2022
ACTIVITY REPORT

Project-Team
PLEIADE

Patterns of diversity and networks of function

IN COLLABORATION WITH: Laboratoire Bordelais de Recherche en Informatique (LaBRI), Biodiversité, Gènes & Communautés (BioGeCo)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Inria

Contents

Project-Team PLEIADE

Creation of the Project-Team: 2019 March 01

Keywords

Computer sciences and digital sciences

A3.1. – Data

A3.2. – Knowledge

A3.3.2. – Data mining

A3.3.3. – Big data analysis

A3.4. – Machine learning and statistics

A6.1. – Methods in mathematical modeling

A6.2. – Scientific computing, Numerical Analysis & Optimization

A8.2. – Optimization

A9.8. – Reasoning

Other research topics and application domains

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B3. – Environment and planet

1 Team members, visitors, external collaborators

Research Scientists

- David Sherman [Team leader, INRIA, Senior Researcher, HDR]

- Alain Franc [INRAE, Emeritus, HDR]

- Clémence Frioux [INRIA, Researcher]

- Simon Labarthe [INRAE, Researcher]

Post-Doctoral Fellow

- Pablo Andres Ugalde Salas [INRIA]

PhD Students

- Mohamed Anwar Abouabdallah [INRIA]

- Chabname Ghassemi Nedjad [UNIV BORDEAUX, from Oct 2022]

- Maxime Lecomte [INRAE]

Technical Staff

- Ariane Badoual [INRIA, until May 2022]

- Jean-Marc Frigerio [INRAE, Engineer]

- Coralie Muller [INRIA, Engineer, from Feb 2022]

- Franck Salin [INRAE, Engineer]

Interns and Apprentices

- Leonard Brindel [INRIA, Apprentice, from Sep 2022]

- Quentin Chapuzet [INRIA, from Feb 2022 until Jul 2022]

- Chabname Ghassemi Nedjad [INRIA, from Feb 2022 until Jul 2022]

- Arie Wortsman Zurich [INRIA, from Apr 2022 until Jul 2022]

Administrative Assistant

- Catherine Cattaert Megrat [INRIA]

Visiting Scientist

- Sebastien Raguideau [EARLHAM INSTITUTE, from Nov 2022 until Nov 2022]

2   Overall objectives

The simplest ecosystem is a complex network of interactions between a diversity of organisms providing a diversity of functions. Each organism is the result of evolutionary processes driven by the creation then the selection of molecular diversity. Every function is a coordinated cascade of biochemical reactions, sensitive to substrate and environment, arising to adapt to the needs of the organism. Pleiade explores the diversity of organisms and the diversity of their functions and, as a fundamental challenge, seeks to formalize the links between them.

Pleiade measures the **diversity of organism**s by comparing DNA sequences and describes it using geometric methods. Amplicon sequences from metabarcoding are compared systematically to produce matrices of taxonomic distances. Mathematical analysis of these distance matrices, made possible by advances in dimension reduction, pattern recognition, and high-performance computing, reveals complex descriptions of the molecular diversity of the organisms in the sampled environment.

Pleiade examines the **diversity of functions** performed by these organisms by identifying the genes responsible for biochemical processes and grouping them into metabolic networks. Annotation of whole genome and metagenomic sequences allows us to delineate the functions provided by individual organisms and describe the interactions between them. Metabolic and process-based models are developed as compact descriptions of functional diversity. In addition to providing a means to simulate a system, a model is a syntactic object that symbolically represents a range of functional behaviors; patterns in the diversity of the encoded functions can be explored inferentially without exhaustively simulating the model to enumerate the set of behaviors it represents.

Comparison of annotated genomes and their associated metabolic networks reveals how functions arose over time: what functions, when they arose, and by which evolutionary mechanisms. Annotation is not performed on genomes individually, but comparatively, taking into account the similarities and differences between related species and strains.

A further challenge, developed recently, is considering the challenge of linking diversity and function in the particular context of *microbial communities*. We are developing a synergistic, iterative combination of a **community-based** strategy for deciphering the diversity in cultures and environmental samples, through metagenomic and metabolomic analysis of functional diversity and metabarcoding analysis of taxonomic diversity; and a **function-based** strategy for constructing *digital twins* of natural or designed communities through numerical models. The goal is a hybrid framework for studying systems dynamics using spatio-temporal models.

Shared methodologies needed to scale up to the complexity of biological systems, include *high-performance computing* (HPC); *machine learning*, including clustering, meta-modeling and classification for knowledge engineering; *machine reasoning*, specifically logical and rule-based methods used for model inference and network analysis. Logicial methods in particular promote *explainable* inference, since the rules are expressed in biological terms and are auditable by biologists, independently from the combinatorial and heuristic optimization techniques used to apply the rules.

Pleiade maintains strong collaborative relations with experimental biologists, and is committed to developing applications in ecology, evolution, biotechnology, and health. Team resources are dedicated to facilitating the adoption of our research by non specialist users, through development of reusable software, integration in HPC frameworks, improvement of web-based environments, and deployment of Jupyter, Galaxy, and Kubernetes interfaces.

3   Research program

Pleiade's mutually reinforcing strengths are a stable foundation of ecology and comparative genomics, and a novel synthesis of new methods extending our reach into microbial and planktonic communities and their dynamics. Our shared aim is to develop both: new challenges in microbial or planktonic communities leverage our solid expertise in foundational methods, and at the same time define new challenges for improving those foundations. We focus on reinforcing of a first set of *disciplinary* activities related to innovations in methods, be they in applied mathematics or in computer sciences, and a second set of *interdisciplinary* activities, to build advantageous assemblies of methods for understanding and managing biological systems of interest.

3.1 Research: A Geometric View of Diversity

Data analysis algorithms and tools must be revisited and scaled up. We mobilize both distributed algorithms and new algorithms, like random projection or column selection methods, to build point clouds in Euclidean spaces from massive data sets, and thus overcome the cubic complexity of computation of eigenvectors and eigenvalues of very large dense matrices. We also link distance geometry [56] with convex optimization procedures through matrix completion [39, 42].

Intercalibration: There is a considerable difference between supervised and unsupervised clustering: in supervised clustering, the result for an item i is independent from the result for an item $j \neq i$, whereas in unsupervised clustering, the result for an item i (e.g. the cluster it belongs to, and its composition) depends on nearby items $j \neq i$. Which means that the result may change if some items are added to or subtracted from the sample. This raises the more global problem of how to merge two studies to yield a more comprehensive view of biodiversity?

