) a quantitative rate that depicts an experimentally observed quantitative variation impacted by at least one event (Sec. 4.4.2), one seeks to learn probabilities to prioritize interactions between events. The overall parameterized model will reproduce variations in compounds (e.g., reactions or products) similar to the experiments. An optimization process, detailed in [20], will compute sets of probabilities for all transitions between each sample within a time series. Thus, ETG of nitrogen cycle sequential biological events will compute probabilities that reproduce the quantitative variation in ammonia and nitrite over four years. It is important to notice herein that searching for optimal probability values is performed by a local search method. Local search methods are sensitive to sub-optimal solutions. Despite the use of a metaheuristic (*i.e.,* Tabu search [85]) that memorizes visited solutions, finding the best solution is complex (NP-hard), which could be prejudicial for larger complex models. However, from a practical viewpoint, models with 15 nodes and 30 edges remain realistic on a personal computer.

Along with probability estimates for transitions between each event, a sensitivity score (S), expressed in percentage, was also computed. The S score associated with a transition expresses the fact that the Euclidean distance between the expected rates (goals from section 4.4.2) and their predictions is modified by S % when its probability value is changed by 1%. Such a sensitivity score permits ranking the transitions according to their respective sensitivities (*i.e.*, a high sensitivity transition implies higher constraints on its corresponding probability value). In practice, sensitivities between two-time points depict in Figure 4.4B are the mean sensitivities of 100 optimal probability estimations that reproduce ammonia and experimental nitrate variations.

Evaluation of the Probabilistic modeling

Probability estimate for simulating the nitrogen cycle Ammonia oxidizing organisms (AOO) mediate the rate-limiting step of nitrification (*i.e.*, $NH_3 \rightarrow NO_2$), a rate-limiting step in the Nitrogen cycle [210, 21, 22]. Herein, we describe AOO integration within a unique ETG of quantitative physicochemical variables and a simulated biological network representing the whole nitrogen cycle. Following an automatic extraction from the KEGG database [104], ETG that covers the whole set of reactions associated with the nitrogen pathways represents 41 nodes and 67

edges. In the present case, for the sake of clarity, the graph is pruned to 14 nodes and 32 edges (Fig. 4.3A). The ETG describes transitions across biochemical pathways with each reaction, or event for the sake of generalization, affecting downstream processes (*e.g.*, each event may produce or consume a compound according to a stoichiometrically balanced reaction equation). In the present case, one substrate can be consumed by several other reactions, which results in multiple edges per node. The cost of each event is then parameterized in order to maintain stable concentrations for each product when transitions of the network are equiprobable (*i.e.*, the null assumption).

To estimate probabilities between reactions and train the ETG, we used an existing environmental dataset representing variations in Chesapeake Bay ammonia, nitrite, and nitrate concentrations (μM) between 2001 and 2004 [22]. The optimization process emphasized a set of probabilities that reproduce observed variations in ammonia and nitrate despite the use of a reduced graph. To test our model, we simulated variations in physicochemical factors using the Gillespie algorithm [84] parameterized with computed probabilities, and compared the predictions with the available time-series data (see Fig.4.3.B). The model accurately replicates ammonia and nitrite physicochemical variables over the period between 2002-2004 but fails to reproduce the observed nitrate dynamics. This point indicates the need for further modeling extensions that could integrate recently discovered new reactions or pathways [121], especially to integrate nitrate concentration variations. It is worth noticing also that no set of probabilities were able to replicate appropriately variations of concentration between April 2003 and October 2003, indicating the sensitivity of our probabilistic modeling to either the time-step or natural perturbations. Indeed, this inability to simulate this particular time slot could be related to the hurricane Isabel, the strongest hurricane in the Atlantic in 2003, that hit the Chesapeake Bay just before sampling. Such a strong perturbation modified the AOO assemblage [22], which could affect, as well, the succession of metabolic reactions compared to normal conditions.

Beyond the probabilistic simulations, the analysis of probabilities between reactions (*i.e.* likelihoods of transitions between two reactions) are of interest. Figure.4.4.A shows the log ratio of computed probabilities over probabilities under the equiprobability assumption, for each transition over the period. Over the four years, some transitions between reactions show similar probability values, or close, to the values corresponding to the equiprobability assumption (*i.e.*, light grey in Figure.4.4.A). Herein, the graph topology remains the main factor to explain the use of these transitions. However, other reaction transitions show probability values very divergent than those obtained under the equiprobability assumption. Transitions depicted in white are underused, whereas those colored in darker grey are overused compared to an equiprobable use of transitions. Among the overused

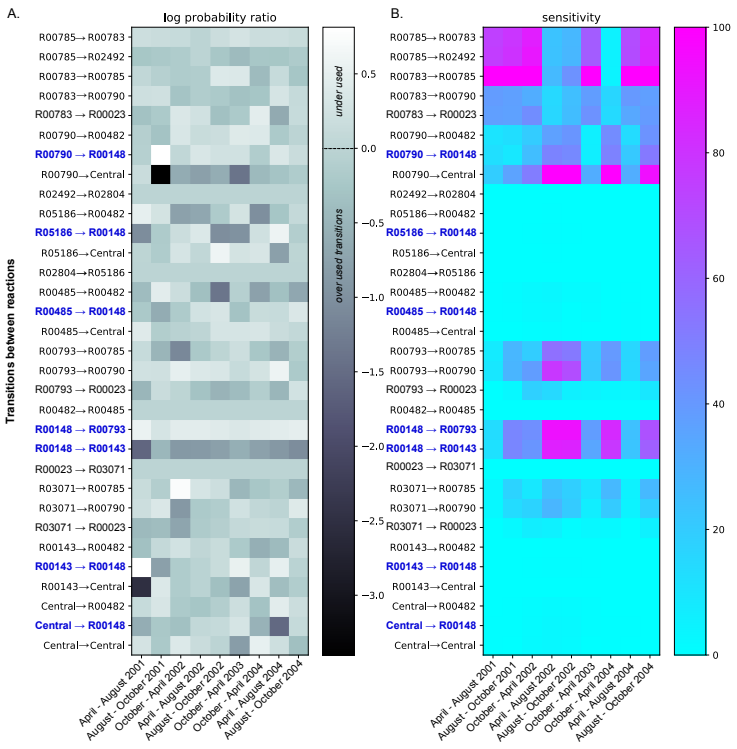


Figure 4.4: Summary of the ETG model Probabilities and Sensibilities trained on ammonia and nitrite concentrations. Panel A shows the log ratio of computed probabilities over probabilities of each transition under the equiprobability assumption. Transitions illustrated in light grey show probabilities in the equiprobability assumption. The dark grey color represents transitions with probabilities lower than those computed under the equiprobability assumption, whereas lighter colors are transitions with higher probabilities. Panel B gives sensitivity values for each transition. Cyan transitions are not sensitive, whereas purple transitions are the most sensitive, i.e., the probability values cannot change without altering the overall predictive accuracy.

reaction transitions, some transitions show substantial variabilities of probability values over the four years, whereas others are overused continuously. In particular, the transition between ammonia-monooxygenase (*amo*) and the hydroxylamine

oxidoreductase reaction (R00143) is of great interest. The transition from *amo* to hydroxylamine oxidoreductase reaction is essential and used continuously over the four years. This relationship is intuitive as the two functions are necessary together for full oxidation of ammonium to regenerate electrons and meet the energetic requirement of the organism to fix CO_2 to biomass. Alternatively, the probability analysis shows a spike of the overuse of the transition from hydroxylamine oxidoreductase reaction to *amo* during Spring 2002, which could alter the above behavior during this period.

Figure 4.4B shows the sensitivity analysis of the model by emphasizing the most constrained transitions; *i.e.* transitions for which the probability values cannot change without altering the training efficiency. These transitions are the most constrained when the system must replicate the quantitative variations used during the training process. Logically, identification of the most sensitive transitions extracts the transition toward *amo* and *nir* nodes, as these events are necessary to mediate NH_3 and NO_2 transformations that are required to reproduce training conditions in Fig. 4.3B. From a biological viewpoint, this result confirms as well the Chesapeake microbial ecosystem as driven by ammonia-oxidizing bacteria. High sensitivities of transitions via R00783, R00785, and R00790, on the other hand, are consequences of the modeling. More importantly, the sensitivity analysis emphasizes seasonal patterns of sensitivities that concern the above-mentioned reactions. Despite the heterogeneous nature of the physical environmental conditions (see Figure **??**B and Table 4.1), the constraints on the transitions are seasonal, unlike most of the observational data, and moreover show an antagonistic pattern between two sets of transitions, with KEGG genes R00790, R00783 and R00785 on one side and *amo*, Central, R00793 and R00143 on another. Overall, the sensitivity analysis emphasizes two antagonistic subsystems that could be resumed as two sets of biochemical reactions. The first subsystem that is the most constrained between April and August implies:

$$\begin{cases} \text{Nitrite} + \text{Reduced azurin} + H^+ \longrightarrow \text{Nitric oxide} + H_2O \\ \qquad\qquad\qquad\qquad\qquad\quad +\text{Oxidized azurin} \\ \text{Nitric oxide} + H_2O + \text{Ferricytochrome C} \longrightarrow \text{Nitrite} + +H^+ \\ \qquad\qquad\qquad\qquad\qquad\quad \text{Ferrocytochrome C} \\ 2 \quad \text{Nitric oxide} + \text{NADH} + H^+ \longrightarrow NAD^+ + H_2O \\ \qquad\qquad\qquad\qquad\qquad\quad +\text{Nitrous oxide} \end{cases}$$

Whereas, another subsystem, driven by ammonia, is very constrained between August and April concerns:

$$\begin{cases} NH_3 + 2H_2O + 6 \quad \text{Oxidized ferredoxin} & \longleftrightarrow & \text{Nitrite} + 6H^+ \\ & & +6 \quad \text{Reduced ferredoxin} \\ \text{Nitrite} + H_2O & \longrightarrow & \text{Hydroxylamine} \\ & & +O_2 \\ \text{Nitric oxide} + H_2O + \text{Oxidized azurin} & \longrightarrow & \text{Nitrite} + H^+ \\ & & +\text{Reduced azurin} \\ NH_3 + NAD^+ + H_2O & \longrightarrow & \text{Hydroxylamine} \\ & & +NADH + H^+ \end{cases}$$

It is worth noting that sensitive transitions and corresponding subsystems may indicate potential constraints (or biochemical trade-offs) on organisms mediating the targeted reactions, which might be related to selective pressures at an ecological level. These pressures occur seasonally and are the results of a mechanistic modeling that represents interactions between reactions.

Learning on a random network The general criticism about probabilistic models concerns their use as a statistical protocol that over-fits observed data. Contrary to other probabilistic modelings, ETG considers a mechanistic interpretation of the systems via the use of a graph of events. The use of a description of events avoids overfitting the data. For the sake of illustration, we propose to randomize the model and to train it on the same dataset. The randomized model consists of building a graph similar to the nitrogen cycle graph for which all edges have been shuffled by permutation. The randomized model is then similar to the ETG nitrogen cycle model in terms of numbers of nodes and edges. We then applied a similar modeling and training procedure to that described above. As pictured in Figure 4.5, the randomized model is unable to predict the seasonal variabilities in ammonia or nitrite. Indeed, no simulations permitted accurate depiction of the ammonia and nitrites experimental data. Furthermore, nitrate quantities remain constant over time, which means that the trained model could not predict changes in nitrate.

Interest of ETG for modeling microbial ecosystems

This ETG framework is ideal for investigating the dynamic and transient nature of microbial ecosystems. It does not begin with an assumption of a community at a steady-state, unlike Flux Balance Analysis techniques (see [158] for review, for metabolic modeling of microbial ecosystems see [224, 28]). This framework is advantageous because, (i) measurements of microbial communities are unlikely to be made at equilibrium, and (ii) most studies focus on changes in communities under an environmental forcing, which is itself a transient behavior. This modeling

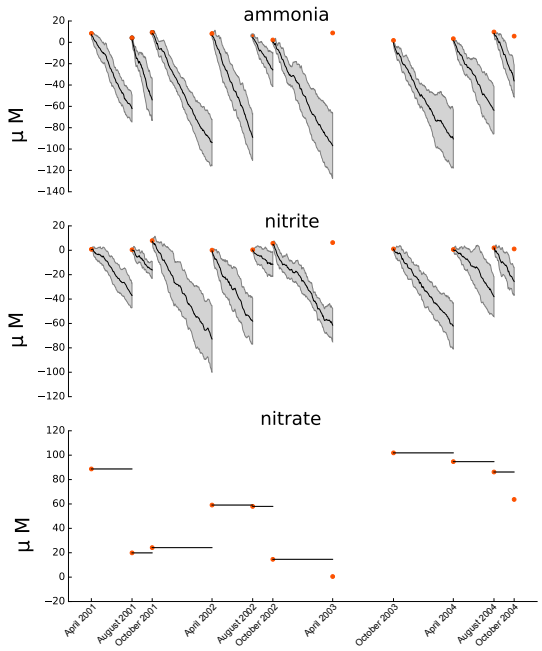


Figure 4.5: Summary of the random ETG model Probabilities and Sensibilities trained on ammonia and nitrites concentrations.

approach, therefore, emphasizes the putative constraints that are applied to a microbial community by inferring the biochemical constraints on metabolic reactions under fluctuating environmental conditions. This dynamical property that emerges from the probabilistic model analysis should be of interest to decipher metabolites of interest from metabolite maps that depict exchanges between microbial strains [19, 18]. In particular, these constraints (*i.e.* sensitivities) could highlight seasonal patterns that could be further compared to co-occurrence patterns [39] for the sake of validation. This study considered a metabolic network as a qualitative description, but other qualitative networks, such as co-occurrence networks [157, 62, 74, 90] or gene associations [35], will be analyzed similarly shortly to refine biogeochemistry models.

Contributions to biological systems modelings

- Damien Eveillard, Nicholas J Bouskill, Damien Vintache, Julien Gras, Bess B Ward, and Jérémie Bourdon. Probabilistic Modeling of Microbial Metabolic Networks for Integrating Partial Quantitative Knowledge Within the Nitrogen Cycle. *Frontiers in microbiology*, 9:395, January 2019

- Domenico D'Alelio, Damien Eveillard, Victoria J Coles, Luigi Caputi, Maurizio Ribera d'Alcalà, and Daniele Iudicone. Modelling the complexity of plankton communities exploiting omics potential: From present challenges to an integrative pipeline. *Current Opinion in Systems Biology*, 13:68–74, February 2019

- Benoît Delahaye, Damien Eveillard, and Nicholas Bouskill. On the Power of Uncertainties in Microbial System Modeling: No Need To Hide Them Anymore. *mSystems*, 2(6), November 2017

- Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE*, 12(2):e0171744, 2017

- Alix Mas, Shahrad Jamshidi, Yvan Lagadeuc, Damien Eveillard, and Philippe Vandenkoornhuyse. Beyond the Black Queen Hypothesis. *The ISME Journal*, 10(9):2085–2091, September 2016

- Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. OPINION PAPER Evolutionary Constraint-Based Formulation Requires New Bi-level Solving Techniques. In *Computational Methods in Systems Biology*, pages 279–281, Nantes, September 2015. Springer International Publishing

- Nicolas Mouquet, Yvan Lagadeuc, Vincent Devictor, Luc Doyen, Anne Duputié, Damien Eveillard, Denis Faure, Eric Garnier, Olivier Gimenez, Philippe Huneman, Franck Jabot, Philippe Jarne, Dominique Joly, Romain Julliard, Sonia Kefi, Gael J Kergoat, Sandra Lavorel, Line Le Gall, Laurence Meslin, Serge Morand, Xavier Morin, Hélène Morlon, Gilles Pinay, Roger Pradel, Frank M Schurr, Wilfried Thuiller, and Michel Loreau. REVIEW: Predictive ecology in a changing world. *Journal of Applied Ecology*, 52(5):1293–1310, July 2015

- Shahrad Jamshidi, Jocelyn E Behm, Damien Eveillard, E Toby Kiers, and Philippe Vandenkoornhuyse. Using hybrid automata modelling to study phenotypic plasticity and allocation strategies in the plant mycorrhizal mutualism. *Ecological Modelling*, 311:11–19, September 2015

- Etienne Z Gnimpieba, Damien Eveillard, Jean-Louis Guéant, and Abalo Chango. Using logic programming for modeling the one-carbon metabolism network to study the impact of folate deficiency on methylation processes. *Molecular bioSystems*, 7(8):2508–2521, August 2011

- A Goldsztejn, O Mullier, Damien Eveillard, and H Hosobe. Including Ordinary Differential Equations Based Constraints in the Standard CP Framework. *Principles and Practice of Constraint Programming, CP2010*, LNCS 6308:221–235, 2010

- Jamil Ahmad, Jérémie Bourdon, Damien Eveillard, Jonathan Fromentin, Olivier Roux, and Christine Sinoquet. Temporal constraints of a gene regulatory network: Refining a qualitative simulation. *BioSystems*, 98(3):149–159, December 2009

- Jonathan Fromentin, Damien Eveillard, and Olivier Roux. Hybrid Modeling of Gene Regulatory Networks: Mixing Temporal and Qualitative Biological Properties. Technical report, BioXiv, December 2008

Chapter 5

Perspectives

I will continue modeling until they don't want me anymore basically because I do love it very much

Claudia Schiffer

In the last decade, we have rooted our work in modeling biological systems and developing computational techniques for the sake of their understandings. From this line of work, scientific interests are dual. From the biological viewpoint, the interest is apparent: gaining biological knowledge from large heterogeneous and incomplete datasets. However, we would like to emphasize here also the interest of this exercise from the computational side. By investigating biological systems, contrary to standard computational studies, one must model the system by choosing the most accurate abstraction with needs dedicated mostly from the application and not necessarily in line with computational customs.

Nevertheless, one must also navigate from one abstraction to another for integrating complex biological knowledge but also to solve biological questions of a single system that occur at different scales. Such plasticity in the choice of the abstraction is a characteristic of biological sciences. On the other hand, this modeling freedom is a significant advantage of Computer Sciences (i.e., switching from one abstraction to another for the sake of automatic solving of a given problem).

From the Computer Sciences viewpoint, developing formal and robust reasonings that must handle such plasticity of abstractions is of particular interest. Previous studies already mentioned the interest of formal computer sciences to tackle the biological complexity, i.e., so-called biocomplexity. Among others, one could notice the seminal works of Alan Turing [203], Stuart Kauffman [110], or more recently an excellent overview by Eric Karsenti [108]. For the sake of modesty, one

can not resume these works in a few lines here. However, one could notice that the use of formal methods enables considering two biological features that are difficult to reach from other scientific fields: (i) Computer Sciences abstract the multiscale organization of Life, for instance by considering different layers of organizations (i.e., from molecules to cellular behavior); (ii) Computer Sciences is accurate to model interactions of elements that are not necessarily quantifiable via the general rules of the information theory (i.e., genes are not quantifiable whereas their protein products are). If studying biological systems via formal methods remains exciting, one advocates herein that studying an ecosystem falls into the same range of interest but amplified by orders of magnitude of scales and the number of elements that one must examine (i.e., from molecules to ecosystems). This change of magnitude calls for a change of the Computer Sciences paradigm to foster ecological questions. In the following, we will decipher this paradigm change into four different sections, that are of distinct scientific interest.