See [48] for some of our recent work linking the distance geometry problem, nonlinear mapping, and weighted least-squares scaling.

Project-team positioning

This research topic is about metabarcoding, i.e. producing inventories through classification, or producing OTUs with clustering, as elementary bricks for diversity, in a way analgous to the role that species play in morphological or molecular based taxonomy. The number of reads produced by NGS facilities is a challenge for bioinformatics and data analysis downstream into these directions: most algorithms scale with the square or the cube of the number of involved reads, which are counted in the millions. Most approaches look for heuristics permitting to produce result within a reasonable time (see e.g. UCLUST or USEARCH as wrappers for many solutions). The Pleiade team has explored another path: relying on HPC (the possibilities of which are underestimated in the community of diversity studies) to derive exact algorithms without heuristics, for computing distances between reads and to analyse distance matrices with dimension reduction.

This has been made possible due to an involvement in scientific computing, associating algorithms, their implemntation and optimisation. Such a bridge between scientific computing and characterisation of the biodiversity is original and still an exploratory field. It is develped in collaboration with project teams HiePACS, Storm and Tadaam in Inria Bordeaux, through two technology development actions: one, ADT Gordon, which is currently being valorized by the redaction of an article, and the second, Diodon, under development.

Collaborations

- In collaboration with the Pasteur Institute in Cayenne and the INRA MIA Research Team in Toulouse, Pleiade is developing a stochastic model for simulation of metacommunities, in the framework of *patch occupancy* models. The objective is a better understanding of zoonose propagation, namely rabies through bat hosts in connection with disturbances of pristine forests in French Guiana, which have an impact on the exposure of human populations to wildlife that act as reservoirs of zoonoses.

- We have co-supervised with Anne Lavergne (Institut Pasteur de Guyane) a PhD student at IPG and University of Cayenne (Sourakhata Tirera, defense on December 17, 2021) on the drivers of the diversity of viromes of rodents and birds. One part is about bioinformatics in order to select viral fragment, which are a minority in shotgun sequencing of viromes, and a second part is about disentangling the role of the habitat and of the phylogeny of the host in viral diversity. This has led to manuscript [64]. A second one is in preparation to make available a new pipeline for dechimerisation of contings after shotgun sequencing and de nova assembly of reads into contigs (a process which is known to create many chimera).

- The Laboratory of excellence (Labex) CEBA promotes innovation in research on tropical biodiversity. It brings together a network of internationally-recognized French research teams, contributes to university education, and encourages scientific collaboration with South American countries. Pleiade has participated in three current international projects funded by CEBA:

> – *MicroBIOMES: Microbial Biodiversities*
>
> – *Neutrophyl: Inferring the drivers of Neotropical diversification*
>
> – *Phyloguianas: Biogeography and pace of diversification in the Guiana Shield*

- We collaborate with Institut Pasteur de Guyane at Cayenne for developing the domain of so-called Ecoviromics for some zoonoses in French Guiana. On top of co-supervising a PhD student at IPG in Cayenne as mentioned above (deciphering the respective roles of host phylogeny and environmental variables in the virome of different hosts (bats, rodents, birds), this collaboration has led to a participation in a synthesis paper on Ecology, evolution, and epidemiology of zoonotic and vector-borne infectious diseases in French Guiana [63].

3.2 Research: Community-scale Metabolic and Omics-based Modeling

Metabolism can be abstracted into sets of metabolic reactions, associated to the genome through gene-protein relationships, and connecting substrates to products thereby forming metabolic networks. **Genome-scale metabolic networks** (GSMNs) contain all the reactions predicted to occur in a organism according to its genomic contents. Combined with additional knowledge on the system, possibly other –omics data and mathematical models, GSMNs are used to **predict the behaviour of an organism or a community** of organisms.

A widely-used mathematical formalism for modeling GSMNs is **constraint-based modeling**, among which flux balance analysis (FBA) is the main representative [58]. Such methods permit a quantitative prediction of activity fluxes in metabolic networks while optimising an objective function and assuming steady-state of the system.

The emergent metabolism of microbial communities can also be **qualitatively modeled using a boolean approximation of metabolic dynamics**[7]. In this approach the behavior of the system is described by logical rules that activate a given reaction as soon as its substrates become available; numerical parameters such as stochiometry or enzyme kinetics are ignored in favor of graph topology and paths. The advantage is that such qualitative models, unlike quantitative methods such as FBA, do not require the assumption that the system is at steady-state and can model systems where cells are constantly growing or constantly reproducing.

Network expansion, introduced in [46] as a recursive traversal of the structure of a metabolic graph, lends itself to concise definition using *answer set programming* (ASP) [53] and thus to efficient implementation using SAT solvers [52]. In practice, using ASP for metabolic modeling makes it possible to define both the activation of metabolic reactions in different conditions, and the constraints and optimizations needed to find solutions in a combinatorically large state space.

We focus in particular on the key question of determining *minimal communities*, subsets of the organisms present in an environment that are sufficient to reproduce a chosen behavior [49]. The methodological goal here is to **identify key species in a community through use of ASP** to rapidly explore the search space and thus, through heuristic resolution of combinatorial problems, provide the guarantees an exhaustive search with a greatly reduced computational cost [4].

Functional and taxonomic diversities, beyond intrinsic specificities encoded in the genetic material, are also strongly shaped by their environment. Spatial nutritional niches, microbial interactions and abiotic constraints lead to **complex spatial structures in the microbial community** that impact its overall dynamics. PDE-based models of the microbiota in its environment allow including in the model these multiple mechanisms in order to decipher their influence on the community faith.

The main methodological developments in this area are related to **mathematical modeling** (in particular the correct level of simplification in the multi-physics description of the microbial environment), **model simplification** (asymptotic approximation), **inference from multi-omics data** (including dimension reduction, statistical learning) and **numerical developments** (in particular fast approximation of metabolic models with machine learning methods). Strong interactions with community-scale metabolic models are sought, specially for multi-omics inference and knowledge-based machine learning constraints.

The goal is to achieve **accurate models of microbial communities** that could be used as *digital twins* of controlled experiments in microbial ecology. Culturomic facilities allow for the acquisition of multi-omics time-series data in controlled conditions, which can be used to build and fit population

dynamics models, and can be used in turn to explore numerically biological assumptions and to help in experimental planning and data analysis.

Project-team positioning

The team has expertise in the state of the art methods for metabolic modeling and in metabolic network reconstruction through earlier works [49, 50, 37, 59]. The team also masters ODE or PDE microbial population dynamics models including their complex environment [41, 55] and parameter inference with experimental data [17]. We work with international or national teams (see below) for combinatorial problem solving (University of Potsdam, Germany), computational biology for health (Quadram Institute Bioscience, UK) and biological applications (Roscoff Biological Station, INRAE teams).

Other international teams working on the subject of community metabolic modeling include but is not restricted to the research groups of: Ines Thiele (U. Galway, Ireland), Kiran Patil (U. Cambridge, UK), Karoline Faust (U. Leuven, Belgium), Daniel Machado (Norwegian University of Science and Technology).

Among Inria project-teams, Dyliss (Rennes) is the one with the closest research themes. Clémence Frioux did her PhD in this team (2015-2018) and stayed for an additional 6 months after the defense. As a result, the majority of her past contributions were done in collaboration with Dyliss members, and current collaborations persist through co-development and maintenance of software, applications etc.

Collaborations

- Quadram Institute Bioscience and Earlham Institute: meta-analyses of metagenomic cohorts for the human gut microbiota.

- University of Potsdam: answer set programming and combinatorial problem solving.

- Station Biologique de Roscoff: algae applications.

- INRAE BFP, STLO, MaIAGE, SAVE, Micalis, IEES: applications and methodological development in computational biology and mathematics.

- CEA. Bio-inspired digital sensors in the framework of the Pherosensor project.

- U. Besançon and U. Orléans. Modeling of the fluidic environment in the gut microbiota.

- U. Paris-Saclay/U.Evry. Machine learning with ANOVA-RKHS.

- Inria Bretagne Atlantique: systems biology and systems ecology.

- Ysopia Bioscience by means of an Inria Tech contract focused the analysis of metagenomic data and the connection to the metabolic screening provided by Metage2Metabo.

- Biomathematica through a CIFRE PhD hosted at MaIAGE (INRAe) co-supervised by Pleiade.

3.3   Research: Bioinformatics, Genomes, and Knowledge Management

The heterogenous data generated in computational molecular biology and ecology are distinguished not only by their volume, but by the richness of the many levels of interpretation that biologists create. The same nucleic acid sequence can be seen as a molecule with a structure, a sequence of base pairs, a collection of genes, an allele, or a molecular fingerprint. To extract the maximum benefit from this treasure trove we must organize the knowledge in ways that facilitate extraction, analysis, and inference. Our focus has been on the efficient representation of relations between biological objects and operations on those representations, in particular heuristic analyses and logical inference.

Pleiade develops applications in comparative genomics of related organisms, using novel mathematical tools for representing compactly, at different scales of difference, comparisons between related genomes. New methods based on distance geometry will refine these comparisons. Compact representations can be stored, exchanged, and combined. They form the basis of novel simultaneous genome annotation methods, that can be linked directly to abductive inference methods for building functional models of the organisms and their communities.

Since a goal of Pleiade is to integrate diversity throughout the analysis process, it is necessary to incorporate **diversity as a form of knowledge** that can be stored in a knowledge base. Diversity can be represented using various compact representations, such as trees and quotient graphs storing nested sets of relations. Extracting structured representations and logical relations from integrated knowledge bases requires domain-specific query methods that can express forms of diversity.

Project-team positioning

Historically, Pleiade members have been pioneers in the development of large-scale eukaryote comparative genomics. We were involved since the late 1990's in the first genome sequencing of eukaryote microorganisms, co-authored 8 articles in the [54] special issue presenting the first large-scale comparative genomics study, and were in the first authors of the landmark Nature article [45] comparing five complete annotated genomes. Our articles in comparative genomics, particularly of the hemiascomycetous yeasts of biotechnological interest, have achieved thousands of citations and continue to do so decades later. A principle that we fought for [44, 61], that comparative genomics must be based on a systematic and mathematical comparison of the genomes, rather than on an opportunistic one-against-all comparison to the model organism *du jour*, is now considered standard practice. We also originally set the standard for web-based tools for comparative genomics, organized around the principle of an interaction design based on the questions asked by by biological user, rather than based on the organization of the underlying database as seen by a computer scientist[62, 61]. Pleiade capitalized on this experience in support of the research efforts describe above, and also through a wide network of collaborators in the biological sciences. Our current work applies the principles of comparative genomics to a series of smaller projects focused on collections of genomes of biotechnological or of health-related interest.

Collaborations

- Institut des Sciences du Vigne et du Vin (ISVV), U. Bordeaux

- Vitapalm (LEAP-Agri)

- Laboratoire de Microbiologie Fondamentale et Pathogénicité (LMFP), UMR 5234 CNRS U. Bordeaux

- Laboratory of Membrane Biology (LBM) UMR 5200 CNRS U. Bordeaux

4   Application domains

4.1   Molecular based systematics and taxonomy

Defining and recognizing the myriads of species occuring in the biosphere has been the focus of phenomenal energy over the past centuries and remains a major goal of Natural History. It is an iconic paradigm in pattern recognition (clustering has coevolved with numerical taxonomy many decades ago). Developments in evolution and molecular biology, as well as in data analysis, have over the past decades enabled a profound revolution, where species can be delimited and recognized by data analysis of sequences. We aim at proposing new tools, in the framework of E-science, which make possible (*i*) better exploration of the diversity in a given clade, and (*ii*) assignment of a place in these patterns for new, unknown organisms, using information provided by sets of sequences. This will require investment in data analysis, machine learning, and pattern recognition to deal with the volumes of data and their complexity.

One example of this project is about the diversity of trees in Amazonian forest, in collaboration with botanists in French Guiana. Protists (unicellular Eukaryots) are by far more diverse than plants, and far less known. Molecular exploration of Eukaryotes diversity is nowadays a standard in biodiversity studies. Data are available, through metagenomics, as an avalanche and make molecular diversity enter the domain of Big Data. Hence, an effort will be invested, in collaboration with other INRIA teams (GenScale, HiePACS) for porting to HPC algorithms of pattern recognition and machine learning, or distance geometry, for these tools to be available as well in metagenomics. This will be developed first

on diatoms (unicellular algae) in collaboration with INRAE team at Thonon and University of Uppsala, on pathogens of tomato and grapewine, within an existing network, and on bacterial communities, in collaboration with University of Pau. For the latter, the studies will extend to correlations between molecular diversity and sets of traits and functions in the ecosystem.