For the sake of illustration, this perspective proposes to root this change of paradigm into the study of plankton. In our opinion, this biological system exhibits several computational interests. While considered as microbial to small organisms, plankton drive marine food webs and global biogeochemical cycles, which are by definition occurring at the enormous scale of Earth. They are thus among the most challenging complex systems to model. For decades, several studies formalized this biological systems via ordinary differential equations (see [170] or [156] for illustrations), which makes a good candidate for further modelings. In particular, systems ecology studies on plankton might actively profit from omics-based systems biology, which would add another dimension to traditional modeling. However, better and adaptive integration with data acquisition and data- analysis is required to achieve the goal. A feasible pipeline will integrate traditional and omics observations, over several time scales, via high-level computational approaches. Such research perspectives will lead to promoting computer sciences for a deeper understanding of our planet and to the provision of knowledge-based directions to stakeholders. The following opinion study describes parts of these research perspectives:

Domenico D'Alelio, Damien Eveillard, Victoria J Coles, Luigi Caputi, Maurizio Ribera d'Alcalà, and Daniele Iudicone. Modelling the complexity of plankton communities exploiting omics potential: From present challenges to an integrative pipeline. *Current Opinion in Systems Biology*, 13:68–74, February 2019

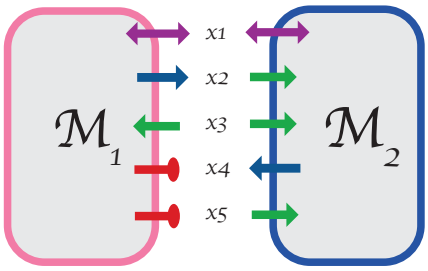


Figure 5.1: Illustration of metabolic interactions between two planktonic metabolic networks. Following FVA, one estimates the directionality of exchanges based on u_b and l_b of exchange reaction. For $u_b > 0 \wedge l_b < 0$, the exchange reaction is reversible (i.e., purple); $u_b > 0 \wedge l_b = 0$, the exchange reaction is irreversible and describes a metabolite production (i.e., blue); $u_b = 0 \wedge l_b < 0$, the exchange reaction is irreversible and describes a metabolite consumption (i.e., green); or $u_b = 0 \wedge l_b = 0$ describes a blocked exchange reaction (i.e., red). Considering the combination of exchange reaction status, the metabolite x_1 could explain the association between $\mathcal{M}_1$ and $\mathcal{M}_2$ but requires further investigations in other environmental conditions. x_2 could explain a causality $\mathcal{M}_1 \longrightarrow \mathcal{M}_2$ like predation or parasitisme. x_3 could explain a competition between $\mathcal{M}_1$ and $\mathcal{M}_2$. The status for metabolites x_4 and x_5 does not allow to explain the association between both organisms.

5.1 Investigation of the ecosystem bio-complexity

Omics measurements provide a large amount of plankton distribution data. Relative abundances are then usually reduced into co-occurrence networks, via several techniques (see Chapter 2). This effort sets the starting information for better characterizing plankton species interactions to shape trophic webs. Plankton species interactions create complex ecological networks, including a copious number of nodes and far more abundant links between them. As mentioned above, a graph, often weighted, describes these interactions, where nodes are either OTUs or genes, and edges depict significant associations between nodes weighted by the score assigned to the link. Beyond the *trendy* modeling of experiments (overall redundant to state-of-the-art multifactorial analysis), the graph remains a formal object that allows further investigations and innovative perspectives.

5.1.1 Fostering graph investigations

From correlation to causality Regardless of the interest of co-occurrence networks to describe putative antagonistic or mutualistic interactions [133], the primary challenge remains to develop networks driven by causal links (i.e., interactions per se) starting from correlation networks. To this purpose, one could take benefit from genomics knowledge that describes the genomic content of each planktonic organism. For each organism, one could build a metabolic network following protocols described in Chapter 3. Considering a given substrate (i.e., chemical parameters such as nutrients), one could then perform a Flux Variability Analysis for two organisms, resp. $\mathcal{M}_1$ and $\mathcal{M}_2$, associated in a given co-occurrence graph (i.e., nodes connected by an edge). For each metabolic model, examining u_b and l_b of each exchange reaction indicates the metabolic flux between two plankton organisms. For instance, Figure 5.1 illustrates that a metabolite x could explain a causality between two associated organisms.

This modeling relies on two distinct models: the co-occurrence network and the metabolic network of each involved organism. The combination of these models could then be used to emphasize causal associations between organisms, rather than solely correlations. However, this combination remains computationally challenging. Indeed, notice here that one must perform such analysis (i) for each association (i.e., edge) and (ii) for several environmental conditions. This induced cost necessitates a change of modeling paradigm, for which the use of constraints programming holds great promises by combining LP problems and motif findings within a graph. In particular, we propose to formulate this problem through Answer Set Programming (ASP). This formulation will allow a flexible encoding that can be easily adjusted to test different pairwise metrics while still being computationally efficient. This technique was already successful in systems biology [78, 15, 36], and we expect a similar efficiency for ecological networks. From a biological viewpoint, this integration could also be challenging. Inevitably, the inference of metabolic causality requires extensive genomic knowledge that is usually out of reach for uncultivable organisms. However, today, one could advocate for the use of recent bioinformatics advances that reconstruct genomes from metagenomes (i.e., MAGS).

Beyond its complexity, the benefit of this modeling is varied. From the technical viewpoint, one must compare this new modeling to previous algorithms that mostly focus on the graph features [19, 55] where herein one could use the quantitative biomass-data variations and metabolic acclimation of each organism. From the theoretical biology viewpoint, one could use our modeling hypothesis to explain causal interactions between organisms that are specific to given conditions or along a gradient (i.e., following the path of a water mass). Hereabouts, the general

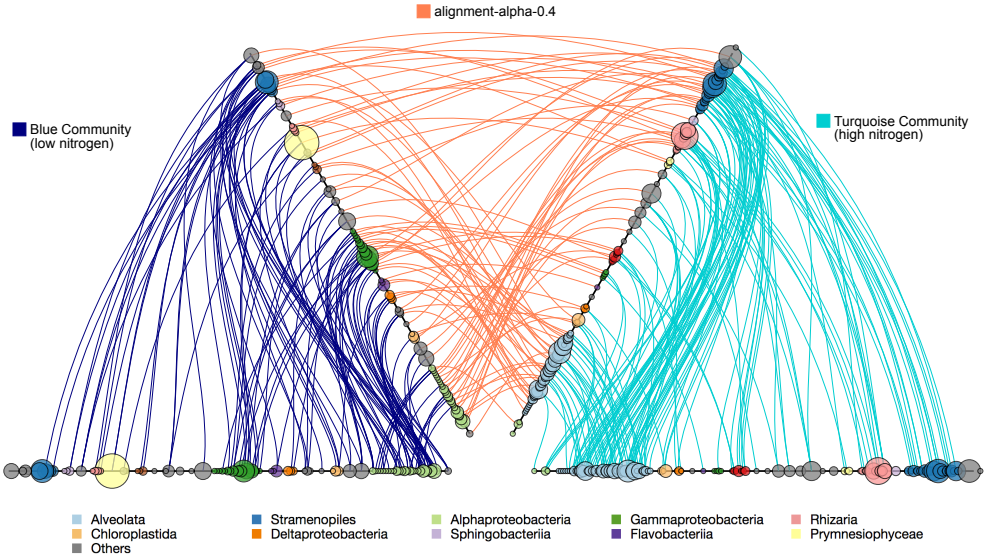


Figure 5.2: Alignment of two planktonic communities from the Sargasso Sea. The blue network depicts a co-occurrence network of a community associated with low nitrogen content, whereas the turquoise one shows a community associated with high nitrogen. The figure describes each network following the Hiveplot nomenclature. Two axes duplicate the nodes of a given community, and edges indicate significant co-occurrence between organisms (from the central ax to the bottom one to avoid edge duplication). The size of the nodes is proportional to the centrality of the organism in its community (betweenness centrality), and their color indicates their respective taxonomy. An orange edge indicates an alignment between organisms of distinct communities. ©Erwan Delage

concept of the ecological niche could be not only applied to the growth of the organism but also on the interaction between organisms. By transitivity, these causal metabolic interactions also impact the use of each metabolic network. This phenomenon could be one of the primary drivers of evolutionary mechanisms such as those resumed by the Black Queen Hypothesis [152, 146] that considers metabolic exchanges as a constraint that explains genome reduction. Thus, the proposed plankton model should deal with the community complexity using a network architecture by performing static analysis or examining dynamical features based on quantitative biomass-data variations.

Graph alignment for comparing ecosystems Assuming the graph as an accurate abstraction of the community structure, we could also extend the investigation of graph properties for the sake of ecological findings. In particular, we propose herein to use it to compare communities from different habitats. The co-occurrence networks enclose the different prevalence of organisms, and the comparison of graphs may emphasize the role of organisms in their respective communities. To examine the extent to which the individual OTUs change the way they interact with other OTUs in the network when living in different environments, one could propose the use of graph-alignment techniques such as L-GRAAL [141]. Initially applied to compare protein-protein interaction networks, this method will align nodes of two graphs when they share both similar topological properties (i.e., for each organism in each network, the graphlet decomposition depicts the number of theoretical motifs in which the given organism is involved), and similar labeling properties (i.e., sequence similarity). As illustrated in Figure 5.2, L-GRAAL will allow us to align organism from two distinct communities if they share: (i) same or similar relations within their co-occurrence network (i.e., a similar number of theoretical motif participation) and (ii) same or similar 16S RNA sequences (i.e., similar taxonomy). A combination of these two features to align organisms will shed light on two different aspects of environmental robustness. First, if the same or closely similar organisms (regarding their DNA sequence) are aligned together, it means that these nodes will establish similar ecological relationships within the community network, despite significant changes in the network structure. Hence, the nodes are themselves ecologically resistant to environmental variability, which will indicate the preservation of specific patterns of ecological interactions, which likely confers robustness to the community network. Second, aligned organisms may be distinctively different in terms of their gene marker sequence (i.e., different 'species'). These organisms, even identified in distinct communities (core species), will establish different relationships with other organisms, and ecological role when environmental conditions shift. Such a pattern would suggest the existence of some level of ecological redundancy and compensatory relationships within the microbial community as a source of network robustness. Conversely, aligned organisms could be unique to each network (non-core OTUs), suggesting that different environmental stressors represent constraints leading to the establishment of similar ecological relations by different planktonic entities.

The results of community alignment remain preliminary [142] but hold great promises to compare, for instance, different oceanic basins from the global ocean. In particular, it is of great interest to compare communities from both poles or to apply such a technique on graphs that depict the co-occurrence of genes for the sake of functional comparison between bioregions. From a computational side, a short-term perspective could integrate the broader genetic diversity within the

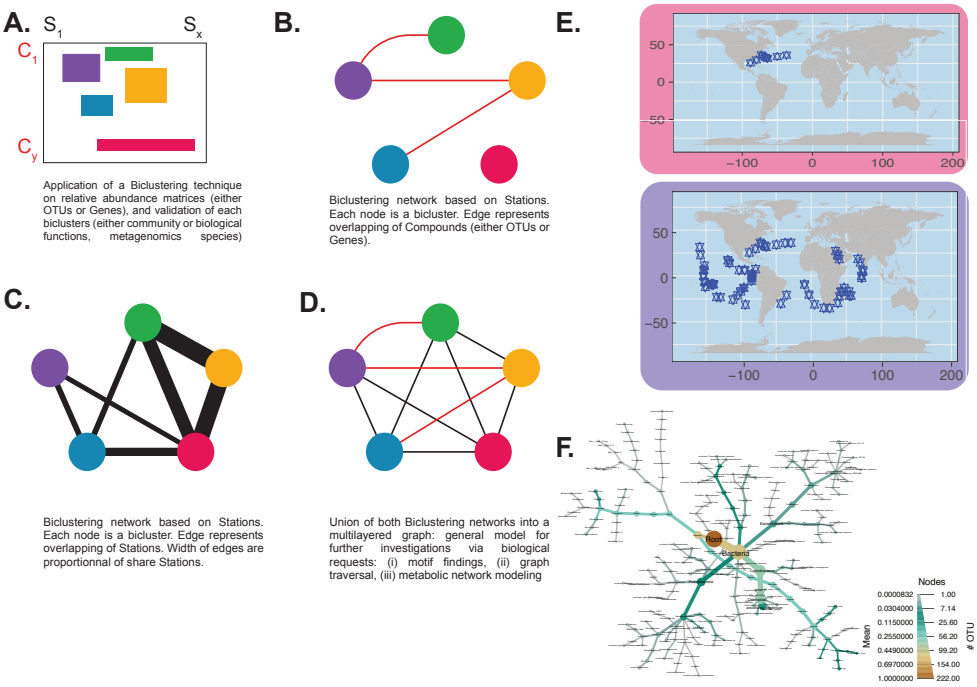


Figure 5.3: Illustration of the identification of bi-clusters. A. depicts the identification of biclusters that cluster given biological components (C_y) on given samples (S_x). Biclusters are linked when they share samples (B) or components (C), which overall help to build a graph of biclusters that emphasize local stabilities (D). When projected on the ocean geography, these biclusters emphasize bioregion based on local interactions (E) between different taxonomy (F). ©Marko Budinich

graph alignment. Thus, instead of resuming an organism by a unique DNA sequence, one could consider sets of sequence variants like proposed by modern sans a priori techniques such as Oligotyping [56] or swarm [140], via the general use of entropy as proposed in [57].

5.1.2 Fostering self-organization properties

The use of systems-biology graph-based methods revealed that one could split the planktonic community into sub-networks. These components represent groups of taxa or genes that correlate to and can predict biogeochemical processes (e.g., carbon flux [90]; iron bioavailability [29], paving the way to future ecosystems mod-

eling that bridges genes, organisms, consortia, and biomes. Beyond the resolution of these techniques to investigate ecosystems, they also change our perceptions of the biological component of the global ocean. Today, we see plankton as a giant, dynamic network (or assemblage of multi-layered metabolic and organismal consortia). This network acclimates and adapts to local ecological conditions principally through multiple modes of interactions, such as red-ox via metabolic networks of organisms or communities of organisms, chemical signaling between organisms, information-transfer through vesicles, viruses, and a broad spectrum of organismal symbioses (from parasitism to mutualism). From an evolutionary viewpoint, some of this knowledge transfer can even be heritable [200], whereas others are the sole consequence of biotic and abiotic interactions (interacting either directly or indirectly). Analyzing this network is complicated because it occurs across various spatiotemporal scales, from local communities to global metacommunities. Indeed, these ecosystems are self-assembling networks of organisms that respond to abiotic and biotic factors, and in turn, feedback on the biological and chemical landscape. To decipher the self-assembling rules that occur in such a multi-layered paradigm, one must focus on the local stability of plankton signals (organism abundance or gene expression). To this purpose, the bi-clustering technique [137] is promising (see Figure 5.3 for illustration). This technique identifies keystone functional subunits that occur at given locations and given biological elements. Preliminary results emphasize these subunits as new indicators of bioregions defined by putative interactions. We will then investigate these subunits via their metabolic and energetic properties across wide-ranging spatiotemporal and biogeochemical dimensions of the ocean ecosystem. Those observations advocate for fostering the biological properties of the interactions at the organismal and genomic scales, which may largely determine ecosystem resilience and dynamics. Answering these fundamental questions of ecosystem self-organization is crucial for predicting global ecosystem change. All these features are additional constraints that one could use to refine the identification of bi-clusters via dedicated modeling in Answer Set Programming.

5.1.3 Adding physical and temporal constraints

Plankton species interactions create complex ecological networks, including a copious number of nodes and far more abundant links between them. As mentioned above, graphs, often weighted, describe such interactions. However, this abstraction does not examine the plankton dynamical features based on quantitative biomass-data variations. Indeed, following the dynamic movement of water masses, planktonic species are subject to changing environmental and ecological interactions at scales ranging from the micro-scale to the global ocean. However,

despite this constant mixing, planktonic organisms are not homogeneous at any scale, from local to global. Local populations show seasonal dynamics and ecological successions, and communities are geographically-differentiated [215]. This local plankton diversity, so-called α-diversity, prevails despite substantial environmental, physical variability. This observation gave rise to the *paradox of the plankton* [96], and still represents a significant challenge for modelers. Whereas high-throughput sequencing of DNA and RNA allows accessing α-diversity by revealing organismal and functional-gene diversities, researchers are still missing a proper abstraction that could combine genomic description with temporal properties. To bridge this conceptual gap, one advocates herein to foster the use of probabilistic models as introduced in Section 4.4 or other abstractions such as continuous-time Markov Chains to model the change of planktonic behavior within a moving water mass that is subject to environmental constraints. An automaton will model each plankton, and its formal investigation could propose new biological insights at the global ocean scale.

5.2 Reduction of the biocomplexity

Arguably the most significant gap in Earth system and climate modeling is the lack of integration of realistic, fine-scaled biological data. Life has dramatically influenced the atmosphere of Earth, and living organisms are at the core of biogeochemical cycles, shaping the structure of ecosystems and climate. Still, while physics and chemistry fairly well constrain current biogeochemical models [6], they are profoundly oversimplified and unrealistic concerning biology. In particular, the NPZD type base model (Nutrients - Phytoplankton - Zooplankton - Detritus) is the standard modeling of this biological component, whose behavior is regulated by the flows between different compartments of the model identified a priori. Despite great modeling successes, this modeling does not take into account (i) biological retroaction processes, nor (ii) the sometimes complex and singular life traits of microorganisms resulting from interactions (from genes to organisms) within an ecosystem. Furthermore, numerous observations and theoretical studies have shown the significant impact that the ecosystem structure has on biogeochemical flows. To fill this gap, new approaches in ecological modeling use the niche concept and functional biodiversity to better formalize the microbial component through the representation of traits (or phytoplankton classes representative of specific traits) in order to understand better the abiotic and biotic constraints regulating biogeochemical flows. However, these traits do not reflect functional diversity, as emphasized by metagenomic and metatranscriptomic knowledge. Indeed, beyond the simple identification of the microbial actors present in an ecosystem (i.e., meta-

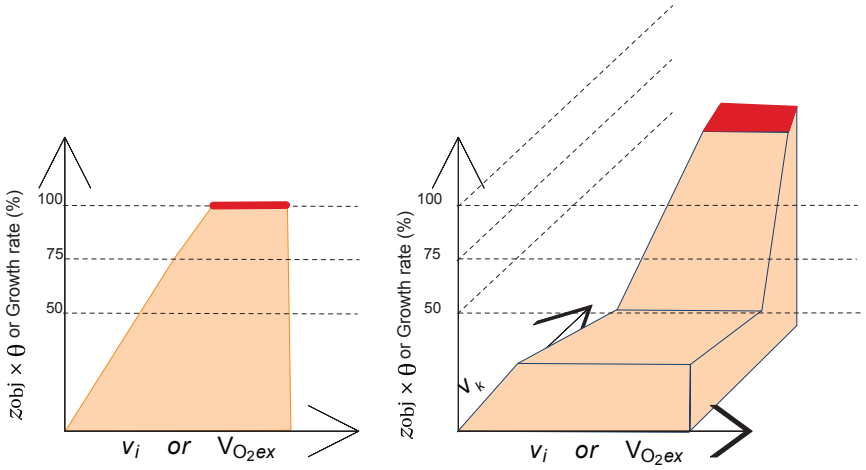


Figure 5.4: Identification of niche from the metabolic model. From Flux Variability Analysis, one estimate fluxes of substrates that are necessary to optimize the growth rate of an organism. By extension, one can determine fluxes necessary for a less optimal growth rate (by a factor θ). As a result, one obtains an estimation of the niche for a given substrate. The right panel depicts a generalization on two substrates, assuming the niche as a cardinal product of substrate-specific niches, which is a particular case.

barcode), high throughput DNA sequencing also makes it possible to acquire almost exhaustive and semi-quantitative data on the functional diversity of microbes (i.e., metagenomics) but also on their adaptations (i.e., metatranscriptomics).

5.2.1 Towards the definition of niche and trait-based model from genomic knowledge

The (meta)genomic information makes it possible to identify new molecular biological markers that provide a better understanding of the fundamental role played by the microbial composition in the biosphere. For instance, in [90], a network analysis identified these biomarkers associated with carbon export, and one could extend their identification via machine learning techniques on satellite images [127]. Notwithstanding of predicting the localization of these biomarkers via statistical inference, one advocates herein to integrate plankton biocomplexity in a mechanistic way to produce next-generation ocean-climate models formally.

Current ocean-climate models are mostly deterministic, ignoring the biodiversity and complex behavior of plankton. As a first step towards incorporating

plankton biocomplexity into ocean models, we propose to refine the concept of niche from omics knowledge. Species distribution models usually depict ecological niches, but these models take into account eco-evo observations for charismatic mega flora/fauna [198], with limited inclusion of microorganisms [53]. Nevertheless, metagenomics and metatranscriptomics of ocean microorganisms [30] may represent a significant advance in shifting towards more dynamical species-distribution models. For instance, a seminal study [149] used genomics data in trait-based models to predict the response of planktonic diatoms to ocean warming. Since a relatively large number of unicellular plankton are genome-wide described, one must pursue shortly the integration of this more accurate description within plankton niche definition, e.g., by predicting niche partitioning from even small genomic differences. However, conceptual models of niche assessment and prediction are still in their infancy and mostly bottom-up driven; i.e., they are based on interactions between organisms and the environment and do not integrate biological interactions. As a near perspective, one proposes herein to use omics knowledge to build a new generation of niche models based on metabolic modeling. In particular, one proposes to redefine the concept of niche via an extension of Flux Variability Analysis. As illustrated by red areas in Figure 5.4, standard FVA consists in identifying upper and lower bounds of exchange fluxes that maximizes an objective function (z_{obj}). By extension, one could perform a similar analysis for an objective function depreciated by a parameter θ. The corresponding upper and lower bounds will there draw a new definition of the niche. Thus, solving the following constraint problem will indicate uncertainties around fluxes exchanges:

Case 1	**Case 2**
maximize v_i	minimize v_i
subject to	subject to
$\mathbf{c}^T\mathbf{v} = z_{\text{obj}} \times \theta, \quad \theta \in [0,1]$	$\mathbf{c}^T\mathbf{v} = z_{\text{obj}} \times \theta, \quad \theta \in [0,1]$
$\mathbf{S}\mathbf{v} = 0$	$\mathbf{S}\mathbf{v} = 0$
$lb_i \leq v_i \leq ub_i, \quad i = 1, ..., n$	$lb_i \leq v_i \leq ub_i, \quad i = 1, ..., n$

Analyzing the metabolic network of an organism via this paradigm will describe a genome-scale niche. Exploration of the corresponding solution space will indicate if the niche of an organism is the cardinal product of niche for all environmental parameters (i.e., exchange fluxes in a metabolic modeling paradigm) or a much more complicated space to investigate. For the sake of validation, this new niche definition will have to be compared to state-of-the-art definitions [198].

5.2.2 New predictive biogeochemical models from metabolic complexity/interactomes

The use of metagenomics and metatranscriptomics will also promote a new generation of biogeochemical models. When not considering organisms and their genomic content, one could foster the functional levels of a given ecosystem. For instance, from the Tara Oceans dataset, one could extract the set of genes [189] that occur in a given sample (i.e., location); or their expression [30]. Beyond the interest of such records, these descriptions represent the ocean genome-scale complexity that one could reduce for the sake of biogeochemical cycling understanding. In particular, amongst the more than150 million genes cataloged by Tara Ocean data, almost 30% encode for enzymes that run the sizeable metabolic engine underlying the redox chemistry of the world ocean. Beyond the description of metabolic potential, we will develop predictive biogeochemical modeling from the corresponding metabolic maps. Via the use of metabolic network reconstruction techniques [135], one will build a prokaryotic functional metabolic network for each Tara Oceans samples. To predict global changes in biogeochemical cycling (notably the flux of carbon to deeper ocean layers by modeling remineralization processes at the metabolic level), one could simulate these networks via state-of-the-art flux balance analysis. One advocates herein that such modeling will represent an accurate proxy of the biogeochemical cycles that occur in the global ocean. Furthermore, such a global, but genome-scale modeling will represent another use of meta-omic experiments.

One could also project these simulation results on the broader picture associated with the Biological Carbon Pump (BCP). In particular, one could resume this complex process by three features: the carbon export already discussed in Chapter 2, the net primary production (NPP), and the flux attenuation that depicts the amount of carbon dissolved along with the sinking process. After normalization, one can plot these measurements in Figure 5.5. Interestingly, the Tara Oceans samples do not cover the whole feasible space of the three BCP features, but rather depict a limited subspace. This result emphasizes the occurrence of a potential system as the three features are interdependent. For the sake of analysis, one could then resume this BCP subspace by a simplex for which each face represents an extreme state of the BCP. In Figure 5.5, the right panel describes the distribution of Tara Oceans samples (dots), and the identification of the simplex and its respective faces (extreme BCP states) in distinct colors. The left panel depicts the metabolic signification of these extreme BCP states in the global ocean. As a perspective, one will investigate the metabolic specificity associated with each extreme BCP state, such as the specific metabolic pathways or specific metabolic coupling that occurs and organisms that could be responsible for them. For validation, one

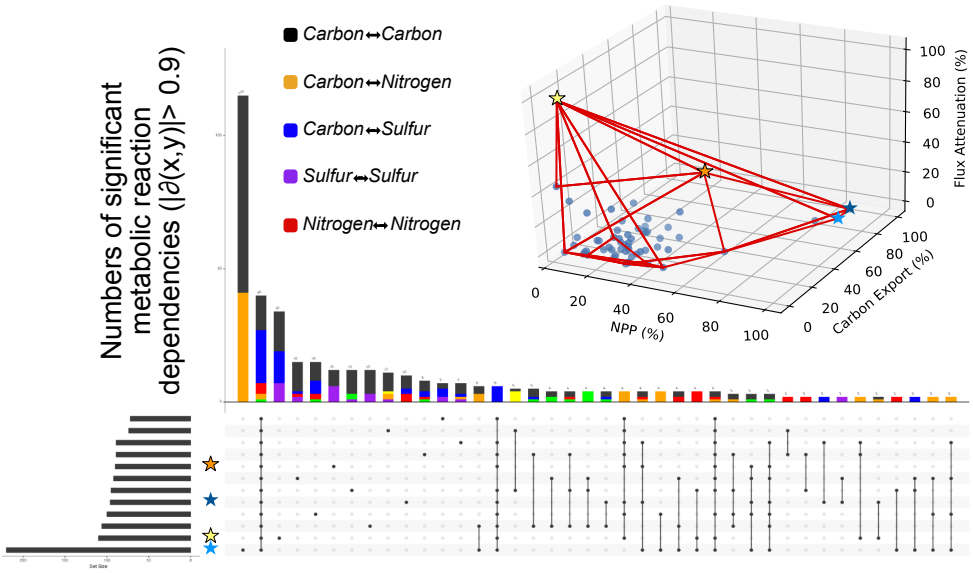


Figure 5.5: Description of a new modeling paradigm for the biological carbon pump. Each Tara Oceans samples are distributed in three complementary features (right panel): the carbon export, the net primary production, and the flux remineralization. One computes the simplex wrapping the whole samples. For each sample, one estimates its metabolic network. For each metabolic network, one computes the dependencies (δ). The most significant dependencies between reaction x and y ($|\delta(x, y)| > 0.9$) are investigated, and associated to the main metabolic pathways and represented in a Uppset diagram (lower left panel). Colored stars depict samples associated with extreme NPP, flux attenuation, or carbon export.

will integrate this new BCP modeling paradigm into next-generation ocean-climate models. Current ocean-climate models are fundamentally deterministic, ignoring the biodiversity and complex behavior of plankton, notably at the ecosystem level. NextGen ocean-climate models integrating these new BCP traits into global-scale ocean physics (e.g., Lagrangian circulation) and chemistry will then be developed, highlighting the role of plankton in climate change, from genes to organisms to ecosystems.

5.3 Synthetic Ecology

As a natural extension, ecosystem modeling could lead to ecosystem engineerings. The use of the constraint modeling paradigm could, therefore, lead to two applicative perspectives that could help to design communities.

Research of optimal nutritional conditions Recent work has reconstructed more than 700 metabolic networks of bacteria associated with the intestinal microbiota [138]. Their availability is an exceptional opportunity to promote community modeling methods. Indeed, beyond the simple description of networks, biologists are committed to understanding interactions and their evolution during nutritional variations. As mentioned above, techniques for simulating metabolic networks are now standard, in particular through the availability of the COBRA software suite [93], which enables FBA and FVA in a context of single organisms. As a natural extension, one could generalize the use of MO-FBA and MO-FVA to simulate the growth of a bacterial co-culture by considering the growth objectives of each strain via a multi-objective optimization paradigm. The simulations explore a Pareto front that describes all possible growth solutions for co-culture under different environmental conditions. Applied on synthetic communities, we will seek to identify the nutritional conditions favorable to each strain of interest, but also a microbial consortium. For example, in the context of the gut microbiome, the nutritional conditions allowing the restoration of damaged flora will then be potential therapies, so-called probiotics.

Ecological capacitance When focusing on a single bacterium, it is possible to determine in silico the genome transformations necessary to modify a given phenotype. In particular, Abdelhalim Larhlimi and collaborators [124] proposed a modeling method called stoichiometric capacitance (SC) that optimally identifies the metabolic transformation that one must add to the metabolic network in order to optimize the production of a metabolite by a given microorganism while maintaining the overall thermodynamic and mass balance properties. By extension, this metabolic engineering approach, which identifies an optimal metabolic transformation, proposes to identify all the reactions producing these transformations as well as the genes encoding them. The genes identified will then be potential candidates for synthetic strain constructions (i.e., genes allowing the encoding of enzymes necessary for capacitance), because they modify the phenotype of a strain optimally. By considering community metabolism as the union of the metabolism of all strains present in an ecosystem, a short term perspective will be to look for one to several SC capacities whose implementation will allow an environment un-

favorable to a given microbial pathogen, or to the benefit to another strain. Each capacitance of interest will then be decomposed and associated with a set of genes necessary for its implementation. Notice here that all genes are not unique and that one will consider several gene combinations. Each gene combination will then be searched in the knowledge bases using web semantics approaches to identify microbial strains that may possess the genes. The challenge herein is to isolate the minimum strains containing the genes necessary for capacitance. At this stage, and for validation of our synthetic modeling protocol, one will analyze the candidate genomes against other complementary techniques based on recently published graph theory [55].

Bibliography

[1] Vicente Acuña, Andrés Aravena, Carito Guziolowski, Damien Eveillard, Anne Siegel, and Alejandro Maass. Deciphering transcriptional regulations coordinating the response to environmental changes. *BMC bioinformatics*, 17(1):35, January 2016.

[2] Jamil Ahmad, Jérémie Bourdon, Damien Eveillard, Jonathan Fromentin, Olivier Roux, and Christine Sinoquet. Temporal constraints of a gene regulatory network: Refining a qualitative simulation. *BioSystems*, 98(3):149–159, December 2009.

[3] Madhur Anand, Andrew Gonzalez, Frédéric Guichard, Jurek Kolasa, and Lael Parrott. Ecological Systems as Complex Systems: Challenges for an Emerging Science. *Diversity*, 2(3):395–410, March 2010.

[4] Robert S Anderson, A P Alivisatos, M J Blaser, E L Brodie, M Chun, J L Dangl, T J Donohue, P C Dorrestein, J A Gilbert, J L Green, J K Jansson, R Knight, M E Maxon, M J McFall-Ngai, J F Miller, K S Pollard, E G Ruby, S A Taha, and Unified Microbiome Initiative Consortium. MICROBIOME. A unified initiative to harness Earth's microbiomes. *Science (New York, NY)*, 350(6260):507–508, October 2015.

[5] Sébastien Angibaud, Damien Eveillard, Guillaume Fertin, and Irena Rusu. Comparing bacterial genomes by searching their common intervals. In *Bioinformatics and Computational Biology*, pages 102–113. Springer, 2009.

[6] Olivier Aumont, Christian Éthé, Alessandro Tagliabue, Laurent Bopp, and Marion Gehlen. Pisces-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development Discussions*, 8(2), 2015.

[7] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews Genetics*, 5(2):101–113, February 2004.

[8] Georg Basler, Zoran Nikoloski, Abdelhalim Larhlimi, Albert-László Barabási, and Yang-Yu Liu. Control of fluxes in metabolic networks. *Genome research*, 2016.

[9] Mathieu Bastian, Sébastien Heymann, and Mathieu Jacomy. Gephi : An Open Source Software for Exploring and Manipulating Networks. In *International AAAI Conference on Weblogs and Social Media*, pages 1–2, 2009.

[10] G Batt, Delphine Ropers, Hidde de de Jong, Johannes Geiselmann, Michel Page, and et al. Qualitative analysis and verification of hybrid models of genetic regulatory networks: Nutritional .... *International Workshop on Hybrid Systems: Computation and Control (HSCC)*, LNCS 3414:134–150, 2005.

[11] Grégory Batt, Delphine Ropers, Hidde de de Jong, Johannes Geiselmann, Radu Mateescu, Michel Page, and Dominique Schneider. Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli. *Bioinformatics (Oxford, England)*, 21 Suppl 1:i19–28, June 2005.

[12] Olivier Bernard and Jean-Luc Gouzé. Transient behavior of biological loop models with application to the Droop model. *Mathematical biosciences*, 127(1):19–43, May 1995.

[13] Matthew B Biggs, Gregory L Medlock, Glynis L Kolling, and Jason A Papin. Metabolic network modeling of microbial communities. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(5):317–334, 2015.

[14] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data-the story so far. *International journal on semantic web and information systems*, 5(3):1–22, 2009.

[15] Philippe Bordron, Damien Eveillard, Alejandro Maass, Anne Siegel, and Sven Thiele. An ASP application in integrative biology: identification of functional gene units. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, 8148:206–218, 2013.

[16] Philippe Bordron, Damien Eveillard, and Irena Rusu. Integrated analysis of the gene neighbouring impact on bacterial metabolic networks. *IET systems biology*, 5(4):261–268, July 2011.

[17] Philippe Bordron, Damien Eveillard, and Irena Rusu. SIPPER: A flexible method to integrate heterogeneous data into a metabolic network. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pages 40–45. IEEE, February 2011.

[18] Philippe Bordron, Mauricio Latorre, Maria Paz Cortés, Mauricio González, Sven Thiele, Anne Siegel, Alejandro Maass, and Damien Eveillard. Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen*, 5(1):106–117, February 2016.

[19] Elhanan Borenstein, Martin Kupiec, Marcus W Feldman, and Eytan Ruppin. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences of the United States of America*, 105(38):14482–14487, September 2008.

[20] Jérémie Bourdon, Damien Eveillard, and Anne Siegel. Integrating quantitative knowledge into a qualitative gene regulatory network. *PLoS computational biology*, 7(9):e1002157, September 2011.

[21] Nicholas J Bouskill, Damien Eveillard, Diana Chien, Amal Jayakumar, and Bess B Ward. Environmental factors determining ammonia-oxidizing organism distribution and diversity in marine environments. *Environmental Microbiology*, 14(3):714–729, March 2012.

[22] Nicholas J Bouskill, Damien Eveillard, Gregory O'mullan, George A Jackson, and Bess B Ward. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environmental Microbiology*, 13(4):872–886, April 2011.

[23] Jennifer L Bowen, Bess B Ward, Hilary G Morrison, John E Hobbie, Ivan Valiela, Linda A Deegan, and Mitchell L Sogin. Microbial community composition in sediments resists perturbation by nutrient enrichment. *The ISME Journal*, 5(9):1540–1548, September 2011.

[24] Frédéric Boyer, Anne Morgat, Laurent Labarre, Joël Pothier, and Alain Viari. Syntons, metabolons and interactons: an exact graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics (Oxford, England)*, 21(23):4209–4215, December 2005.

[25] Mark V Brown, Gayle K Philip, John A Bunge, Matthew C Smith, Andrew Bissett, Federico M Lauro, Jed A Fuhrman, and Stuart P Donachie. Microbial community structure in the North Pacific ocean. *The ISME Journal*, 3(12):1374–1386, July 2009.

[26] Jennifer R Brum, J Cesar Ignacio-Espinoza, Simon Roux, Guilhem Doulcier, Silvia G Acinas, Adriana Alberti, Samuel Chaffron, Corinne Cruaud, Colomban de Vargas, Josep M Gasol, Gabriel Gorsky, Ann C Gregory, Lionel Guidi, Pascal Hingamp, Daniele Iudicone, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Bonnie T Poulos, Sarah M Schwenck, Sabrina Speich, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Sarah Searson, Tara Oceans coordinators, Peer Bork, Chris Bowler, Shinichi Sunagawa, Patrick Wincker, Eric Karsenti, and Matthew B Sullivan. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science (New York, NY)*, 348(6237):1261498–1261498, May 2015.

[27] Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. OPINION PAPER Evolutionary Constraint-Based Formulation Requires New Bi-level Solving Techniques. In *Computational Methods in Systems Biology*, pages 279–281, Nantes, September 2015. Springer International Publishing.

[28] Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE*, 12(2):e0171744, 2017.

[29] Luigi Caputi, Quentin Carradec, Damien Eveillard, Amos Kirilovsky, Eric Pelletier, Juan J Pierella Karlusich, Fabio Rocha Jimenez Vieira, Emilie Villar, Samuel Chaffron, Shruti Malviya, Eleonora Scalco, Silvia G Acinas, Adriana Alberti, Jean-Marc Aury, Anne Sophie Benoiston, Alexis Bertrand, Tristan Biard, Lucie Bittner, Martine Boccara, Jennifer R Brum, Christophe Brunet, Greta Busseni, Anna Carratalà, Hervé Claustre, Luis Pedro Coelho, Sébastien Colin, Salvatore D'Aniello, Corinne Da Silva, Marianna Del Core, Hugo Doré, Stéphane Gasparini, Florian Kokoszka, Jean-Louis Jamet, Christophe Lejeusne, Cyrille Lepoivre, Magali Lescot, Gipsi Lima-Mendez, Fabien Lombard, Julius Lukeš, Nicolas Maillet, Mohammed-Amin Madoui, Elodie Martinez, Maria Grazia Mazzocchi, Mario B Néou, Javier Paz Yepes, Julie Poulain, Simon Ramondenc, Jean-Baptiste Romagnan, Simon Roux, Daniela Salvagio Manta, Remo Sanges, Mario Sprovieri, Vincent Taillandier, Atsuko Tanaka, Leila Tirichine, Camille Trottier, Julia Uitz, Alaguraj Veluchamy, Jana Veselá, Flora Vincent, Sheree Yau, Stefanie Kandels-Lewis, Sarah Searson, Céline Dimier, Marc Picheral, Peer Bork, Lionel Guidi, Paolo Sordino, Matthew B Sullivan, Alessandro Tagliabue, Adriana Zingone, Laurence Garczarek, Fabrizio d'Ortenzio, Pierre Testor, Maurizio Ribera d'Alcalà, Patrick Wincker, Chris Bowler, Tara Oceans coordinators, Emmanuel Boss, Colomban Vargas, Michael J Follows, Gabriel

Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Lee Karp-Boss, Eric Karsenti, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen Raes, Emmanuel G Reynaud, Christian Sardet, Mike Sieracki, Sabrina Speich, Lars Stemmann, Shinichi Sunagawa, Didier Velayoudon, and Jean Weissenbach. Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Global Biogeochemical Cycles*, 4(30):10,438, March 2019.