4.2   Genome and transcriptome annotation

Sequencing genomes and transcriptomes provides a picture of how a biological system can function, or does function under a given physiological condition. Simultaneous sequencing of a group of related organisms is now a routine procedure in biological laboratories for studying a behavior of interest, and provides a marvelous opportunity for building a comprehensive knowledge base of the relations between genomes[1, 11]. Key elements in mining these relations are: classifying the genes in related organisms and the reactions in their metabolic networks, recognizing the patterns that describe shared features, and highlighting specific differences.

Pleiade develops applications in comparative genomics of related organisms, using new mathematical tools for representing compactly, at different scales of difference, comparisons between related genomes. New methods based on computational geometry refine these comparisons. Compact representations can be stored, exchanged, and combined. They will form the basis of new simultaneous genome annotation methods, linked directly to abductive inference methods for building functional models of the organisms and their communities.

Our ambition in biotechnology is to permit the design of synthetic or genetically selected organisms at an abstract level, and guide the modification or assembly of a new genome. Our effort is focused on two main applications: genetic engineering and synthetic biology of oil-producing organisms (biofuels in CAER, palm oils), and improving and selecting starter microorganisms used in winemaking (collaboration with the ISVV and the BioLaffort company).

4.3   Community ecology

Community assembly models how species can assemble or diassemble to build stable or metastable communities. It has grown out of inventories of countable organisms. Using *metagenomics* one can produce molecular based inventories at rates never reached before. Most communities can be understood as pathways of carbon exchange, mostly in the form of sugar, between species. Even a plant cannot exist without carbon exchange with its rhizosphere. Two main routes for carbon exchange have been recognized: predation and parasitism. In predation, interactions–even if sometimes dramatic–may be loose and infrequent, whereas parasitism requires what Claude Combes has called intimate and sustainable interactions [43]. About one decade ago, some works [60] have proposed a comprehensive framework to link the studies of biodiversity with community assembly. This is still incipient research, connecting community ecology and biogeography.

We aim at developing graph-based models of co-occurence between species from NGS inventories in metagenomics, i.e. recognition of patterns in community assembly, and as a further layer to study links, if any, between diversity at different scales and community assemblies, starting from current, but oversimplified theories, where species assemble from a regional pool either randomly, as in neutral models, or by environmental filtering, as in niche modeling. We propose to study community assembly as a multiscale process between nested pools, both in tree communities in Amazonia, and diatom communities in freshwaters. This will be a step towards community genomics, which adds an ecological flavour to metagenomics.

Next-generation sequencing technologies are now an essential tool in population and community genomics, either for making evolutionary inferences or for developing SNPs for population genotyping analyses. Two problems are highlighted in the literature related to the use of those technologies for population genomics: variable sequence coverage and higher sequencing error in comparison to the Sanger sequencing technology. Methods are developed to develop unbiased estimates of key parameters, especially integrating sequencing errors [57]. An additional problem can be created when sequences are mapped on a reference sequence, either the sequenced species or an heterologous one, since paralogous genes are then considered to be the same physical position, creating a false signal of diversity [51]. Several approaches were proposed to correct for paralogy, either by working directly on the sequences issued from

mapped reads [51] or by filtering detected SNPs. Finally, an increasingly popular method (RADseq) is used to develop SNP markers, but it was shown that using RADseq data to estimate diversity directly biases estimates [38]. Workflows to implement statistical methods that correct for diversity biases estimates now need an implementation for biologists.

5 Social and environmental responsibility

5.1 Footprint of research activities

Pleiade's policy is to rely on shared computing platforms for computations that consume significant energy, for two reasons. First, those platforms have greater leverage over total energy consumption, and have the technical means to implement green computing on a useful scale. Second, those platforms have staff with the requisite skills to develop and implement policies as part of their service offer. Our partner platforms are mesocenters in Bordeaux and Grenoble, and national centers including the Idris and the CEA. Some of them charge back the cost of CO_2 generation. The scientific committee of the Bordeaux mesocenter MCIA, which has a member from Pleiade, actively debates green computing implementation. Pursuant to our team policy, Pleiade does not use high-powered workstations on the desktop.

6 Highlights of the year

- First publications from the SLIMMEST exploratory action (§8.1) combining discrete reasoning models of metabolism with numerical metamodels and PDEs for the simulation of microbial communities in time and space [29, 35, 36].

- A new method for analyzing time-lapse confocal laser scanning images of *bacterial swimmers* in pathogenic exogeneous host biofilms, based on the inference of a kinetic model of swimmer populations including mechanistic interactions with the host biofilm [17].

- Definition and formal start of the flagship project **MISTIC** (§10.3.2), coordinated by Pleiade, in the *Programme et équipements prioritaires de recherche* (PEPR) Agroecology and ICT. MISTIC will elucidate the role of microbial community dynamics in the adaptation of crop plants to environmental stresses including climate change.

- Dissemination of two new monographs by Alain Franc: *Linear Dimensionality Reduction*, reviewing the broad domain of linear methods in multivariate data analysis, and showing how the diversity of methods and tools boil down to a single core method, PCA with SVD, on which efforts to optimize codes for analyzing massive data sets can focus [28]; and *Tensor Ranks for the Pedestrian*, reviewing tensors and presenting in detail an integrated and progressive approach to approximate a given tensor by a tensor of lower rank [47].

7 New software and platforms

7.1 New software

7.1.1 Metage2Metabo

Keywords: Metabolic networks, Microbiota, Metagenomics, Workflow

Scientific Description: Flexible pipeline for the metabolic screening of large scale microbial communities described by reference genomes or metagenome-assembled genomes. The pipeline comprises several main steps. (1) Automatic and parallel reconstruction of metabolic networks. (2) Computation of individual metabolic potentials (3) Computation of collective metabolic potential (4) Calculation of the cooperation potential described as the set of metabolites producible by species only in a cooperative context (5) Computation of minimal-sized communities sastifying a metabolic objective (6) Extraction of key species (essential and alternative symbionts) associated to a metabolic function

Functional Description: Metabolic networks are graphs which nodes are compounds and edges are biochemical reactions. To study the metabolic capabilities of microbiota, Metage2Metabo uses multiprocessing to reconstruct metabolic networks at large-scale. The individual and collective metabolic capabilities (number of compounds producible) are computed and compared. From these comparisons, a set of compounds only producible by the community is created. These newly producible compounds are used to find minimal communities that can produce them. From these communities, the keystone species in the production of these compounds are identified.