[30] Quentin Carradec, Eric Pelletier, Corinne Da Silva, Adriana Alberti, Yoann Seeleuthner, Romain Blanc-Mathieu, Gipsi Lima-Mendez, Fabio Rocha, Leila Tirichine, Karine Labadie, Amos Kirilovsky, Alexis Bertrand, Stefan Engelen, Mohammed-Amin Madoui, Raphaël Méheust, Julie Poulain, Sarah Romac, Daniel J Richter, Genki Yoshikawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Sarah Searson, Tara Oceans coordinators, Olivier Jaillon, Jean-Marc Aury, Eric Karsenti, Matthew B Sullivan, Shinichi Sunagawa, Peer Bork, Fabrice Not, Pascal Hingamp, Jeroen Raes, Lionel Guidi, Hiroyuki Ogata, Colomban de Vargas, Daniele Iudicone, Chris Bowler, and Patrick Wincker. A global ocean atlas of eukaryotic genes. *Nature communications*, 9(1):373, January 2018.

[31] Sean B Carroll. *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom*. Number 54. WW Norton & Company, 2005.

[32] Ron Caspi, Tomer Altman, Kate Dreher, Carol A Fulcher, Pallavi Subhraveti, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Quang Ong, Suzanne Paley, Anuradha Pujar, Alexander G Shearer, Michael Travers, Deepika Weerasinghe, Peifen Zhang, and Peter D Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, 40(Database issue):D742–53, January 2012.

[33] Samuel Chaffron, Hubert Rehrauer, Jakob Pernthaler, and Christian von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959, July 2010.

[34] Abalo Chango, Afif Abdel Nour, Souad Bousserouel, Damien Eveillard, Pauline M Anton, and Jean-Louis Guéant. Time course gene expression in the one-carbon metabolism network using HepG2 cell line grown in folate-deficient medium. *The Journal of nutritional biochemistry*, 20(4):312–320, April 2009.

[35] V J Coles, M R Stukel, M T Brooks, A Burd, B C Crump, M A Moran, J H Paul, B M Satinsky, P L Yager, B L Zielinski, and R R Hood. Ocean biogeochemistry modeled with emergent trait-based genomics. *Science (New York, NY)*, 358(6367):1149–1154, December 2017.

[36] Guillaume Collet, Damien Eveillard, Martin Gebser, Sylvain Prigent, Torsten Schaub, Anne Siegel, and Sven Thiele. Extending the Metabolic Network of Ectocarpus Siliculosus using Answer Set Programming. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, LNAI8148:245–256, September 2013.

[37] Samuel Collombet, Chris van Oevelen, Jose Luis Sardina Ortega, Wassim Abou-Jaoudé, Bruno Di Stefano, Morgane Thomas-Chollier, Thomas Graf, and Denis Thieffry. Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, 114(23):5792–5799, 2017.

[38] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[39] Jacob A Cram, Cheryl-Emiliane T Chow, Rohan Sachdeva, David M Needham, Alma E Parada, Joshua A Steele, and Jed A Fuhrman. Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *The ISME journal*, 9(3):563, 2015.

[40] Domenico D'Alelio, Damien Eveillard, Victoria J Coles, Luigi Caputi, Maurizio Ribera d'Alcalà, and Daniele Iudicone. Modelling the complexity of plankton communities exploiting omics potential: From present challenges to an integrative pipeline. *Current Opinion in Systems Biology*, 13:68–74, February 2019.

[41] Charles Darwin. *On the origin of species, 1859*. Routledge, 2004.

[42] Luis F de Figueiredo, Adam Podhorski, Angel Rubio, Christoph Kaleta, John E Beasley, Stefan Schuster, and Francisco J Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics (Oxford, England)*, 25(23):3158–3165, December 2009.

[43] Hidde de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology : a journal of computational molecular cell biology*, 9(1):67–103, 2002.

[44] Hidde de de Jong, Jean-Luc Gouzé, Céline Hernandez, Michel Page, Tewfik Sari, and Johannes Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bulletin of mathematical biology*, 66(2):301–340, March 2004.

[45] Colomban de Vargas, Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, Cédric Berney, Noan Le Bescot, Ian Probert, Margaux Carmichael, Julie Poulain, Sarah Romac, Sébastien Colin, Jean-Marc Aury, Lucie Bittner, Samuel Chaffron, Micah Dunthorn, Stefan Engelen, Olga Flegontova, Lionel Guidi, Aleš Horák, Olivier Jaillon, Gipsi Lima-Mendez, Julius Lukeš, Shruti Malviya, Raphael Morard, Matthieu Mulot, Eleonora Scalco, Raffaele Siano, Flora Vincent, Adriana Zingone, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Silvia G Acinas, Peer Bork, Chris Bowler, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen Raes, Michael E Sieracki, Sabrina Speich, Lars Stemmann, Shinichi Sunagawa, Jean Weissenbach, Patrick Wincker, and Eric Karsenti. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science (New York, NY)*, 348(6237):1261605–1261605, May 2015.

[46] Géraldine Del Mondo, Damien Eveillard, and Irena Rusu. Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions. *Bioinformatics (Oxford, England)*, 25(7):926–932, April 2009.

[47] Benoît Delahaye, Damien Eveillard, and Nicholas Bouskill. On the Power of Uncertainties in Microbial System Modeling: No Need To Hide Them Anymore. *mSystems*, 2(6), November 2017.

[48] Tom O Delmont, Christopher Quince, Alon Shaiber, Özcan C Esen, Sonny TM Lee, Michael S Rappé, Sandra L MacLellan, Sebastian Lücker, and A Murat Eren. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, June 2018.

[49] Edward F DeLong and David M Karl. Genomic perspectives in microbial oceanography. *Nature*, 437(7057):336–342, September 2005.

[50] Simon M Dittami, Damien Eveillard, and Thierry Tonon. A metabolic approach to study algal-bacterial interactions in changing environments. *Molecular ecology*, 23(7):1656–1660, April 2014.

[51] Erwan Drezen, Guillaume Rizk, Rayan Chikhi, Charles Deltel, Claire Lemaitre, Pierre Peterlongo, and Dominique Lavenier. Gatb: Genome assembly & analysis tool box. *Bioinformatics*, 30(20):2959–2961, 2014.

[52] MR Droop. Some thoughts on nutrient limitation in algae 1. *Journal of Phycology*, 9(3):264–272, 1973.

[53] Sally Eaton, Christopher Ellis, David Genney, Richard Thompson, Rebecca Yahr, and Daniel T Haydon. Adding small species to the big picture: Species distribution modelling in an age of landscape scale conservation. *Biological Conservation*, 217:251–258, 2018.

[54] J S Edwards and B O Palsson. Robustness analysis of the Escherichia coli metabolic network. *Biotechnology progress*, 16(6):927–939, 2000.

[55] Alexander Eng and Elhanan Borenstein. An algorithm for designing minimal microbial communities with desired metabolic capacities. *Bioinformatics (Oxford, England)*, 32(13):2008–2016, July 2016.

[56] A Murat Eren, Loïs Maignien, Woo Jun Sul, Leslie G Murphy, Sharon L Grim, Hilary G Morrison, and Mitchell L Sogin. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in ecology and evolution / British Ecological Society*, 4(12), December 2013.

[57] A Murat Eren, Hilary G Morrison, Pamela J Lescault, Julie Reveillaud, Joseph H Vineis, and Mitchell L Sogin. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, 9(4):968–979, April 2015.

[58] Ana E Escalante, María Rebolleda-Gómez, Mariana Benítez, and Michael Travisano. Ecological perspectives on synthetic biology: insights from microbial population biology. *Frontiers in microbiology*, 6(219):143, 2015.

[59] Damien Eveillard, Nicholas J Bouskill, Damien Vintache, Julien Gras, Bess B Ward, and Jérémie Bourdon. Probabilistic Modeling of Microbial Metabolic Networks for Integrating Partial Quantitative Knowledge Within the Nitrogen Cycle. *Frontiers in microbiology*, 9:395, January 2019.

[60] Damien Eveillard, Delphine Ropers, Hidde de de Jong, Christiane Branlant, and Alexander Bockmayr. A multi-scale constraint programming model of alternative splicing regulation. *Theoretical Computer Science*, 325(1):3–24, September 2004.

[61] Karoline Faust, Gipsi Lima-Mendez, Jean-Sébastien Lerat, Jarupon F Sathirapongsasuti, Rob Knight, Curtis Huttenhower, Tom Lenaerts, and Jeroen Raes. Cross-biome comparison of microbial association networks. *Frontiers in microbiology*, 6:1200, 2015.

[62] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, August 2012.

[63] Karoline Faust, J Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS computational biology*, 8(7):e1002606, July 2012.

[64] G Fertin, A Labarre, I Rusu, E Tannier, and S Vialette. Combinatorics of genome rearrangements, ser. computational molecular biology, 2009.

[65] Guillaume Fertin, Anthony Labarre, Irena Rusu, Eric Tannier, and StÃ phane Vialette. *Combinatorics of Genome Rearrangements (Computational Molecular Biology)*. The MIT Press, 1 edition, August 2009.

[66] Jasmin Fisher, David Harel, and Thomas A Henzinger. Biology as reactivity. *Communications of the ACM*, 54(10), October 2011.

[67] Michael J Follows and Stephanie Dutkiewicz. Modeling diverse communities of marine microbes. *Annual review of marine science*, 3:427–451, 2011.

[68] Christopher A Francis, Gregory D O'Mullan, and Bess B Ward. Diversity of ammonia monooxygenase (amoA) gene across environmental gradients in Chesapeake Bay sediments. *Geobiology*, 1:129–140, 2003.

[69] Shiri Freilich, Raphy Zarecki, Omer Eilam, Ella Shtifman Segal, Christopher S Henry, Martin Kupiec, Uri Gophna, Roded Sharan, and Eytan Ruppin. Competitive and cooperative metabolic interactions in bacterial communities. *Nature communications*, 2:589, December 2011.

[70] Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687, 2012.

[71] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[72] Jonathan Fromentin, Damien Eveillard, and Olivier Roux. Hybrid Modeling of Gene Regulatory Networks: Mixing Temporal and Qualitative Biological Properties. Technical report, BioXiv, December 2008.

[73] Jonathan Fromentin, Damien Eveillard, and Olivier Roux. Hybrid modeling of biological networks: mixing temporal and qualitative biological properties. *BMC systems biology*, 4:79, 2010.

[74] Jed A Fuhrman, Jacob A Cram, and David M Needham. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13(3):133, 2015.

[75] Jed A Fuhrman, Ian Hewson, Michael S Schwalbach, Joshua A Steele, Mark V Brown, and Shahid Naeem. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences*, 103(35):13104–13109, 2006.

[76] Julien Gagneur and Steffen Klamt. Computation of elementary modes: a unifying framework and the new binary approach. *BMC bioinformatics*, 5:175, November 2004.

[77] Julien Gagneur, Roland Krause, Tewis Bouwmeester, and Georg Casari. Modular decomposition of protein-protein interaction networks. *Genome Biology*, 5(8):R57, 2004.

[78] M Gebser, A. Konig, T Schaub, S Thiele, and P Veber. The BioASP library: ASP solutions for systems biology. *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, 1:383–389, 2010.

[79] Martin Gebser, Benjamin Kaufmann, Roland Kaminski, Max Ostrowski, Torsten Schaub, and Marius Schneider. Potassco: The potsdam answer set solving collection. *Ai Communications*, 24(2):107–124, 2011.

[80] Martin Gebser, Benjamin Kaufmann, Andre Neumann, and Torsten Schaub. Conflict-driven answer set solving. In *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc, January 2007.

[81] R Ghosh and C Tomlin. Symbolic reachable set computation of piecewise affine hybrid automata and its application to .... *IEEE Systems biology*, 2004.

[82] Tara A Gianoulis, Jeroen Raes, Prianka V Patel, Robert Bjornson, Jan O Korbel, Ivica Letunic, Takuji Yamada, Alberto Paccanaro, Lars J Jensen, Michael Snyder, Peer Bork, and Mark B Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1374–1379, February 2009.

[83] Jack A Gilbert, Robert A Quinn, Justine Debelius, Zhenjiang Z Xu, James Morton, Neha Garg, Janet K Jansson, Pieter C Dorrestein, and Rob Knight. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103, July 2016.

[84] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58:35–55, 2007.

[85] Fred Glover. Future paths for integer programming and links to artificial intelligence. *Computers & operations research*, 13(5):533–549, 1986.

[86] Etienne Z Gnimpieba, Damien Eveillard, Jean-Louis Guéant, and Abalo Chango. Using logic programming for modeling the one-carbon metabolism network to study the impact of folate deficiency on methylation processes. *Molecular bioSystems*, 7(8):2508–2521, August 2011.

[87] Oscar Godoy, Ignasi Bartomeus, Rudolf P Rohr, and Serguei Saavedra. Towards the Integration of Niche and Network Theories. *Trends in ecology & evolution (Personal edition)*, 33(4):287–300, April 2018.

[88] A Goldsztejn, O Mullier, Damien Eveillard, and H Hosobe. Including Ordinary Differential Equations Based Constraints in the Standard CP Framework. *Principles and Practice of Constraint Programming, CP2010*, LNCS 6308:221–235, 2010.

[89] Volker Grimm, Uta Berger, Finn Bastiansen, Sigrunn Eliassen, Vincent Ginot, Jarl Giske, John Goss-Custard, Tamara Grand, Simone K Heinz, Geir Huse, et al. A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, 198(1-2):115–126, 2006.

[90] Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline, Jennifer R Brum, Jennifer Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Tara Oceans Consortium Coordinators, Lars Stemmann, Fabrice Not, Pascal Hingamp,

Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stéphane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Peer Bork, Colomban de Vargas, Daniele Iudicone, Matthew B Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470, February 2016.

[91] Gutiérrez-Ríos, Rosa María, Rosenblueth, David A, Loza, José Antonio, Huerta, Araceli M, Glasner, Jeremy D, Blattner, Fred R, and Collado-Vides, Julio. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Research*, 13(11):2435–2443, November 2003.

[92] Joshua J Hamilton and Jennifer L Reed. Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, 16(1):49–59, January 2014.

[93] Laurent Heirendt, Sylvain Arreckx, Thomas Pfau, Sebastian N Mendoza, Anne Richelle, Almut Heinken, Hulda S Haraldsdottir, Jacek Wachowiak, Sarah M Keating, Vanja Vlasov, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nature protocols*, page 1, 2019.

[94] Manuela Hische, Abdelhalim Larhlimi, Franziska Schwarz, Antje Fischer-Rosinský, Thomas Bobbert, Anke Assmann, Gareth S Catchpole, Andreas FH Pfeiffer, Lothar Willmitzer, Joachim Selbig, and Joachim Spranger. A distinct metabolic signature predicts development of fasting plasma glucose. *Journal of clinical bioinformatics*, 2:3, 2012.

[95] Michael Huerta, Gregory Downing, Florence Haseltine, Belinda Seto, and Yuan Liu. Nih working definition of bioinformatics and computational biology. *US National Institute of Health*, 2000.

[96] G Evelyn Hutchinson. The paradox of the plankton. *The American Naturalist*, 95(882):137–145, 1961.

[97] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, and Gil McVean. De novo assembly and genotyping of variants using colored de bruijn graphs. *Nature genetics*, 44(2):226, 2012.

[98] Franck Jabot and Jérôme Chave. Analyzing tropical forest tree species abundance distributions using a nonneutral model and through approximate bayesian inf ce. *The American Naturalist*, 178(2):E37–E47, 2011.

[99] Shahrad Jamshidi, Jocelyn E Behm, Damien Eveillard, E Toby Kiers, and Philippe Vandenkoornhuyse. Using hybrid automata modelling to study phenotypic plasticity and allocation strategies in the plant mycorrhizal mutualism. *Ecological Modelling*, 311:11–19, September 2015.

[100] H Jeong, S P Mason, A L Barabási, and Z N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[101] Talia NM Jewell, Ulas Karaoz, Eoin L Brodie, Kenneth H Williams, and Harry R Beller. Metatranscriptomic evidence of pervasive and diverse chemolithoautotrophy relevant to c, s, n and fe cycling in a shallow alluvial aquifer. *The ISME journal*, 10(9):2106, 2016.

[102] Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating'omics' data sets. *Nature reviews Molecular cell biology*, 7(3):198, 2006.

[103] Tammi Kaeberlein, Kim Lewis, and Slava S Epstein. Isolating" uncultivable" microorganisms in pure culture in a simulated natural environment. *Science*, 296(5570):1127–1129, 2002.

[104] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[105] MB Karner, EF DeLong, and DM Karl. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature*, 409(6819):507–510, 2001.

[106] Peter D Karp, Suzanne M Paley, Markus Krummenacker, Mario Latendresse, Joseph M Dale, Thomas J Lee, Pallavi Kaipa, Fred Gilham, Aaron Spaulding, Liviu Popescu, Tomer Altman, Ian Paulsen, Ingrid M Keseler, and Ron Caspi. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 11(1):40–79, January 2010.

[107] Eric Karsenti. Self-organization in cell biology: a brief history. *Nature Reviews Molecular Cell Biology*, 9(3):255–262, March 2008.

[108] Éric Karsenti. *Aux sources de la vie*. Flammarion, 2018.

[109] Eric Karsenti, Silvia G Acinas, Peer Bork, Chris Bowler, Colomban de Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean-Michel Claverie, Mick Follows, Gaby Gorsky, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels-Lewis, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Emmanuel Georges Reynaud,

Christian Sardet, Michael E Sieracki, Sabrina Speich, Didier Velayoudon, Jean Weissenbach, Patrick Wincker, and Tara Oceans Consortium. A holistic approach to marine eco-systems biology. *PLoS biology*, 9(10):e1001177, October 2011.

[110] Stuart A Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

[111] Ruchir A Khandelwal, Brett G Olivier, Wilfred F M Röling, Bas Teusink, and Frank J Bruggeman. Community flux balance analysis for microbial consortia at balanced growth. *PLoS ONE*, 8(5):e64567, 2013.

[112] Zachary A King, Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A Lerman, Ali Ebrahim, Bernhard O Palsson, and Nathan E Lewis. Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522, 2015.

[113] Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, November 2002.

[114] Hiroaki Kitano. Systems biology: a brief overview. *Science (New York, NY)*, 295(5560):1662–1664, March 2002.

[115] Steffen Klamt and Ernst Dieter Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics (Oxford, England)*, 20(2):226–234, January 2004.

[116] Niels Klitgord and Daniel Segrè. The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Informatics. International Conference on Genome Informatics*, 22(22):41–55, January 2010.

[117] Niels Klitgord and Daniel Segrè. Ecosystems biology of microbial metabolism. *Current opinion in biotechnology*, 22(4):541–546, August 2011.

[118] Kishori M Konwar, Niels W Hanson, Antoine P Pagé, and Steven J Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC bioinformatics*, 14:202, 2013.

[119] Martin Krzywinski, Inanc Birol, Steven J M Jones, and Marco A Marra. Hive plots–rational approach to visualizing networks. *Briefings in Bioinformatics*, 13(5):627–644, September 2012.

[120] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS computational biology*, 11(5):e1004226, May 2015.

[121] Marcel MM Kuypers, Hannah K Marchant, and Boran Kartal. The microbial nitrogen-cycling network. *Nature Reviews Microbiology*, 16(5):263, 2018.

[122] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[123] Julie Laniau, Clémence Frioux, Jacques Nicolas, Caroline Baroukh, Maria Paz Cortés, Jeanne Got, Camille Trottier, Damien Eveillard, and Anne Siegel. Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, 5:e3860, 2017.

[124] Abdelhalim Larhlimi, Georg Basler, Sergio Grimbs, Joachim Selbig, and Zoran Nikoloski. Stoichiometric capacitance reveals the theoretical capabilities of metabolic networks. *Bioinformatics (Oxford, England)*, 28(18):i502–i508, September 2012.

[125] Abdelhalim Larhlimi, Sylvain Blachon, Joachim Selbig, and Zoran Nikoloski. Robustness of metabolic networks: a review of existing definitions. *Biosystems*, 106(1):1–8, 2011.

[126] Abdelhalim Larhlimi and Alexander Bockmayr. A new constraint-based description of the steady-state flux cone of metabolic networks. *Discrete Applied Mathematics*, 157(10):2257–2266, 2009.

[127] Peter E Larsen, Nicole Scott, Anton F Post, Dawn Field, Rob Knight, Yuki Hamada, and Jack A Gilbert. Satellite remote sensing data can be used to model marine microbial metabolite turnover. *The ISME Journal*, 9(1):166–179, January 2015.

[128] Axel Legay, Benoît Delahaye, and Saddek Bensalem. Statistical Model Checking - An Overview. *RV*, 6418(6):122–135, 2010.

[129] Pierre Legendre and Louis Legendre. *Numerical Ecology*. Elsevier, 2012.

[130] Jennifer Levering, Christopher L Dupont, Andrew E Allen, Bernhard Ø Palsson, and Karsten Zengler. Integrated Regulatory and Metabolic Networks of the Marine Diatom Phaeodactylum tricornutum Predict the Response to Rising CO2 Levels. *mSystems*, 2(1), January 2017.

[131] Peng Li, Chaoyang Zhang, Edward J Perkins, Ping Gong, and Youping Deng. Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks. *BMC bioinformatics*, 8 Suppl 7:S13, 2007.

[132] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321, 2015.

[133] Gipsi Lima-Mendez, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, Fabrizio Carcillo, Samuel Chaffron, J Cesar Ignacio-Espinosa, Simon Roux, Flora Vincent, Lucie Bittner, Youssef Darzi, Jun Wang, Stéphane Audic, Léo Berline, Gianluca Bontempi, Ana M Cabello, Laurent Coppola, Francisco M Cornejo-Castillo, Francesco d'Ovidio, Luc De Meester, Isabel Ferrera, Marie-José Garet-Delmas, Lionel Guidi, Elena Lara, Stéphane Pesant, Marta Royo-Llonch, Guillem Salazar, Pablo Sánchez, Marta Sebastian, Caroline Souffreau, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Gabriel Gorsky, Fabrice Not, Hiroyuki Ogata, Sabrina Speich, Lars Stemmann, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Shinichi Sunagawa, Peer Bork, Matthew B Sullivan, Eric Karsenti, Chris Bowler, Colomban de Vargas, and Jeroen Raes. Ocean plankton. Determinants of community structure in the global plankton interactome. *Science (New York, NY)*, 348(6237):1262073–1262073, May 2015.

[134] Richard A Long, Damien Eveillard, Shelli L M Franco, Eric Reeves, and James L Pinckney. Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. *FEMS microbiology ecology*, 83(1):74–81, January 2013.

[135] Daniel Machado, Sergej Andrejev, Melanie Tramontano, and Kiran Raosaheb Patil. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research*, 5:320, June 2018.

[136] Daniel Machado, Rafael S Costa, Miguel Rocha, Eugenio C Ferreira, Bruce Tidor, and Isabel Rocha. Modeling formalisms in Systems Biology. *AMB Express*, 1(1):45, December 2011.

[137] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[138] Stefanía Magnúsdóttir, Almut Heinken, Laura Kutt, Dmitry A Ravcheev, Eugen Bauer, Alberto Noronha, Kacy Greenhalgh, Christian Jäger, Joanna Baginska, Paul Wilmes, Ronan M T Fleming, and Ines Thiele. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1):81–89, January 2017.

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

[140] Frédéric Mahé, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3(2):e1420, 2015.

[141] Noël Malod-Dognin and Nataša Pržulj. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics (Oxford, England)*, 31(13):2182–2189, July 2015.

[142] Dinka Mandakovic, Claudia Rojas, Jonathan Maldonado, Mauricio Latorre, Dante Travisany, Erwan Delage, Audrey Bihouée, Géraldine Jean, Francisca P Díaz, Beatriz Fernández-Gómez, Pablo Cabrera, Alexis Gaete, Claudio Latorre, Rodrigo A Gutiérrez, Alejandro Maass, Verónica Cambiazo, Sergio A Navarrete, Damien Eveillard, and Mauricio González. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific reports*, 8(1):5875, April 2018.

[143] Glenn Marion, Greg J McInerny, Jörn Pagel, Stephen Catterall, Alex R Cook, Florian Hartig, and Robert B O'Hara. Parameter and uncertainty estimation for process-oriented population and distribution models: data, statistics and the niche. *Journal of Biogeography*, 39(12):2225–2239, 2012.

[144] Kim Marriott, Peter J Stuckey, and Peter J Stuckey. *Programming with constraints: an introduction*. MIT press, 1998.

[145] Susan F Martinez, Axelle Renodon-Cornière, Julian Nomme, Damien Eveillard, Fabrice Fleury, Masayuki Takahashi, and Pierre Weigel. Targeting human Rad51 by specific DNA aptamers induces inhibition of homologous recombination. *Biochimie*, 92(12):1832–1838, December 2010.

[146] Alix Mas, Shahrad Jamshidi, Yvan Lagadeuc, Damien Eveillard, and Philippe Vandenkoornhuyse. Beyond the Black Queen Hypothesis. *The ISME Journal*, 10(9):2085–2091, September 2016.

[147] Folker Meyer, Daniel Paarmann, Mark D'Souza, Robert Olson, Elizabeth M Glass, Michael Kubal, Tobias Paczian, Alex Rodriguez, Rick Stevens, Andreas Wilke, et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.

[148] Shira Mintz-Oron, Sagit Meir, Sergey Malitsky, Eytan Ruppin, Asaph Aharoni, and Tomer Shlomi. Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 109(1):339–344, January 2012.

[149] Thomas Mock, Robert P Otillar, Jan Strauss, Mark McMullan, Pirita Paajanen, Jeremy Schmutz, Asaf Salamov, Remo Sanges, Andrew Toseland, Ben J Ward, Andrew E Allen, Christopher L Dupont, Stephan Frickenhaus, Florian Maumus, Alaguraj Veluchamy, Taoyang Wu, Kerrie W Barry, Angela Falciatore, Maria I Ferrante, Antonio E Fortunato, Gernot Glöckner, Ansgar Gruber, Rachel Hipkin, Michael G Janech, Peter G Kroth, Florian Leese, Erika A Lindquist, Barbara R Lyon, Joel Martin, Christoph Mayer, Micaela Parker, Hadi Quesneville, James A Raymond, Christiane Uhlig, Ruben E Valas, Klaus U Valentin, Alexandra Z Worden, E Virginia Armbrust, Matthew D Clark, Chris Bowler, Beverley R Green, Vincent Moulton, Cock van Oosterhout, and Igor V Grigoriev. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature*, 541(7638):536–540, January 2017.

[150] Jacques Monod. The growth of bacterial cultures. *Annual Reviews in Microbiology*, 3(1):371–394, 1949.

[151] Jacques Monod. La technique de culture continue. th´orie et applications. *Ann. Inst. Pasteur*, 79:390–410, 1950.

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

[153] Nicolas Mouquet, Yvan Lagadeuc, Vincent Devictor, Luc Doyen, Anne Duputié, Damien Eveillard, Denis Faure, Eric Garnier, Olivier Gimenez, Philippe Huneman, Franck Jabot, Philippe Jarne, Dominique Joly, Romain Julliard, Sonia Kefi, Gael J Kergoat, Sandra Lavorel, Line Le Gall, Laurence Meslin, Serge Morand, Xavier Morin, Hélène Morlon, Gilles Pinay, Roger Pradel, Frank M Schurr, Wilfried Thuiller, and Michel Loreau. REVIEW: Predictive ecology in a changing world. *Journal of Applied Ecology*, 52(5):1293–1310, July 2015.

[154] Silvia E Newell, Damien Eveillard, Mark J McCarthy, Wayne S Gardner, Zhanfei Liu, and Bess B Ward. A shift in the archaeal nitrifier community in response to natural and anthropogenic disturbances in the northern Gulf of Mexico. *Environmental Microbiology Reports*, 6(1):106–112, February 2014.

[155] H Bjørn Nielsen, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Damian R Plichta, Laurent Gautier, Anders G Pedersen, Emmanuelle Le Chatelier, Ida Bonde, Trine Nielsen, Chaysavanh Manichanh, Manimozhiyan Arumugam, Jean-Michel Batto, Marcelo B Quintanilha dos Santos, Nikolaj Blom, Natalia Borruel, Kristoffer S Burgdorf, Fouad Boumezbeur, Francesc Casellas, Joël Doré, Piotr Dworzynski, Francisco Guarner, Torben Hansen, Falk Hildebrand, Rolf S Kaas, Sean Kennedy, Karsten Kristiansen, Pierre Léonard, Florence Levenez, Ole Lund, Bouziane Moumen, Denis Le Paslier, Nicolas Pons, Oluf Pedersen, Edi Prifti, Junjie Qin, Jeroen Raes, Søren Sørensen, Julien Tap, Sebastian Tims, David W Ussery, Pierre Renault, Thomas Sicheritz-Ponten, Peer Bork, Jun Wang, Søren Brunak, Shinichi Sunagawa, S Dusko Ehrlich, MetaHIT Consortium, Eric Pelletier, Jens Roat Kultima, and Takuji Yamada. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, August 2014.

[156] Philippe Normand, Robert Duran, Xavier Le Roux, Cindy Morris, and Jean-Christophe Poggiale. Biodiversity and microbial ecosystems functioning. In *Environmental microbiology: fundamentals and applications*, pages 261–291. Springer, 2015.

[157] Prianka V Patel, Tara A Gianoulis, Robert D Bjornson, Kevin Y Yip, Donald M Engelman, and Mark B Gerstein. Analysis of membrane proteins in metagenomics: networks of correlated environmental features and protein families. *Genome Research*, 20(7):960–971, July 2010.

[158] Octavio Perez-Garcia, Gavin Lear, and Naresh Singhal. Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Frontiers in microbiology*, 7:673, 2016.

[159] Marc Picheral, Lionel Guidi, Lars Stemmann, David M Karl, Ghizlaine Iddaoud, and Gabriel Gorsky. The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods*, 8(9):462–473, September 2010.

[160] Cristian Picioreanu, Mark CM Van Loosdrecht, and Joseph J Heijnen. Mathematical modeling of biofilm structure with a hybrid differential-discrete cellular automaton approach. *Biotechnology and bioengineering*, 58(1):101–116, 1998.

[161] Frédéric Plewniak, Sandrine Koechler, Benjamin Navet, Eric Dugat-Bony, Olivier Bouchez, Pierre Peyret, Fabienne Séby, Fabienne Battaglia-Brunet, and Philippe N Bertin. Metagenomic insights into microbial metabolism affecting arsenic dispersion in Mediterranean marine sediments. *Molecular ecology*, 22(19):4870–4883, October 2013.

[162] Nathan D Price, Iman Famili, Daniel A Beard, and Bernhard Ø Palsson. Extreme pathways and kirchhoff's second law. *Biophysical journal*, 83(5):2879–2882, 2002.

[163] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, November 2004.

[164] Sylvain Prigent, Clémence Frioux, Simon M Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, Frédéric Plewniak, Thierry Tonon, and Anne Siegel. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. *PLoS computational biology*, 13(1):e1005276, January 2017.

[165] Jeroen Raes and Peer Bork. Molecular eco-systems biology: towards an understanding of community function. *Nature Reviews Microbiology*, 6(9):693–699, September 2008.

[166] Jeroen Raes, Ivica Letunic, Takuji Yamada, Lars Juhl Jensen, and Peer Bork. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Molecular Systems Biology*, 7:473, March 2011.

[167] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, 2009.

[168] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *Science (New York, NY)*, 334(6062):1518–1524, December 2011.

[169] Adrien Richard and Jean-Paul Comet. Necessary conditions for multistationarity in discrete dynamical systems. *Discrete Applied Mathematics*, 155(18):2403–2413, 2007.

[170] Tammi L Richardson and George A Jackson. Small phytoplankton and carbon export from the surface ocean. *Science (New York, NY)*, 315(5813):838–840, February 2007.

[171] Francisco Rodriguez-Valera. Environmental genomics, the big picture? *FEMS Microbiology Letters*, 231:153–158, 2004.

[172] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, November 2017.

[173] PR Romero and P Karp. Nutrient-related analysis of pathway/genome databases. In *Biocomputing 2001*, pages 471–482. World Scientific, 2000.

[174] Delphine Ropers, Hidde de de Jong, Michel Page, Dominique Schneider, and Johannes Geiselmann. Qualitative simulation of the carbon starvation response in Escherichia coli. *BioSystems*, 84(2):124–152, May 2006.

[175] Quansong Ruan, Debojyoti Dutta, Michael S Schwalbach, Joshua A Steele, Jed A Fuhrman, and Fengzhu Sun. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics (Oxford, England)*, 22(20):2532–2538, October 2006.

[176] Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, Karen Beeson, Bao Tran, Hamilton Smith, Holly Baden-Tillson, Clare Stewart, Joyce Thorpe, Jason Freeman, Cynthia Andrews-Pfannkoch, Joseph E Venter, Kelvin Li, Saul Kravitz, John F Heidelberg, Terry Utterback, Yu-Hui Rogers, Luisa I Falcón, Valeria Souza, Germán Bonilla-Rosso, Luis E Eguiarte, David M Karl, Shubha Sathyendranath, Trevor Platt, Eldredge Bermingham, Victor Gallardo, Giselle Tamayo-Castillo, Michael R Ferrari, Robert L Strausberg, Kenneth Nealson, Robert Friedman, Marvin Frazier, and J Craig Venter. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, 5(3):e77, March 2007.

[177] Wouter Saelens, Robrecht Cannoodt, and Yvan Saeys. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications*, 9(1):1090, March 2018.

[178] Tanvir Sajed, Ana Marcu, Miguel Ramirez, Allison Pon, An Chi Guo, Craig Knox, Michael Wilson, Jason R Grant, Yannick Djoumbou, and David S Wishart. Ecmdb 2.0: A richer resource for understanding the biochemistry of e. coli. *Nucleic acids research*, 44(D1):D495–D501, 2015.

[179] Emanuel Salazar-Cavazos and Moisés Santillán. Optimal performance of the tryptophan operon of e. coli: A stochastic, dynamical, mathematical-modeling approach. *Bulletin of mathematical biology*, 76(2):314–334, 2014.

[180] Erwin Schrodinger. *What is life?* University Press: Cambridge, 1943.

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

[182] Stefan Schuster, David Fell, T Pfeiffer, Thomas Dandekar, and Peter Bork. Elementary modes analysis illustrated with human red cell metabolism. In *BioThermoKinetics in the Post Genomic Era*, pages 332–339. Göteborg, 1998.

[183] Ilya Shmulevich, Edward R Dougherty, Seungchan Kim, and Wei Zhang. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics (Oxford, England)*, 18(2):261–274, February 2002.

[184] Heike Siebert and Alexander Bockmayr. Temporal constraints in the logical analysis of regulatory networks. *Theoretical Computer Science*, 391(3):258–275, 2008.

[185] E Simão, E Remy, Denis Thieffry, and C Chaouiya. Qualitative modelling of regulated metabolic pathways: application to the tryptophan biosynthesis in E.coli. *Bioinformatics (Oxford, England)*, 21 Suppl 2:ii190–6, September 2005.

[186] Adam Alexander Thil Smith, Eugeni Belda, Alain Viari, Claudine Medigue, and David Vallenet. The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. *PLoS computational biology*, 8(5):e1002540, May 2012.

[187] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big Data: Astronomical or Genomical? *PLoS biology*, 13(7):e1002195, July 2015.

[188] Sergey Stolyar, Steve Van Dien, Kristina Linnea Hillesland, Nicolas Pinel, Thomas J Lie, John A Leigh, and David A Stahl. Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, 3:92, 2007.

[189] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, Francisco M Cornejo-Castillo, Paul I Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Sabrina Speich, Lars Stemmann, Matthew B Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G Acinas, and Peer Bork. Ocean plankton. Structure and function of the global ocean microbiome. *Science (New York, NY)*, 348(6237):1261359–1261359, May 2015.

[190] T Surrey, F Nédélec, S Leibler, and E Karsenti. Physical properties determining self-organization of motors and microtubules. *Science (New York, NY)*, 2001.

[191] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

[192] Reed Taffs, John E Aston, Kristen Brileya, Zackary Jay, Christian G Klatt, Shawn McGlynn, Natasha Mallette, Scott Montross, Robin Gerlach, William P Inskeep, David M Ward, and Ross P Carlson. In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC systems biology*, 3(1):114, 2009.

[193] Ines Thiele and Bernhard Ø Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93–121, January 2010.

[194] Stanislas Thiriet-Rupert, Gregory Carrier, Camille Trottier, Damien Eveillard, Benoit Schoefs, Gael Bougaran, Jean-Paul Cadoret, Benoit Chénais, and Bruno Saint-Jean. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of Tisochrysis lutea. *Algal Research*, 30:59–72, March 2018.

[195] François Thomas, Philippe Bordron, Damien Eveillard, and Gurvan Michel. Gene Expression Analysis of Zobellia galactanivorans during the Degradation of Algal Polysaccharides Reveals both Substrate-Specific and Shared Transcriptome-Wide Responses. *Frontiers in microbiology*, 8:1808, 2017.

[196] R Thomas. Laws for the dynamics of regulatory networks. *The International journal of developmental biology*, 42(3):479–485, 1998.

[197] René Thomas, Denis Thieffry, and Marcelle Kaufman. Dynamical behaviour of biological regulatory networks—i. biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bulletin of mathematical biology*, 57(2):247–276, 1995.