News of the Year: 2022: (1) Fix the number of colors used to color the taxon in the powergraph.(2) Change the shapes of the nodes in the html output of m2m_analysis: circle for essential symbionts and square for alternative symbionts. (3) Release of version 1.5.3.

URL: https://github.com/AuReMe/metage2metabo

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.2 MiSCoTo

Name: Microbiota Screening and COmmunity Selection with TOpology

Keywords: Metabolic networks, ASP - Answer Set Programming, Logic programming

Scientific Description: MiSCoTo solves combinatorial problems using Answer Set Programming. It aims at minimizing either the number of selected species or both the number of selected species and the cost of the interaction between them, characterized by the number of metabolic exchanges. In the first case, the level of modeling is called lumped or mixed-bag, in the latter, it is compartmentalized.

Functional Description: Metabolic networks are composed of biochemical reactions and gather the expected metabolic capabilities of species. For organisms that live in interaction altogether (microbiotas), complementarity between these networks can be exploited to predict cooperation events. This software takes as inputs metabolic networks for various species (host, symbionts of the microbiota), components of the growth medium and a metabolic objective (metabolites to be produced), and aims at selecting a minimal set of symbionts to ensure the metabolic objective can be achieved. The software can use two types of modelings: a simplified one and another that takes into account the cost of metabolic exchanges and aims at minimizing it.

Release Contributions: Memory usage optimization. Fix issues with input file formats.

News of the Year: (1) Release of version 3.1.1 (2) New functionality: miscoto-focus determines the metabolic potential of symbionts of interest in the community

URL: https://github.com/cfrioux/miscoto

Publication: hal-01871600

Contact: Clemence Frioux

Participants: Clemence Frioux, Anne Siegel, Enora Fremy, Camille Trottier, Arnaud Belcour

7.1.3 MeneTools

Name: Metabolic networks Topological tools

Keywords: Metabolic networks, Graph, Topology, Bioinformatics, Systems Biology, ASP - Answer Set Programming

Scientific Description: MeneTools are a set of tools for the exploration of the producibility potential in a metabolic network using the network expansion algorithm. The MeneTools can: - assess whether targets are producible starting from nutrients (Menecheck) - get all compounds that are producible starting from nutrients (Menescope) - get all reactions that are activable from nutrients (Meneacti) - get production paths of specific compounds (Menepath) - obtain compounds that if added to the nutrients, would ensure the producibility of targets (Menecof) - identify metabolic deadends, i.e. metabolites that act as reactants of reactions but never as products, or metabolites that act as products of reactions but never as reactants. This is a purely structural analysis. All MeneTools using modelling follow the producibility in metabolic networks as defined by the network expansion algorithm.

Functional Description: MeneTools consists in four topological tool to analyze metabolic models in a graph-based perspective. Menecheck verifies the producibility of target compounds from available substrates (growth medium) of the metabolic network. Menescope gives the whole range of accessible compounds in the metabolic network starting from substrates. Menepath give the production paths of given compounds in the model. Menecof proposes compounds that need to be produced or added as substrate for ensuring the producibility of targets.

News of the Year: (1) Release of version 3.2.0 (2) New functionality: mene-seed identifies external compounds from the topology of the network

URL: https://github.com/cfrioux/MeneTools

Publications: hal-01819150, hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Anne Siegel, Arnaud Belcour

7.1.4 Fluto

Keywords: ASP - Answer Set Programming, Answer Set Programming, Metabolic networks, Flux Balance Analysis, Linear programming

Scientific Description: Fluto performs metabolic network completion with respect to topological and linear reaction rate constraints based on the stoichiometry of metabolic reactions.

Functional Description: Fluto relies on Answer Set Programming (ASP) and a hybrid modelling that associates to ASP a Linear Programming (LP) constraint propagator. Models satisfying the qualitative constraints of network expansion are tested for satisfiability of flux constraints with the LP propagator. Resulting answer sets permit the completion of a metabolic network that ensures the metabolic reaction of interest is activated according to both formalisms.

News of the Year: Reorganisation of the code. Implementation of continuous integration. Addition of the Sagot & Acuna formalism in the software.

URL: https://github.com/cfrioux/fluto/

Publications: hal-01936778, hal-01557347

Contact: Clemence Frioux

Participant: Sven Thiele

Partners: Max Planck Institute Magdeburg, University of Potsdam

7.1.5   Emapper2GBK

Keywords: Bioinformatics, Metabolic networks, Functional annotation

Functional Description: Starting from FASTA and Eggnog-mapper annotation files, Emapper2GBK builds a GBK file that is suitable for metabolic network reconstruction with Pathway Tools, and adds the GO terms and EC numbers annotations in the GenBank file.

News of the Year: 2022: Using gffutils region to speed up emapper2gbk. Supporting for gmove and eggnog GFF format. Adding gff-type option: add mRNA and gene parameters. Allowing the tool to use the IDs in "mRNA" or "gene" field in the gff to match the faa file IDs. Improving error messages and updating the readme.

URL: https://github.com/AuReMe/emapper2gbk

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.6   Biodiversiton

Name: Biodiversiton

Keywords: Biodiversity, Comparative metagenomics, Clustering, Dimensionality reduction, Masses of data

Functional Description: Biodiversiton is a suite of tools for biodiversity composed by Rsyst, pairwise_dis, diagno_syst, and yapotu. The global project provides tutorials, datasets, and a readme for the whole suite.