[198] David Tilman. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences*, 101(30):10854–10861, July 2004.

[199] Laura Tipton, Christian L Müller, Zachary D Kurtz, Laurence Huang, Eric Kleerup, Alison Morris, Richard Bonneau, and Elodie Ghedin. Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, 6(1):12, January 2018.

[200] Leïla Tirichine and Chris Bowler. Decoding algal genomes: tracing back the history of photosynthetic life on earth. *The Plant Journal*, 66(1):45–57, 2011.

[201] Thierry Tonon and Damien Eveillard. Marine systems biology. *Frontiers in genetics*, 6(20):181, 2015.

[202] Thierry Tonon, Damien Eveillard, Sylvain Prigent, Jérémie Bourdon, Philippe Potin, Catherine Boyen, and Anne Siegel. Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment. *Omics : a journal of integrative biology*, 15(12):883–892, December 2011.

[203] Alan Mathison Turing. The chemical basis of morphogenesis. *Phil. Trans. R. Soc. Lond. B*, 237(641):37–72, 1952.

[204] David Vallenet, Laurent Labarre, Zoe Rouy, Valerie Barbe, Stephanie Bocs, Stephane Cruveiller, Aurelie Lajus, Geraldine Pascal, Claude Scarpelli, and Claudine Medigue. Mage: a microbial genome annotation system supported by synteny results. *Nucleic acids research*, 34(1):53–65, 2006.

[205] Amit Varma and Bernhard Ø Palsson. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology*, 12(10):994–998, September 1994.

[206] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jeff M Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, NY)*, 304(5667):66–74, April 2004.

[207] Emilie Villar, Gregory K Farrant, Michael Follows, Laurence Garczarek, Sabrina Speich, Stéphane Audic, Lucie Bittner, Bruno Blanke, Jennifer R Brum, Christophe Brunet, Raffaella Casotti, Alison Chase, John R Dolan, Fabrizio d'Ortenzio, Jean-Pierre Gattuso, Nicolas Grima, Lionel Guidi, Christopher N Hill, Oliver Jahn, Jean-Louis Jamet, Hervé Le Goff, Cyrille

Lepoivre, Shruti Malviya, Eric Pelletier, Jean-Baptiste Romagnan, Simon Roux, Sébastien Santini, Eleonora Scalco, Sarah M Schwenck, Atsuko Tanaka, Pierre Testor, Thomas Vannier, Flora Vincent, Adriana Zingone, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Silvia G Acinas, Peer Bork, Emmanuel Boss, Colomban de Vargas, Gabriel Gorsky, Hiroyuki Ogata, Stéphane Pesant, Matthew B Sullivan, Shinichi Sunagawa, Patrick Wincker, Eric Karsenti, Chris Bowler, Fabrice Not, Pascal Hingamp, and Daniele Iudicone. Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science (New York, NY)*, 348(6237):1261447–1261447, May 2015.

[208] Eberhard O Voit. *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*, volume 1. Cambridge University Press, 2000.

[209] Bess B Ward, Damien Eveillard, Julie D Kirshtein, Joshua D Nelson, Mary A Voytek, and George A Jackson. Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology*, 9(10):2522–2538, October 2007.

[210] Bess B Ward, Damien Eveillard, Julie D Kirshtein, Joshua D Nelson, Mary A Voytek, and George A Jackson. Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology*, 9(10):2522–2538, 2007.

[211] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, Amanda Birmingham, Jacob A Cram, Jed A Fuhrman, Jeroen Raes, Fengzhu Sun, and Rob Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669–1681, July 2016.

[212] Hans V Westerhoff and Bernhard Ø Palsson. The evolution of molecular biology into systems biology. *Nature Biotechnology*, 22(10):1249–1252, October 2004.

[213] William B Whitman, David C Coleman, and William J Wiebe. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, 1998.

[214] Sharon Wiback and Bernhard Ø Palsson. Extreme Pathway Analysis of Human Red Blood Cell Metabolism. *Biophysical journal*, 83:808–818, August 2002.

[215] Timothy Wyatt. Margalef's mandala and phytoplankton bloom strategies. *Deep Sea Research Part II: Topical Studies in Oceanography*, 101:32–49, 2014.

[216] Jun Yan, Haifang Wang, Yuting Liu, and Chunxuan Shao. Analysis of gene regulatory networks in the mammalian circadian rhythm. *PLoS computational biology*, 4(10):e1000193, October 2008.

[217] Jonathan P Zehr and Raphael M Kudela. Nitrogen cycle of the open ocean: from genes to ecosystems. *Annual review of marine science*, 3:197–225, 2011.

[218] Aleksej Zelezniak, Sergej Andrejev, Olga Ponomarova, Daniel R Mende, Peer Bork, and Kiran Raosaheb Patil. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 112(20):6449–6454, May 2015.

[219] Karsten Zengler and Bernhard Ø Palsson. A road map for the development of community systems (CoSy) biology. *Nature Reviews Microbiology*, 10(5):366–372, May 2012.

[220] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[221] Jun Zhao, Alistair Miles, Graham Klyne, and David Shotton. Linked data and provenance in biological data webs. *Briefings in bioinformatics*, 10(2):139–152, 2008.

[222] Ali R Zomorrodi, Mohammad Mazharul Islam, and Costas D Maranas. d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology*, 3(4):247–257, April 2014.

[223] Ali R Zomorrodi and Costas D Maranas. Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS computational biology*, 8(2):e1002363, 2012.

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

Chapter 6

Extended *Curriculum Vitae*

Damien Eveillard

Maitre de Conférences

2 rue de la Houssinière
44322 Nantes
France
+33 (0)6 16 09 59
+33 (0)2 51 12 59 85
FAX +33 (0)2 51 12 58 12
damien.eveillard@univ-nantes.fr
http://pagesperso.lina.univ-nantes.fr/~eveillard-d
eveillard

Expérience Professionnelle

2006 – ... **Maitre de Conférences**, *Université de Nantes*, LS2N UMR 6004.
Membre de l'équipe ComBi

2004 – 2006 **Chercheur Post-doctorant**, *Texas A&M University*, Environmental Modeling Group.
Travail sous la direction de George A. Jackson (Texas A&M University) et Bess B. Ward (Princeton University)

2000 – 2004 **Ingénieur d'études - doctorant**, *Université de Nancy 1*, LORIA - INRIA Lorraine.
Membre de l'équipe MODBIO, sous la direction d'Alexander Bockmayr
et de Christiane Branlant

Habilitation à Diriger des Recherches

Spécialité *Bioinformatique*

Jury Philippe Vandenkoornhuyse (Prof U. Rennes), Alexander Bockmayr (Prof. Freie Universität Berlin), Jérémie Bourdon (Prof. U. Nantes), Karoline Faust (Ass. Prof. KU Leuven), Christopher Quince (Ass. Prof U. Warwick), Claudine Medigue (DR CNRS) et Eric Rival (DR CNRS)

Titre From Systems Biology to Systems Ecology: a computational journey

Univ. Nantes soutenue le 13 octobre 2020

Doctorat

Spécialité *Analyse et Modélisation des Systèmes Biologiques*

Encadrants Alexander Bockmayr (Prof. Freie Universität Berlin) et Christiane Branlant (DR CNRS)

Titre Modélisation statistique et formelle de la régulation de l'épissage alternatif chez HIV-1

Univ. Nancy soutenue le 14 mai 2004

Diplome d'Etudes Approfondies

Spécialité *Océanologie Biologique*

Encadrants Antoine Sciandra (DR CNRS) et Olivier Bernard (DR INRIA)

Titre Modélisation de l'effet de la limitation conjuguée de la lumière et de l'azote sur la croissance autotrophe

Univ. Paris 6 juin 2000, mention AB

Maitrise de Biologie des Populations

Spécialité *Océanographie physique et écologie marine*

Univ. Paris 6 juin 1999, mention AB

Distinctions et rayonnement international

- PEDR (2015 - 2024) du ministère de la Recherche et de l'Enseignement Supérieur
- Lauréat NSF (USA National Science Fundation) : *"Integrative Biology and adaptation of antarctic marine organisms"* (sélection de 24 dossiers sur 480). Mission en Antarctique (décembre 2009 - février 2010) qualifiante pour les missions polaires et pour l'étude des systèmes microbiens dans leur environnement
- Prime de l'Excellence Scientifique (2010 - 2014) du ministère de la Recherche et de l'Enseignement Supérieur
- Lauréat du Programme Initiative Post-Doc (2006) : Modélisation des systèmes biologiques avec la programmation par contraintes

Responsabilités Scientifiques

- Responsable de l'équipe ComBi du LS2N (2019 –)

Participations à des groupements de recherche

- Membre du comité de direction de la Fédération de Recherche CNRS TARA Oceans Systems Ecology & Evolution (GO-SEE) FR2022 (2018 – 2023)
- Co-Responsable avec Philippe Vandenkoornhuyse du groupe de travail Ecologie des Systèmes pour Biogenouest (2017 - 2018)
- Membre du Bureau du Groupe de Recherche (GDR) Génomique Environnementale (2014 – 2018)
- Membre du Réseau Thématique Prioritaire (RTP) du CNRS en Génomique Environnementale (depuis 2012), coordinateur de la prospective *modélisation et données NGS*
- Co-Responsable avec Charles Pineau du groupe de travail Génomique Intégrative pour Biogenouest (mars - décembre 2011)

Mobilités

USA Dept of Energy, Lawrence Berkeley National Laboratory, 17 jours, février 2018

Chili Center for Mathematical Modeling - Universidad de Chile, 10 jours, septembre 2017

USA Dept of Energy, Lawrence Berkeley National Laboratory, 13 jours, octobre 2016

USA Dept of Energy, Lawrence Berkeley National Laboratory, 12 jours, décembre 2015

USA Dept of Energy, Lawrence Berkeley National Laboratory, 14 jours, mai 2014

Chili Center for Mathematical Modeling - Universidad de Chile, 15 jours, décembre 2013

Chili Center for Mathematical Modeling - Universidad de Chile, 19 jours, décembre 2012

Chili	Center for Mathematical Modeling - Universidad de Chile, 17 jours, avril 2011
USA	Bodega Bay Marine Laboratory - UC Davis, 15 jours, avril 2010
Japon	National Institute of Informatics (NII) - Tokyo, 13 jours mars 2010
Antarctique	Mc Murdo Scientific Station, USAP, 6 semaines, janvier 2010
USA	Department of Ecology and Evolutionary Biology, Princeton University, 11 jours, juin 2009
Japon	National Institute of Informatics (NII) - Tokyo, 10 jours mars 2009
UK	Marine Biology Laboratory, University of Plymouth, 10 jours, août 2008
Allemagne	DFG Research - Matheon, Freie Universität Berlin - (juillet 2006)
USA	Membre du projet *Biocomplexity* (septembre 2004 à août 2006) dans le groupe relatif à la modélisation et à l'analyse des données. Dans ce cadre, j'ai été amené à différents séjours entre Texas A&M University sous la direction de George A. Jackson et Princeton University en collaboration avec Bess B. Ward.

Participation à des projets scientifiques

Projets internationaux

- Projet H2020 ATLANTECO, membre du consortium (2020 – 2024)
- Tara Magallanes – co-porteur de la campagne océanographique sous l'égide de la Fondation Tara Océan (2020 – 2021)
- CNRS PICS France-Berkeley – Responsable du projet EMBASSY en collaboration avec Lawrence Berkeley National Laboratory (2014 – 2017)
- TARA Oceans – Membre du Consortium TARA Oceans depuis 2015
- IntegrativeBioChile – Membre de l'équipe associée (2011 - 2014), (2014 - 2018) en collaboration avec le Centro de Modialemento Matematico, Chile Universidad
- INRIA - CONYCIT – Membre du programme de collaboration internationale (2011 - 2012), membre
- CONICYT – Membre de projet financé par la Commission Nationale de la Recherche Scientifique et Technologique du Chili (2010 - 2012), membre
- PHC Sakura – avec le NII (Japon), et Waseda University (2008 - 2010): *Hybrid Constraints Programming for biological systems*, membre
- PHC Procope – avec Université de Nice et Freie Universität Berlin (2008 - 2010): *Temporal properties of discrete biological models*, co-porteur
- Projet NSF *BioComplexity* – Membre du projet sous la direction de Bess B. Ward, Princeton University (2000 - 2008)

Participation à des projets de recherche nationaux

- MEGALODOM - projet interdisciplinaire MASTODOM (2017 - ...), co-porteur de projet avec Lionel Guidi (CR CNRS - Villefranche-sur-Mer)
- CINAMON (2018 – 2022) porté par Laurence Garzareck (DR CNRS - Roscoff), Responsable d'un Workpackage Modélisation de la physiologie de *Synnechococus sp.*

- ANR IMPEKAB (2016 – 2020) porté par Fabrice Not (CR CNRS - Roscoff), Responsable d'un Workpackage Modélisation de la symbiose chez les radiolaires
- ANR SAMOSA (2014 – 2017) porté par Laurence Garzareck (DR CNRS - Roscoff), Responsable d'un Workpackage Modélisation de la physiologie de *Synnechococus sp.*
- Projet Structurant CNRS EC2CO, Commerce (2013 – 2014), co-responsable avec Fredéric Plewniak, modélisation du métabolisme d'un écosystème microbien
- Projet Structurant Project Lab INRIA *Algae In Silico* : modélisation de la réponse photosynthétique des phytoplancton avec objectif de valorisation industrielle (production de biocarburant). Projet porté par Olivier Bernard (DR INRIA - Sophia Antipolis)
- PEPII AQUASYST (2011 – 2012) : modélisation des systèmes microbiens des nappes phréatiques. Co-Responsable avec Alexis Dufresne.
- PEPS MANIFOLD (2011 – 2012) coordonné par A. Goldstejn, membre
- Investissement d'avenir IDEALG (2011 – 2021) membre associé au projet : reconstruction et modélisation du métabolisme des algues brunes soumis à un stress abiotique
- ANR Blanche (2010 – 2014) : BIOTEMPO Langages, concepts de temps et modèles hybrides pour l'analyse de modèles incomplets en biologie moléculaire. Coordonné par A. Siegel. Responsable du Workpackage Intégration de données
- ANR SYSTERRA (2010 – 2014) : ECS Evolution du comportement de coopération plante-symbiontes dans la perspective d'un usage étendu en agriculture écologiquement intensive. Coordonné par P. Vandenkoornhuyse. Responsable d'un Workpackage Modélisation
- PEPS QuantOursin (2010 – 2011) : Modélisation quantitative de l'initiation de traduction de l'oursin. Coordonné par A. Siegel, membre
- Membre de l'ARC INRIA Calcul de Processus et Biologie Moléculaire (CPBio) (2002 – 2004)

Participation à des projets de recherche (bi)régionaux

- PROLIFIC (2020 – 2024). Membre du Consortium
- MIBIOGATE - Région Pays de la Loire (2017 – 2021). Membre du comité de pilotage
- ProBioSTIC Projet RFI Numerique - Région Pays de la Loire (2017 – 2019), porteur de projet
- ProBioSTIC - projet interdisciplinaire de l'Université de Nantes (2017), porteur de projet
- ECOSYST (2016 - 2018) - projet structurant de BioGenOuest. Co-porteur avec Philippe Vandenkoornhuyse (Pr Université de Rennes 1)
- Projet Régional Structurant GRIOTE (2013 – 2017), coordonné par Jérémie Bourdon, Richard Redon & Dominique Tessier, membre du comité scientifique, animateur du pôle valorisation.

- Projet de la fédération de recherche ATLANSTIC: BioATLANSTIC pour une collaboration avec l'École Centrale de Nantes sur la modélisation des propriétés temporelles des systèmes biologiques (2006 - 2008)
- Projet régional structurant Bioinformatique ligérienne (BIL) (2006-2010) coordonné par R. Houlgatte

Enseignement et responsabilités pédagogiques

Je suis Maître de Conférences au département d'Informatique de l'Université de Nantes depuis 2006 avec une *moyenne de 243 heures ETD d'enseignement par an*. Les domaines enseignés couvrent les Bases de Données, Langage Script, Langage Objet, initiation à l'algorithmique pour un public d'étudiants en Informatique, mais aussi la bioinformatique pour des étudiants en biologie. J'interviens également en Master Bioinformatique pour enseigner la Bioanalyse, les scripts Perl en Bioinformatique, la gestion de projets.

- Co-porteur du parcours Master 2 *Génétique, Génomique, Biologie des Systèmes* du Master Biologie-Santé de l'Université de Nantes (∼14 étudiants) depuis 2017
- responsable de Licence 1 à l'Université de Nantes (∼860 étudiants et 23 enseignants) de 2009 à 2012, et depuis 2014
- responsable du module initiation à l'informatique pour le portail Biologie Géologie Chimie de Licence 1 (∼820 étudiants et 16 intervenants de TD) depuis 2008
- Co-responsable du module géologie quantitative - Licence 1(∼ 36 étudiants) 2008 – 2009
- Co-responsable du module informatique pour la Chimie - Licence 2 (∼ 36 étudiants) depuis 2009
- Responsable du module langage Script pour la Bionformatique Master 2 Bioinformatique (∼15 étudiants) depuis 2006
- Responsable du module développement de projet en Bionformatique Master 1 Bioinformatique (∼ 36 étudiants) depuis 2006
- Co-responsable du module bioanalyse Master 2 Bioinformatique (∼ 25 étudiants) depuis 2006
- Intervenant dans le module de biologie quantitative – option BioSTIC de l'École Centrale de Nantes (∼ 25 étudiants) depuis 2016

Par ailleurs, je suis intervenu dans d'autres formations

- Université de Nancy 1 (2001 - 2004) : 123 heures ETD sur 3 ans dans le DESS Ressources Génomiques et Traitements Informatiques (Programmation en C, Bioanalyse et Modélisation Statistique)
- Texas A&M University (2005 - 2006) : interventions ponctuelles pour enseigner la modélisation des systèmes biologiques pour des *graduate students in Marine microbiology*

Encadrement d'activités de recherche

Encadrement Doctoral

- Jakez Rolland - Bourse CIFRE (mai 2021 –), étudiant en thèse à l'université de Nantes et salarié de la société Bio-Logbook, encadré à 50% avec Benoit Delahaye (LS2N): *Science des données pour la médecine personalisée*
- Marinna Gaudin - Bourse H2020/Région Pays de la Loire (novembre 2020 –), étudiant en thèse à l'université de Nantes, encadré à 50% avec Samuel Chaffron (LS2N): *Modélisation des interactions planctoniques*
- Anna Lambert - Bourse projet PROLOFIC (novembre 2020 –), étudiant en thèse à l'université de Nantes, encadré à 50% avec Samuel Chaffron (LS2N): *Modélisation du microbiote intestinale en inyteraction avec la paroi intestinale*
- Antoine Regimbeau - Bourse CNRS 80 Prime (octobre 2019 –), étudiant en thèse à l'université de Nantes, encadré à 60% avec Laurent Memery (DR CNRS, LEMAR) et & Olivier Aumont (DR IRD, LOCEAN): *Introduction de la Complexité planctonique dans les modèles biogéochimiques*
- Ulysse Guyet - Bourse Paris Universitat / Region Bretagne (octobre 2016 – juin 2020), étudiant en thèse à l'université de Paris 6, encadré à 50% avec Laurence Garkzarek (DR CNRS, Roscoff): *Modélisation de la réponse transcriptomique de Synechococcus sp. aux stress abiotiques*
- Julien Fradin - Bourse MESR (depuis octobre 2015), étudiant en thèse à l'Université de Nantes, encadré à 30% avec Guillaume Fertin (Pr Univ. Nantes), et Géraldine Jean (MCU Univ. Nantes): *Comparaison de graphes en Biologie*
- Julie Laniau - Bourse INRIA (Octobre 2013 – Octobre 2017), étudiante en thèse à l'Université de Rennes, encadré à 50% avec Anne Siegel (DR CNRS) : *Reconstruction du métabolisme fonctionnel des communautés*
- Marko Budinich - bourse CNRS / région Pays de la Loire (octobre 2013 - avril 2017), étudiant en thèse à l'Université de Nantes, encadré à 50% avec Jérémie Bourdon (Pr. Université de Nantes), inscrit à l'école doctorale Sciences et Technologies de l'Information et de Mathématiques : *Modélisation du métabolisme des écosystèmes microbiens : vers une modélisation multi-objectif du métabolisme*
- Philippe Bordron - Bourse ministérielle (septembre 2008 - juin 2012), étudiant en thèse à l'Université de Nantes, encadré à 50% avec Irena Rusu (Pr. Université de Nantes), inscrit à l'école doctorale Sciences et Technologies de l'Information et de Mathématiques : *Utilisation de la théorie des graphes pour l'intégration de données omics*.
- Etienne Zohim - Bourse de la Ligue Contre le Cancer (décembre 2008 - juillet 2011), étudiant en thèse à l'Université de Compiègne, encadré à 50% avec Abalo Chango (Pr. Institut LaSalle Beauvais), inscrit à l'école doctorale de Biologie-Santé : *Biologie des Systèmes pour la modélisation du métabolisme des monocarbones*.

Etudiants de Master

- Adrien Bonneterre (mars 2017 - juillet 2017) étudiant de Master de Bioinformatique de l'Université de Nantes: *Intégration des données de génomique environnementale et satelitaires*
- Julie Haguait (mars 2015 - septembre 2016) étudiante de Master Bioinformatique de l'Université de Nantes: *Analyse des données transcriptomique de Synnechococcus sp.*

- Soizic Auffray (mars 2015 - juin 2015) étudiante de Master Bioinformatique de l'Université de Nantes: *Analyse de la réponse transcriptomique de Synnechococcus sp. aux stress abiotiques*
- Sebastien Charneau (mars 2014 - juin 2014) étudiant de Master Bioinformatique de l'Université de Nantes: *Intégration des données transcriptomiques dans les modèles métaboliques*
- Géraldine Del Mondo (février 2007 - décembre 2007) étudiante de Master Informatique de l'Université de Nantes, encadrée à 50% avec Irena Rusu (Pr. Université de Nantes): *décomposition homogène des graphes d'interactions biologiques*.
- Myriam Vezain (septembre 2002 - juin 2004), étudiante en DESS de bioinformatique EGOIST de Rouen, co-encadrée avec A. Bockmayr : *Modélisation de la régulation l'épissage alternatif au site A7 de HIV-1*.
- Deo Pakrash Pandey (juin 2002 - août 2002), étudiant en dernier cycle d'ingénieur en informatique de l'université de Kanpur (Inde), actuellement étudiant en thèse de biologie moléculaire University of Southern Denmark (SDU) : *Modeling with hybrid constraints the HIV-1 life cycle*.
- Stéphanie Billaut (janvier 2003 - mars 2003), étudiante en Maîtrise de Biologie Cellulaire et Génétique de Nancy 1 : *Recherche de motifs de régulation de l'épissage alternatif : apprentissage statistique sur HIV-1*. (février 2004 - septembre 2004) étudiante en Master de Bioinformatique de Bordeaux: modélisation du métabolisme de la vitamine B12, approche topologique, co-encadrée avec A. Bockmayr.
- Sébastien Vachenc (juin 2002 - février 2003), étudiant en DESS RGTI à Nancy 1, actuellement Ingénieur attaché au département bioinformatique des laboratoires Fournier : *Combinaison d'approches statistiques pour la découverte de motifs de régulation d'épissage*.
- Carole Dossat (octobre 2001 - février 2002), étudiante en DESS de Génomique RGTI à Nancy 1, actuellement Ingénieur au Génoscope : *Comparaison et coordination des algorithmes d'alignements dédiés aux séquences nucléiques*.

Participation à des jurys de thèse

- Examinateur de la thèse de Simon Ramondenc *Analyse des variations spatio-temporelles du zooplancton gélatineux et son effet sur les flux de matières à l'aide d'une approche combinant expérimentation et écologie numérique*. Sorbonne Universités, 24 Novembre 2017
- Rapporteur de la thèse de Dhivyaa Rajasundaram *Integrative analysis of heterogeneous plant cell wall related data using canonical- and network- based approaches*. Potsdam Universitat, 4 juin 2015.
- Examinateur de la thèse de Gael Bougaran, *Co-limitation par l'azote et le phosphore : étude des mécanismes chez la microalgue Tisochrysis lutea*. Université de Nantes, le 24 octobre 2014. Thèse de l'école doctorale VENAM
- Examinateur de la thèse de David Thybert, *Identification des potentialités fonctionnelles dans les génomes procaryotes : Application au sous-système de détoxication des radicaux libres de l'oxygène et de l'azote*. Université de Rennes 1, le 8 juillet 2010. Thèse de l'école doctorale Vie Agro-Santé.

- Examinateur de la thèse de Jamil Ahmad, *Modélisation Hybride et analyse des dynamiques des réseaux de régulations biologiques en tenant compte des délais*. École Centrale de Nantes, le 5 février 2009. Thèse de l'école doctorale Sciences et Technologies de l'Information et Mathématiques.

Participation à des comités de thèse

- Clarisse Lemonier (2016 – 2017) doctorat d'Océanographie Biologique, encadrée par Lois Magnien, Université de Brest
- Alix Mas (2015 – 2016) doctorat en Ecologie, encadrée par Philippe Vandenkoornhuyse & Yvan Lagadeuc, Université de Rennes
- Loren Méar (2015 – 2016) doctorat en Biologie Moléculaire, encadré par Charle Pineau (INSERM), Université de Paris Dauphine
- Simon Ramondenc (2015 – 2016) doctorat en Océanographie Biologique, encadré par Lionel Guidi (CNRS) et Lars Stemman (Univ. Paris 6), Université Paris 6
- Stanislas Thiriet-Rupert (2014 – 2015) doctorat en Physiologie Végétale, encadré par Jean-Paul Cadoret (IFREMER), Université de Nantes.
- Vincent Picard (2014 – 2015) doctorat en Informatique, encadré par Anne Siegel (CNRS) et Jérémie Bourdon (Univ. Nantes), Université de Rennes & Ecole Normale de Rennes
- Louis Fippo Fitime (2013 – 2014) doctorat en Informatique, encadré par Olivier Roux, Ecole Centrale de Nantes.
- Nicolas Henry (2013 – 2014) doctorat en Océanographie, encadré par Colomban de Vargas, Université Paris 6
- Caroline Baroukh (2012 – 2013) doctorat en Génie des Procédés Biologiques, encadré par Olivier Bernard, Université de Montpellier 1

Responsabilités Administratives

Organisation de conférences et ateliers scientifiques

- Co-organisation de JOBIM 2019 à Nantes (∼450 personnes)
- Co-organisation des "journées réseaux" pour le GDR Génomique Environnemental, 2 juin 2017, Nantes, France (∼ 102 personnes)
- Membre du comité d'organisation de la Conférence Computational Methods in Systems Biology, 16 – 18 septembre 2015, Nantes, France (∼ 85 personnes)
- Co-organisation du Workshop international on Integrative-omics, 9 – 12 décembre 2013, Pucón, Chili (∼ 50 personnes)
- Membre du comité d'organisation de JOBIM 2009 à Nantes (∼450 personnes)
- Co-organisation de l'école thématique bi-régionale en biologie intégrative BIGOU (7-9 novembre 2011, ∼50 personnes).
- Co-organisation de trois éditions des journées satellites "Modélisation dynamique et simulation des réseaux biologiques" (∼60 personnes en moyenne),
  - Lille 2008,
  - Nantes 2009,
  - Montpellier 2010

Membre de Comités de Selection

J'ai participé à des comités de sélection pour le recrutement de chercheurs et enseignant-chercheurs en Biologie et en Informatique.

avril 2021	Paris Sorbonnes Université, section 65 CNU (Biologie Moléculaire)	*Membre extérieur*
avril 2018	Paris Nord Université, section 65 CNU (Biologie Moléculaire)	*Membre extérieur*
mars 2015	INRIA Rhône-Alpes (Informatique)	*Membre extérieur, Chargé de Recherche*
juin 2012	Université de Nantes, section 27 CNU (Informatique)	
juin 2012	Université de Nantes, section 27 CNU (Informatique)	
juin 2012	Ecole Centrale de Nantes, section 27 CNU (Informatique)	*Membre extérieur, Chaire CNRS*
septembre 2011	Université de Nantes, section 27 CNU (Informatique)	
juin 2011	Université de Nice, section 27 CNU (Informatique)	*Membre extérieur*
juin 2011	Université de Nantes, section 27 CNU (Informatique)	
juin 2010	Université de Nantes, section 27 CNU (Informatique)	
juin 2009	Université de Nantes, section 27 CNU (Informatique)	
juin 2009	Université de Rennes 1, section 65 CNU (Biologie Cellulaire)	*Membre extérieur*

Vie du laboratoire

- Responsable de l'équipe ComBi (2019 –)
- Chargé de Mission Développement Durable pour le LS2N (2021 –)
- Membre du bureau de la celllule *"Qualité de Vie au Laboratoire"*
- Membre élu du conseil du laboratoire LS2N (2017 – ....)
- Animateur du thème transversal Bioinformatique au LINA (2013 – 2016)
- Membre nommé du conseil de laboratoire du LINA (2011 – 2016)
- Membre élu du conseil scientifique de l'UFR Sciences et Techniques de l'Université de Nantes (2012 – 2017)
- Membre du comité d'organisation de la fête de la Science (2012 – 2016)
- Responsable des séminaires mensuels "jeunes chercheurs" internes au laboratoire (2009 – 2011)

Vulgarisation Scientifique

- Conférencier invité à la nuit blanche des chercheurs : approches du vivant à l'ère numérique. 27 janvier 2021 lien ici
- Conférencier invité à la Nantes Digital Week : IA et objectifs de développement durables. 25 septembre 2020 lien ici
- Conférencier invité aux Rencontres GreenTech du ministère de la transition écologique et solidaire : *Comment créer de la valeur avec les données*. 8 juin 2017
- Conférencier invité à l'université d'été du Modem: *Tara Océans et les enjeux de l'océan dans la compréhension du climat*

- Presse régionale (Presse Ocean) : lien ici
- Interview radio – mars 2016 – Radio SUN lien ici
- Emission régionale : TV Nantes avril 2016s lien ici
- Interview radio – Podcast National Interstices lien ici
- Weekly Scientific international journals Nature (Research Highlight) Feb 2016
- From Molecular Oceanography to Ocean systems modeling. Principles of Systems Biology - N^{o}5, Cell Systems, 2016
- Parole de Chercheurs, Université de Nantes lien ici

Publications avec comité de lecture

Journaux

[1] Yajuan Lin, Carly Moreno, Adrian Marchetti, Hugh Ducklow, Oscar Schofield, Erwan Delage, Michael Meredith, Zuchuan Li, Damien Eveillard, Samuel Chaffron, and Nicolas Cassar. Decline in plankton diversity and carbon flux with reduced sea ice extent along the Western Antarctic Peninsula. *Nature Communications*, 12(1):4948, dec 2021.

[2] Seaver Wang, Weiyi Tang, Erwan Delage, Scott Gifford, Hannah Whitby, Aridane G González, Damien Eveillard, Hélène Planquette, and Nicolas Cassar. Investigating the microbial ecology of coastal hotspots of marine nitrogen fixation in the western North Atlantic. *Scientific Reports*, 11(1):5508, dec 2021.

[3] Marta and Royo-Llonch, Pablo Sánchez, Clara Ruiz-González, Guillem Salazar, Carlos Pedrós-Alió, Marta Sebastián, Karine Labadie, Lucas Paoli, Federico M. Ibarbalz, Lucie Zinger, Benjamin Churcheward, Marcel Babin, Peer Bork, Emmanuel Boss, Guy Cochrane, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Lionel Guidi, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Nicole Poulton, Jeroen Raes, Christian Sardet, Sabrina Speich, Lars Setmmann, Matthew B. Sullivan, Samuel Chaffron, Damien Eveillard, Eric Karsenti, Shinichi Sunagawa, Patrick Wincker, Lee Karp-Boss, Chris Bowler, and Silvia G. Acinas. Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. *Nature Microbiology*, 2021.

[4] Samuel Chaffron, Erwan Delage, Marko Budinich, Damien Vintache, Nicolas Henry, Charlotte Nef, Mathieu Ardyna, Ahmed A. Zayed, Pedro C. Junger, Pierre E. Galand, Connie Lovejoy, Alison E. Murray, Hugo Sarmento, Silvia G. Acinas, Marcel Babin, Daniele Iudicone, Olivier Jaillon, Eric Karsenti, Patrick Wincker, Lee Karp-Boss, Matthew B. Sullivan, Chris Bowler, Colomban De Vargas, and Damien Eveillard. Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Science Advances*, 7(35):1–16, 2021.

[5] Simon M Dittami, Enrique Arboleda, Jean-Christophe Auguet, Arite Bigalke, Enora Briand, Paco Cárdenas, Ulisse Cardini, Johan Decelle, Aschwin H Engelen, Damien Eveillard, Claire M M Gachon, Sarah M Griffiths, Tilmann Harder, Ehsan Kayal, Elena Kazamia, François H Lallier, Monica Medina, Ezequiel M Marzinelli, Teresa Maria

Morganti, Laura Núñez Pons, Soizic Prado, José Pintado, Mahasweta Saha, Marc-André Selosse, Derek Skillings, Willem Stock, Shinichi Sunagawa, Eve Toulza, Alexey Vorobev, Catherine Leblanc, and Fabrice Not. A community perspective on the concept of marine holobionts: current status, challenges, and future directions. *PeerJ*, 9:e10911, 2021.

[6] Shinichi Sunagawa, Silvia G Acinas, Peer Bork, Chris Bowler, Tara Oceans coordinators, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, Fabien Lombard, Hiroyuki Ogata, Stéphane Pesant, Matthew B Sullivan, Patrick Wincker, and Colomban de Vargas. Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, 18(8):428–445, August 2020.

[7] Clarisse Lemonnier, Morgan Perennou, Damien Eveillard, Antonio Fernàndez-Guerra, Aude Leynaert, Louis Marié, Hilary G Morrison, Laurent Memery, Christine Paillard, and Loïs Maignien. Linking Spatial and Temporal Dynamic of Bacterioplankton Communities With Ecological Strategies Across a Coastal Frontal Area. *Frontiers in Marine Science*, 7:258, June 2020.

[8] Alison E Murray, Nicole E Avalon, Lucas Bishop, Karen W Davenport, Erwan Delage, Armand E K Dichosa, Damien Eveillard, Mary L Higham, Sofia Kokkaliari, Chien-Chi Lo, Christian S Riesenfeld, Ryan M Young, Patrick S G Chain, and Bill J Baker. Uncovering the Core Microbiome and Distribution of Palmerolide in Synoicum adareanum Across the Anvers Island Archipelago, Antarctica. *Marine drugs*, 18(6), June 2020.

[9] Alison E Murray, John Freudenstein, Simonetta Gribaldo, Roland Hatzenpichler, Philip Hugenholtz, Peter Kämpfer, Konstantinos T Konstantinidis, Christopher E Lane, R Thane Papke, Donovan H Parks, Ramon Rossello-Mora, Matthew B Stott, Iain C Sutcliffe, J Cameron Thrash, Stephanus N Venter, William B Whitman, Silvia G Acinas, Rudolf I Amann, Karthik Anantharaman, Jean Armengaud, Brett J Baker, Roman A Barco, Helge B Bode, Eric S Boyd, Carrie L Brady, Paul Carini, Patrick S G Chain, Daniel R Colman, Kristen M DeAngelis, Maria Asuncion de Los Rios, Paulina Estrada-de Los Santos, Christopher A Dunlap, Jonathan A Eisen, David Emerson, Thijs J G Ettema, Damien Eveillard, Peter R Girguis, Ute Hentschel, James T Hollibaugh, Laura A Hug, William P Inskeep, Elena P Ivanova, Hans-Peter Klenk, Wen-Jun Li, Karen G Lloyd, Frank E Löffler, Thulani P Makhalanyane, Duane P Moser, Takuro Nunoura, Marike Palmer, Victor Parro, Carlos Pedrós-Alió, Alexander J Probst, Theo H M Smits, Andrew D Steen, Emma T Steenkamp, Anja Spang, Frank J Stewart, James M Tiedje, Peter Vandamme, Michael Wagner, Feng-Ping Wang, Brian P Hedlund, and Anna-Louise Reysenbach. Roadmap for naming uncultivated Archaea and Bacteria. *Nature Microbiology*, 1:16048–7, June 2020.

[10] Simon Ramondenc, Damien Eveillard, Lionel Guidi, Fabien Lombard, and Benoît Delahaye. Probabilistic modeling to estimate jellyfish ecophysiological properties and size distributions. *Scientific reports*, 10(1):6074, April 2020.

[11] Simon M Dittami, Enrique Arboleda, Jean-Christophe Auguet, Arite Bigalke, Enora Briand, Paco Cárdenas, Ulisse Cardini, Johan Decelle, Aschwin Engelen, Damien

Eveillard, Claire M M Gachon, Sarah Griffiths, Tilmann Harder, Ehsan Kayal, Elena Kazamia, François H Lallier, Mónica Medina, Ezequiel M Marzinelli, Teresa Morganti, Laura Núñez Pons, Soizic Prado, José Pintado Valverde, Mahasweta Saha, Marc-André Selosse, Derek Skillings, Willem Stock, Shinichi Sunagawa, Eve Toulza, Alexey Vorobev, Catherine Leblanc, and Fabrice Not. A community perspective on the concept of marine holobionts: current status, challenges, and future directions. *Peer Community in Ecology*, pages 1–30, March 2020.

[12] Ulysse Guyet, Ngoc A Nguyen, Hugo Doré, Julie Haguait, Justine Pittera, Maël Conan, Morgane Ratin, Erwan Corre, Gildas Le Corguillé, Loraine Brillet-Guéguen, Mark Hoebeke, Christophe Six, Claudia Steglich, Anne Siegel, Damien Eveillard, Frédéric Partensky, and Laurence Garczarek. Synergic Effects of Temperature and Irradiance on the Physiology of the Marine Synechococcus Strain WH7803. *Frontiers in microbiology*, 11:1707, 2020.