URL: https://gitlab.inria.fr/biodiversiton

Authors: Alain Franc, Jean-Marc Frigerio, Franck Salin

Contact: Alain Franc

7.1.7   Yapotu

Name: Yet Another Pipeline for OTU building

Keywords: Metagenomics, Biodiversity, Dimensionality reduction, Masses of data

Functional Description: The main functionalities are as follows: 1) building OTUs from a fasta file (swarm, vsearch, ..) or a distannce file (yapotu) for an environmenal sample 2) building a fasta file and a distance file per OTU 3) checking the consistency of the OTUs by displaying them as a graph (see OTU as a graph below) 4) displaying the shape of an OTU or of a set of OTUs by Multidimensional Scaling 5) implementing Hierachical Aggregative Clustering of an OTU or a set of OTUs with various aggregation methods

News of the Year: Ugraded from an older version, fusion with declic now deprecated, new functionalities for working with massive data sets

URL: https://gitlab.inria.fr/biodiversiton/yap

Authors: Alain Franc, Jean-Marc Frigerio, Franck Salin

Contact: Alain Franc

Partner: INRAE

7.1.8 pydiodon

Name: pydiodon

Keywords: Dimensionality reduction, Data analysis

Functional Description: Most of dimension reduction methods inherited from Multivariate Data Analysis, and currently implemented as element in statistical learning for handling very large datasets (the dimension of spaces is the number of features) rely on a chain of pretreatments, a core with a SVD for low rank approximation of a given matrix, and a post-treatment for interpreting results. The costly part in computations is the SVD, which is in cubic complexity. Diodon is a list of functions and drivers which implement (i) pre-treatments, SVD and post-treatments on a large diversity of methods, (ii) random projection methods for running the SVD which permits to bypass the time limit in computing the SVD, and (iii) an implementation in C++ of the SVD with random projection at prescribed rank or precision, connected to MDS.

Release Contributions: - completed documentation with sphynx - library now public through Inria git - availability of a readme - making a few "toy" datasets available - delivering a few jupyter notebooks as tutorials

URL: https://gitlab.inria.fr/diodon/pydiodon

Authors: Alain Franc, Florent Pruvost, Romain Peressoni, Romain Peressoni

Contact: Alain Franc

7.1.9 swim-infer

Keywords: Bayesian estimation, Trajectory Modeling, Trajectory Generation

Scientific Description: Bayesian inference of swimming trajectories described by population-wide movement conservation equations taking into account interaction terms with the host biofilm (speed and direction selection, stochastic term).

Functional Description: Bayesian inference solver of swimming trajectories of microbial micro-swimmers in biofilms from confocal microscopy images. The specificity of the solver is to take into account the interactions of the micro-swimmers with the host biofilm.

News of the Year: First release

URL: https://forgemia.inra.fr/bioswimmers/swim-infer

Publication: hal-03695580

Contact: Simon Labarthe

Participants: Simon Labarthe, Guillaume Ravel

7.1.10 mpwt

Keywords: Metabolic networks, Multi-processor

Functional Description: mpwt is a Python package for running Pathway Tools on multiple genomes using multiprocessing. More precisely, it launches one PathoLogic process for each organism. This allows to increase the speed of draft metabolic network reconstruction when working on multiple organisms.

News of the Year: 2022: Changing mpwt to run the processes independently. Now it can take as input PathoLogic files without fasta. This should make mpwt compatible with PathoLogic files created by EsMeCaTa. Adding the -dump-flat-files-biopax option compatible with Pathway Tools 26.0.

Publication: hal-02395024

Contact: Anne Siegel

Participants: Arnaud Belcour, Anne Siegel, Clemence Frioux, Meziane Aite

7.1.11 TANGO

Keywords: Computational biology, Systems Biology, Metabolic networks, Bacterial strains

Functional Description: The organoleptic properties that provide the added value of fermented dairy products result from specific metabolites that are produced by metabolic processes performed in concert by consortia of microbial species. TANGO enable a deeper understanding of the molecular and cooperative mechanisms underlying the production of organoleptic compounds. Tango uses a combination of whole-genome metabolic modeling and dynamic numerical simulation to assemble a complete, precise model of cheese production using lactic acid and propionic acid bacteria. The results of this modeling reveal interactions between the members of the bacterial community, follow dynamically organoleptic compounds and fit with experimental data.

Contact: Maxime Lecomte

7.2 New platforms

Participants: David Sherman, Ariane Badoual.

As a founding principle, Pleiade supports reproducible scientific analyses and promotes a declarative approach using reusable software modules, rigorous documentation of data provenance, and systematic recording of workflows. The latter is a challenge when interactive interfaces are used, but can be addressed, to cite two examples, in Galaxy by extracting workflows, and in other systems by using Jupyter notebooks. Part of Pleiade's mission is to automate the deployment of environments that support these goals, for non-technical end users.

Pleiade relies on specific computing resources to support our work and that of our collaborators. There are four main use cases:

- Fast deployment of **containerized user environments**, combining biological data and databases, software modules specified by version, a CWL executor, and interactive tools including web front ends, notebooks, or Galaxy. A user environment will provide at least one specific HTTPS endpoint, created dynamically. A single researcher may deploy several different environments in the course of one day.

- Support for **development and testing of workflows**, as above but configured for team members who are developing software modules or interfaces, and who must often deploy several different environments simultaneously.

- Dynamically allocated **containerized compute tasks**, including both individual analysis steps in workflows and GitLab runner containers used for continuous integration. These tasks arrive in bursts that often cannot be planned in advance.

- Long-running **stream preprocessing**, a low-priority background task that watches external databases for changes, chooses pertinent data, precomputes representations and ingests them into local data bases.

We support community best practices for reproducible computing in bioinformatics, using biocontainers generated by bioconda, in CWL or Galaxy workflows. For internal use we provide TES endpoints and host JupyterHub environments.

Pleiade's environment is built on OKD 4, the community distribution of Kubernetes developed alongside of RedHat Openshift. OKD4 in particular uses the CRI-O runtime, not Docker, and containers run unprivileged. Software-defined storage and S3 endpoints are provided by Ceph.

8 New results

8.1 Modeling the metabolic of the gut microbiome in time and space

Participants: Pablo Ugalde-Salas, Simon Labarthe, Clémence Frioux, Coralie Muller,
Arie Wortsman Zurich.

In 2021, an **Inria Exploratory Action** - SLIMMEST – carried out by Simon Labarthe and Clémence Frioux – was initiated. This project aims at combining discrete reasoning models of metabolism to numerical metamodels and PDEs for the simulation of microbial communities in time and space. The selected methodology is the coupling between PDE-based microbial population dynamics model with metamodels of complex optimizations predicting their metabolism. The main difficulty here is to ensure the scalability of the simulation and the selection of relevant metabolic functions and species to be tracked over time. The grant included the participation to the 2021 CEMRACS summer school with a research projet, a postdoc (Pablo Ugalde) and an engineer (Coralie Muller) position.