[13] Antonietta Capotondi, Michael Jacox, Chris Bowler, Maria Kavanaugh, Patrick Lehodey, Daniel Barrie, Stephanie Brodie, Samuel Chaffron, Wei Cheng, Daniela F Dias, Damien Eveillard, Lionel Guidi, Daniele Iudicone, Nicole S Lovenduski, Janet A Nye, Ivonne Ortiz, Douglas Pirhalla, Mercedes Pozo Buil, Vincent Saba, Scott Sheridan, Samantha Siedlecki, Aneesh Subramanian, Colomban de Vargas, Emanuele Di Lorenzo, Scott C Doney, Albert J Hermann, Terrence Joyce, Mark Merrifield, Arthur J Miller, Fabrice Not, and Stéphane Pesant. Observational Needs Supporting Marine Ecosystems Modeling and Forecasting: From the Global Ocean to Regional and Coastal Systems. *Frontiers in Marine Science*, 6:383, October 2019.

[14] Luigi Caputi, Quentin Carradec, Damien Eveillard, Amos Kirilovsky, Eric Pelletier, Juan J Pierella Karlusich, Fabio Rocha Jimenez Vieira, Emilie Villar, Samuel Chaffron, Shruti Malviya, Eleonora Scalco, Silvia G Acinas, Adriana Alberti, Jean-Marc Aury, Anne Sophie Benoiston, Alexis Bertrand, Tristan Biard, Lucie Bittner, Martine Boccara, Jennifer R Brum, Christophe Brunet, Greta Busseni, Anna Carratalà, Hervé Claustre, Luis Pedro Coelho, Sébastien Colin, Salvatore D'Aniello, Corinne Da Silva, Marianna Del Core, Hugo Doré, Stéphane Gasparini, Florian Kokoszka, Jean-Louis Jamet, Christophe Lejeusne, Cyrille Lepoivre, Magali Lescot, Gipsi Lima-Mendez, Fabien Lombard, Julius Lukeš, Nicolas Maillet, Mohammed-Amin Madoui, Elodie Martinez, Maria Grazia Mazzocchi, Mario B Néou, Javier Paz Yepes, Julie Poulain, Simon Ramondenc, Jean-Baptiste Romagnan, Simon Roux, Daniela Salvagio Manta, Remo Sanges, Mario Sprovieri, Vincent Taillandier, Atsuko Tanaka, Leila Tirichine, Camille Trottier, Julia Uitz, Alaguraj Veluchamy, Jana Veselá, Flora Vincent, Sheree Yau, Stefanie Kandels-Lewis, Sarah Searson, Céline Dimier, Marc Picheral, Peer Bork, Lionel Guidi, Paolo Sordino, Matthew B Sullivan, Alessandro Tagliabue, Adriana Zingone, Laurence Garczarek, Fabrizio d'Ortenzio, Pierre Testor, Maurizio Ribera d'Alcalà, Patrick Wincker, Chris Bowler, Tara Oceans coordinators, Emmanuel Boss, Colomban Vargas, Michael J Follows, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Lee Karp-Boss, Eric Karsenti, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Jeroen Raes, Emmanuel G Reynaud, Christian Sardet, Mike Sieracki, Sabrina Speich, Lars Stemmann, Shinichi Sunagawa, Didier Velayoudon, and Jean Weissenbach. Community-Level Responses to Iron

Availability in Open Ocean Plankton Ecosystems. *Global Biogeochemical Cycles*, 4(30):10,438, March 2019.

[15] Domenico D'Alelio, Damien Eveillard, Victoria J Coles, Luigi Caputi, Maurizio Ribera d'Alcalà, and Daniele Iudicone. Modelling the complexity of plankton communities exploiting omics potential: From present challenges to an integrative pipeline. *Current Opinion in Systems Biology*, 13:68–74, February 2019.

[16] Damien Eveillard, Nicholas J Bouskill, Damien Vintache, Julien Gras, Bess B Ward, and Jérémie Bourdon. Probabilistic Modeling of Microbial Metabolic Networks for Integrating Partial Quantitative Knowledge Within the Nitrogen Cycle. *Frontiers in microbiology*, 9:395, January 2019.

[17] Stanislas Thiriet-Rupert, Gregory Carrier, Camille Trottier, Damien Eveillard, Benoit Schoefs, Gael Bougaran, Jean-Paul Cadoret, Benoit Chénais, and Bruno Saint-Jean. Identification of transcription factors involved in the phenotype of a domesticated oleaginous microalgae strain of Tisochrysis lutea. *Algal Research*, 30:59–72, March 2018.

[18] Dinka Mandakovic, Claudia Rojas, Jonathan Maldonado, Mauricio Latorre, Dante Travisany, Erwan Delage, Audrey Bihouée, Géraldine Jean, Francisca P. Díaz, Beatriz Fernández-Gómez, Pablo Cabrera, Alexis Gaete, Claudio Latorre, Rodrigo Gutierrez, Alejandro Maass, Verónica Cambiazo, Sergio Navarrete, Damien Eveillard, and Mauricio Gonzalez. Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports*, in press, 2018.

[19] Benoît Delahaye, Damien Eveillard, and Nicholas Bouskill. On the Power of Uncertainties in Microbial System Modeling: No Need To Hide Them Anymore. *mSystems*, 2(6), November 2017.

[20] Sylvain Prigent, Clémence Frioux, Simon M Dittami, Sven Thiele, Abdelhalim Larhlimi, Guillaume Collet, Fabien Gutknecht, Jeanne Got, Damien Eveillard, Jérémie Bourdon, Frédéric Plewniak, Thierry Tonon, and Anne Siegel. Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks. *PLoS computational biology*, 13(1):e1005276, January 2017.

[21] Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. *PLoS ONE*, 12(2):e0171744, 2017.

[22] Julie Laniau, Clémence Frioux, Jacques Nicolas, Caroline Baroukh, Maria Paz Cortés, Jeanne Got, Camille Trottier, Damien Eveillard, and Anne Siegel. Combining graph and flux-based structures to decipher phenotypic essential metabolites within metabolic networks. *PeerJ*, 5(3):e3860, 2017.

[23] François Thomas, Philippe Bordron, Damien Eveillard, and Gurvan Michel. Gene Expression Analysis of Zobellia galactanivorans during the Degradation of Algal Polysaccharides Reveals both Substrate-Specific and Shared Transcriptome-Wide Responses. *Frontiers in microbiology*, 8:1808, 2017.

[24] Alix Mas, Shahrad Jamshidi, Yvan Lagadeuc, Damien Eveillard, and Philippe Vandenkoornhuyse. Beyond the Black Queen Hypothesis. *The ISME Journal*, 10(9):2085–2091, September 2016.

[25] Philippe Bordron, Mauricio Latorre, Maria Paz Cortés, Mauricio González, Sven Thiele, Anne Siegel, Alejandro Maass, and Damien Eveillard. Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach. *MicrobiologyOpen*, 5(1):106–117, February 2016.

[26] Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline, Jennifer R Brum, Luis Pedro Coelho, Julio Cesar Ignacio Espinoza, Shruti Malviya, Shinichi Sunagawa, Céline Dimier, Stefanie Kandels-Lewis, Marc Picheral, Julie Poulain, Sarah Searson, Tara Oceans Consortium Coordinators, Lars Stemmann, Fabrice Not, Pascal Hingamp, Sabrina Speich, Mick Follows, Lee Karp-Boss, Emmanuel Boss, Hiroyuki Ogata, Stéphane Pesant, Jean Weissenbach, Patrick Wincker, Silvia G Acinas, Peer Bork, Colomban de Vargas, Daniele Iudicone, Matthew B Sullivan, Jeroen Raes, Eric Karsenti, Chris Bowler, and Gabriel Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465–470, February 2016.

[27] Vicente Acuña, Andrés Aravena, Carito Guziolowski, Damien Eveillard, Anne Siegel, and Alejandro Maass. Deciphering transcriptional regulations coordinating the response to environmental changes. *BMC bioinformatics*, 17(1):35, 2016.

[28] Shahrad Jamshidi, Jocelyn E Behm, Damien Eveillard, E Toby Kiers, and Philippe Vandenkoornhuyse. Using hybrid automata modelling to study phenotypic plasticity and allocation strategies in the plant mycorrhizal mutualism. *Ecological Modelling*, 311:11–19, September 2015.

[29] Nicolas Mouquet, Yvan Lagadeuc, Vincent Devictor, Luc Doyen, Anne Duputié, Damien Eveillard, Denis Faure, Eric Garnier, Olivier Gimenez, Philippe Huneman, Franck Jabot, Philippe Jarne, Dominique Joly, Romain Julliard, Sonia Kefi, Gael J Kergoat, Sandra Lavorel, Line Le Gall, Laurence Meslin, Serge Morand, Xavier Morin, Hélène Morlon, Gilles Pinay, Roger Pradel, Frank M Schurr, Wilfried Thuiller, and Michel Loreau. REVIEW: Predictive ecology in a changing world. *Journal of Applied Ecology*, 52(5):1293–1310, July 2015.

[30] Thierry Tonon and Damien Eveillard. Marine systems biology. *Frontiers in genetics*, 6:181, 2015.

[31] Sylvain Prigent, Guillaume Collet, Simon M Dittami, Ludovic Delage, Floriane Ethis de Corny, Olivier Dameron, Damien Eveillard, Sven Thiele, Jeanne Cambefort, Catherine Boyen, Anne Siegel, and Thierry Tonon. The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond. *The Plant journal : for cell and molecular biology*, 80(2):367–381, October 2014.

[32] Simon M Dittami, Damien Eveillard, and Thierry Tonon. A metabolic approach to study algal-bacterial interactions in changing environments. *Molecular ecology*, 23(7):1656–1660, April 2014.

[33] Silvia E Newell, Damien Eveillard, Mark J McCarthy, Wayne S Gardner, Zhanfei Liu, and Bess B Ward. A shift in the archaeal nitrifier community in response to natural and anthropogenic disturbances in the northern Gulf of Mexico. *Environmental Microbiology Reports*, 6(1):106–112, February 2014.

[34] Richard A Long, Damien Eveillard, Shelli L M Franco, Eric Reeves, and James L Pinckney. Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. *FEMS microbiology ecology*, 83(1):74–81, January 2013.

[35] Nicholas J Bouskill, Damien Eveillard, Diana Chien, Amal Jayakumar, and Bess B Ward. Environmental factors determining ammonia-oxidizing organism distribution and diversity in marine environments. *Environmental Microbiology*, 14(3):714–729, March 2012.

[36] Thierry Tonon, Damien Eveillard, Sylvain Prigent, Jérémie Bourdon, Philippe Potin, Catherine Boyen, and Anne Siegel. Toward systems biology in brown algae to explore acclimation and adaptation to the shore environment. *Omics : a journal of integrative biology*, 15(12):883–892, December 2011.

[37] Jérémie Bourdon, Damien Eveillard, and Anne Siegel. Integrating quantitative knowledge into a qualitative gene regulatory network. *PLoS computational biology*, 7(9):e1002157, September 2011.

[38] Etienne Z Gnimpieba, Damien Eveillard, Jean-Louis Guéant, and Abalo Chango. Using logic programming for modeling the one-carbon metabolism network to study the impact of folate deficiency on methylation processes. *Molecular bioSystems*, 7(8):2508–2521, August 2011.

[39] Philippe Bordron, Damien Eveillard, and Irena Rusu. Integrated analysis of the gene neighbouring impact on bacterial metabolic networks. *IET systems biology*, 5(4):261–268, July 2011.

[40] Nicholas J Bouskill, Damien Eveillard, Gregory O'mullan, George A Jackson, and Bess B Ward. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. *Environmental Microbiology*, 13(4):872–886, April 2011.

[41] Susan F Martinez, Axelle Renodon-Cornière, Julian Nomme, Damien Eveillard, Fabrice Fleury, Masayuki Takahashi, and Pierre Weigel. Targeting human Rad51 by specific DNA aptamers induces inhibition of homologous recombination. *Biochimie*, 92(12):1832–1838, December 2010.

[42] Jonathan Fromentin, Damien Eveillard, and Olivier Roux. Hybrid modeling of biological networks: mixing temporal and qualitative biological properties. *BMC systems biology*, 4(1):79, June 2010.

[43] Jamil Ahmad, Jérémie Bourdon, Damien Eveillard, Jonathan Fromentin, Olivier Roux, and Christine Sinoquet. Temporal constraints of a gene regulatory network: Refining a qualitative simulation. *BioSystems*, 98(3):149–159, December 2009.

[44] Abalo Chango, Afif Abdel Nour, Souad Bousserouel, Damien Eveillard, Pauline M Anton, and Jean-Louis Guéant. Time course gene expression in the one-carbon metabolism network using HepG2 cell line grown in folate-deficient medium. *The Journal of nutritional biochemistry*, 20(4):312–320, April 2009.

[45] Géraldine Del Mondo, Damien Eveillard, and Irena Rusu. Homogeneous decomposition of protein interaction networks: refining the description of intra-modular interactions. *Bioinformatics (Oxford, England)*, 25(7):926–932, April 2009.

[46] Bess B Ward, Damien Eveillard, Julie D Kirshtein, Joshua D Nelson, Mary A Voytek, and George A Jackson. Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental Microbiology*, 9(10):2522–2538, 2007.

[47] Damien Eveillard, Delphine Ropers, Hidde de Jong, Christiane Branlant, and Alexander Bockmayr. A multi-scale constraint programming model of alternative splicing regulation. *Theoretical Computer Sciences*, 325(1):3–24, 2004.

Conference

[48] Marko Budinich, Jérémie Bourdon, Abdelhalim Larhlimi, and Damien Eveillard. OPINION PAPER Evolutionary Constraint-Based Formulation Requires New Bi-level Solving Techniques. In *Computational Methods in Systems Biology*, pages 279–281, Nantes, September 2015. Springer International Publishing.

[49] Guillaume Collet, Damien Eveillard, Martin Gebser, Sylvain Prigent, Torsten Schaub, Anne Siegel, and Sven Thiele. Extending the Metabolic Network of Ectocarpus Siliculosus using Answer Set Programming. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, LNAI8148:245–256, September 2013.

[50] Philippe Bordron, Damien Eveillard, Alejandro Maass, Anne Siegel, and Sven Thiele. An ASP application in integrative biology: identification of functional gene units. *LPNMR - 12th Conference on Logic Programming and Nonmonotonic Reasoning - 2013, Corunna : Spain*, 8148:206–218, 2013.

[51] Philippe Bordron, Damien Eveillard, and Irena Rusu. SIPPER: A flexible method to integrate heterogeneous data into a metabolic network. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pages 40–45, February 2011.

[52] A Goldsztejn, O Mullier, Damien Eveillard, and H Hosobe. Including Ordinary Differential Equations Based Constraints in the Standard CP Framework. *Principles and Practice of Constraint Programming, CP2010*, LNCS 6308:221–235, 2010.

[53] Sébastien Angibaud, Damien Eveillard, Guillaume Fertin, and Irena Rusu. Comparing Bacterial Genomes by Searching Their Common Intervals. *BICoB*, pages 102–113, 2009.

[54] Alexander Bockmayr, Arnaud Courtois, Damien Eveillard, and Myriam Vezain. Building and Analysing an Integrative Model of HIV-1 RNA Alternative Splicing. In

Vincent Danos and Vincent Schaechter, editors, *Computational Methods in Systems Biology, CMSB'04*, pages 43–57, Paris, France, May 2004. Springer LNBI.

[55] Damien Eveillard, Abdelhalim Larhlimi, Delphine Ropers, Stéphanie Billaut, and Sandrine Peyrefitte. KOALAB: A new method for regulatory motif search. In *JOBIM-04'*, Montréal, Canada, 2004.

[56] Damien Eveillard, Delphine Ropers, Hidde de Jong, Christiane Branlant, and Alexander Bockmayr. Multiscale modeling of alternative splicing regulation. In C Priami, editor, *Computational Methods in Systems Biology, CMSB'03*, pages 75–87, Rovereto, Italy,, 2003. Springer LNCS.

[57] Damien Eveillard and Yann Guermeur. Traitement statistique des résultats SELEX. In Jacques Nicolas and Claude Thermes, editors, *JOBIM*, pages 277–283, St Malo, 2002.

Contents

List of Tables

List of Figures

Titre : De la biologie systémique à l'écologie systémique : un voyage informatique à travers les échelles biologiques

Mots clés : biologie systémique, modélisation informatique, contraintes, écologie microbienne

Résumé : Les progrès récents de la métagénomique ont favorisé un changement de paradigme dans l'étude des écosystèmes microbiens. Ces écosystèmes sont aujourd'hui analysés par leur contenu génétique qui permet notamment de mettre en évidence la composition microbienne en terme de taxonomie ou plus récemment leurs fonctions putatives. Cependant, comprendre suffisamment bien les interactions entre les communautés microbiennes et l'environnement pour prédire la diversité à partir de paramètres physico-chimiques est une quête fondamentale de l'écologie microbienne qui nous échappe encore. Cette tâche nécessite de déchiffrer les règles mécanistes qui prévalent au niveau moléculaire. Une telle tâche doit être accomplie par des approches ou des modélisations informatiques dédiées, inspirées de la Biologie Systémique. Néanmoins, l'application directe des approches standard de la biologie des systèmes cellulaires est une tâche complexe. En effet, la description métagénomique des écosystèmes montre un grand nombre de variables à étudier. De plus, les communautés sont (i) complexes, (ii) le plus souvent décrites qualitativement, et (iii) la compréhension quantitative de la façon dont les communautés interagissent avec leur environnement reste incomplète. Dans ce résumé de recherche, nous illustrerons comment les approches de la biologie systémique doivent être adaptées pour surmonter ces points de différentes manières. Dans un premier temps, nous présenterons l'application du protocole bioinformatique aux données de métagénomique, avec un accent particulier sur l'analyse des réseaux. Deuxièmement, nous décrirons comment intégrer les connaissances hétérogènes en omique par programmation logique. Cette intégration mettra l'accent sur les unités fonctionnelles présumées au niveau communautaire. Troisièmement, nous illustrerons la conception et l'utilisation de la modélisation quantitative à partir de ce réseau. En particulier, la modélisation basée sur les contraintes sera utilisée pour prédire la structure de la communauté microbienne et ses comportements à partir des connaissances à l'échelle du génome.

Title: From Systems Biology to Systems Ecology: a computational journey through biological scales

Keywords: systems biology, computational modeling, constraints, microbial ecology

Abstract: Recent progress in metagenomics has promoted a change of paradigm to investigate microbial ecosystems. These ecosystems are today analyzed by their gene content that, in particular, allows to emphasize the microbial composition in term of taxonomy (i.e., «who is there and who is not») or more recently their putative functions. However, understanding the interactions between microbial communities and their environment well enough to be able to predict diversity based on physicochemical parameters is a fundamental pursuit of microbial ecology that still eludes us. This task requires to decipher the mechanistic rules that prevail at the molecular level. Such a task must be achieved by dedicated computational approaches or modeling, as inspired by Systems Biology. Nevertheless, the direct application of standard cellular systems biology approaches is a complicated task. Indeed, the metagenomic description of ecosystems shows a large number of variables to investigate. Furthermore, communities are complex, mostly described qualitatively, and the quantitative understanding of the way communities interacts with their surroundings remains incomplete. Within this research summary, we will illustrate how systems biology approaches must be adapted to overcome these points in different manners. First, we will present the application of bioinformatics protocol on metagenomics data, with a particular emphasis on network analysis. Second, we will describe how to integrate heterogeneous omics knowledge via logic programming. Such integration will emphasize putative functional units at the community level. Third, we will illustrate the design and the use of quantitative modeling from this network. In particular, constraint-Based modeling will be used to predict microbial community structure and its behaviors based on genome-scale knowledge.