Metabolic models allow the prediction of the metabolic behaviour of microbial strains from the knowledge of their genomes. Quantitative methods such as Flux Balance Analysis provide a framework to compute metabolic fluxes (i.e. consumption and production rates of metabolites present in the environment of the microbe) and biomass growth at the cost of an optimization problem. When modeling the dynamics of microbial communities, a FBA model must be solved for each microbial strain of the community and at each time step of the dynamics, representing an important computational load.

The aim of the project is to articulate qualitative methods to analyse the functional potential of a microbial community by analysing and simplifying the community-wide metabolic capabilities, to surrogate models of the individual FBA models to provide fast and accurate simulations of the community dynamics. Qualitative methods include Answer Set Programming (ASP) to predict metabolic interactions in the community, metabolic exploration and simplifications. Surrogate models are obtained with a statistical method named ANOVA-RKHS that build a specific RKHS allowing for feature selection and promising trade-off between speed and accuracy.

In 2022, the work achieved during the CEMRACS project has been accepted for publication in the CEMRACS proceedings [29]. Figure 1 illustrates the use of surrogate metamodels learned by ANOVA-RKHS to speed up dynamic flux balance analysis. The team hosted Arie Worstman as intern to work on the project. The project results have been presented by Coralie Muller, Arie Worstman and Pablo Ugalde during several national (JOBIM, journées du GDR MathSav) [35] and international conferences (Newton Institute conference on microbial communities) [36].

In related work concerning gut microbiota, we constructed an individual-based model of epithelial cells interacting with the microbiota-derived chemicals diffusing in the crypt lumen, formalized as a piecewise deterministic Markov process [13]. It accounts for local interactions due to cell contact, and for cell proliferation, differentiation and extrusion that are regulated spatially or by chemical concentrations. It also includes chemicals diffusing and reacting with cells. A deterministic approximated model is also introduced for a large population of small cells, expressed as a system of porous media type equations. Both models were extensively studied through numerical exploration.

8.2 Characterization of Molecular Biodiversity

Participants: Mohammed Anwar Abouabdallah, Romain Peressonni, Alain Franc.

In 2021, Pleiade continued the development and refinement of new methods for chacterizing molecular biodiversity. Two approaches are being pursued, each with a PhD student in their third year.

- The central focus of Mohammed Anwar Abouabdallah's PhD is building OTUs from a pairwise distance matrix using Stochastic Block Models (SBM). Building OTUs is traditionally seen as a form of unsupervised clustering. This work is done in collaboration with the MIAT INRAE research unit
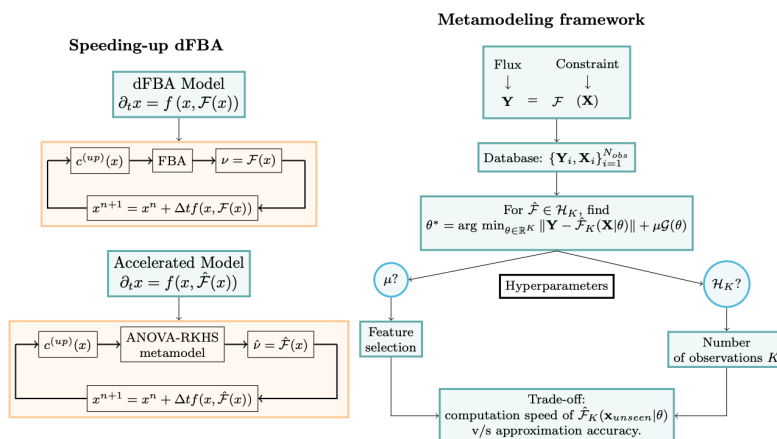
Figure 1: A surrogate metamodel is learned from many simulation runs using ANOVA-RKHS, and used in the place of the FBA model to speed up dynamic flux balance analysis (from [29]).

in Toulouse and HiePACS. It represents a connection between metabarcoding and statistical modeling, a topic which deserves investigation. In [12], we studied whether morphological-based and molecular-based approaches are in agreement, and provided evidence that yes, automatic clustering and group identification can be done reliably using barcoding. Using Aggregative Hierarchical Clustering and Stochastic Block Models, we found that the agreement between morphological-based and molecular-based classifications ranges in most cases from good to very good at taxonomic levels above species (figure 2).

- A major goal of Pleiade is to develop a geometric view of biodiversity. The tool selected up to now is to associate a point cloud to a dataset (pairwise distances between sequences) and to study its shape. This approach has expanded and been developed in 2020 as a collaboration with HiePACS through the cosupervision of Romain Peressonni's PhD, which aims to provide new approaches and algorithms for computing distances between two point clouds.

8.3 Dimension Reduction

Participants: Alain Franc, Jean-Marc Frigerio.

Metabarcoding is a series of technical procedures to build molecular based inventories from large datasets of amplicons. The underlying information needs to be compacted without losing its information content before it can be further processed with domain-specific tools. This links metabarcoding tools to dimension reduction techniques, which is an important topic in Pleiade. This has been implemented through a participation in following research projects:

- Contribution to and finalization of the **ADT Gordon** project in Inria BSO. The objective of this project (partners: Tadaam (coordinator), Storm, HiePACS, Pleiade) is to integrate SVD as a tool available in Chameleon, starPU and new Madeleine. The contribution of Pleiade is to bring in metabarcoding as a use case, and random projection as a method for scaling Multidimensional Scaling (which requires an SVD) in collaboration with HiePACS with a template implemented in Diodon. A MDS on a 106 x 106 matrix has been succesfully run at the end of ADT Gordon, on Occigen, in 900 seconds including I/O.

- A consequence of this involvement is the ADT Diodon, for extending to a diversity of linear dimension reduction techniques what has been aquired in ADT Gordon for MDS, namely a significant
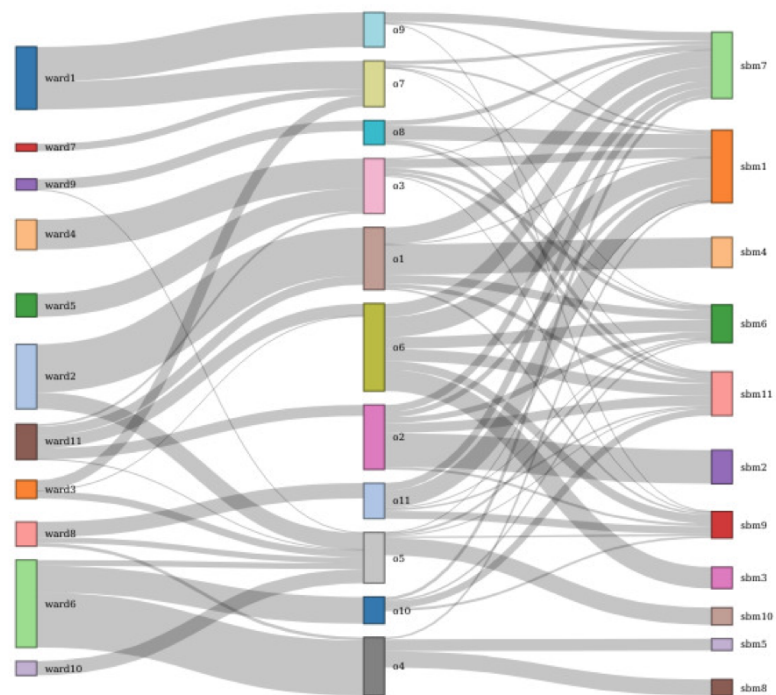
Figure 2: Sankey plot of correspondences between classifications using Aggregative Hierarchical Clustering with Ward (left column), botanical (central column), and Stochastic Block Models (right column), at the Order level. The width of a flow between two classes is proportional to the number of sequences belonging to the two classes (from [12]).

progress in speed and memory management brought by random SVD, which can be integrated into a diversity of methods : PCA, CoA, etc.

- Pleiade is involved in the EU project EOSC-Pillar, in a task for better connecting data to calculation, currently data in Inrae Dataverse system connected to tools running on a INRAE local server or on PlaFRIM, on a testbed on biodiversity assessment with metabarcoding. This task in done in collaboration with INRAE DipSO (Direction Science Ouverte) and the Inria HiePACS project-team.

8.4   Reasoning-based Modeling of interactions and metabolic redundancy in microbial communities

Participants: Maxime Lecomte, David Sherman, Clémence Frioux, Chabname Ghassemi Nedjad.

Over the past few years, we demonstrated how reasoning approaches can prove helpful for the analysis of microbial community metabolism through the Modeling of complementarity accross metabolic networks that permit the selection of minimal communities [49] and more generally screening metabolic potential in microbiomes [40].

These previous works illustrated the need for scalable methods in order to tackle large collections of genome, an objective that is hardly reachable with numerical models. Our recent work focuses on going further in characterizing microbial communities and in particular the interactions that occur among species and with their environment. We take advantage again of the expressivity of the logic paradigm of Answer Set Programming as we provide a model of competitive and cooperative interactions among species in order to compare communities. As such, in the context of Maxime Lecomte's PhD, we developped scores assessing both competition and cooperation potentials based on the genome-scale metabolic networks of microbial species. This work was presented at the French national conference in bioinformatics (JOBIM) [33], and as seminars [23, 22].

A second axis of research relates to the follow up of the Metage2Metabo paper [40] in which we identified interesting patterns of associations in the minimal communities that we computed for groups of metabolic targets using genome-scale metabolic networks of cultivable species from the human gut microbiome. We observed that when compressing association graphs of the assemblies of minimal communities, groups of equivalent bacteria were highlighted, illustrating their similar roles in these communities. We therefore enquired the underlying mechanisms explaining the equivalency between bacteria, what role each group played in the community. To that end, we used Answer Set Programming to query the metabolic potential of each species, in relation to members of its equivalence group, members of other groups or species that were not selected in minimal communities in order to highlight the specific functions responsible for the assemblies. Preliminary results demonstrate that while the combinatorics of shared functions are high, we can suggest functional roles for most groups in microbial assemblies. These results were presented at the French national conference in bioinformatics (JOBIM) and as seminars [20, 31].

8.5   Metabolic Modeling of marine algae: methodological developments and applications

Participants: Clémence Frioux.

Brown algae, especially the species *Ectocarpus siliculosus*, are important models for deciphering the complex interactions within marine holobionts, with the goal of studying their metabolism together with the metabolism of the bacteria that inhabit their direct environment. Because performing wet lab experiments on such systems is technocally challenging, there is need for bioinformatic predictive methods for assessing the putative roles and interdependences between species. In parallel, addressing the difficulties brought by the study of these organisms is also a means to enhance and calibrate the

tools devlopped in the team. Hence a fruitful collaboration for the past years has been developed with scientists from the Roscoff Biological station and the Inria project team Dyliss (Rennes).

In 2022, we published a work investigating the potential for mutualistic and harmful host–microbe interactions affecting brown alga freshwater acclimation [15]. This work combined wet lab experiments and computational models in order to establish how changes in the microbiome composition of the alga can impact the ability of the latter to grow in low-salinity environments.

More generally, the metabolism of algae themselves is challenging to decipher, as it is for most eukaryotic species. A bottleneck resides in the difficulty of annotating their genomes, leading to manual annotations being of importance for these species. However, when aiming at comparing the metabolism of several organisms, it is preferable to annotate all of their genomes homogeneously, usually automatically, in order to prevent the propagation of biases during metabolic reconstruction. This entails ignoring the manual curation work performed on the genomes. AuCoMe is a pipeline aiming at preserving and propagating manual annotations as it permits working with heterogeneously-annotated genomes and efficiently compare their metabolic contents [26]. It was tested on several datasets of genomes, including algal ones, and demonstrated that it builds good quality metabolic networks as it corrects the inconsistencies brought by heterogenous annotation efforts.

8.6 Bacterial swimming in biofilms

Participants: Simon Labarthe.

Biofilms are spatially organized colonies on microorganisms, embedded in a self-produced matrix that confers on the community resistance to environmental stresses. Motile bacteria have been observed swimming in the matrix of pathogenic exogeneous host biofilms, suggesting a potential use of *bacterial swimmers* to increase biofilm vascularization, and thus enhance chemical treatment or biocontrol agent delivery. To study swimmer trajectories in biofilm matrices, we developed a new method for analyzing time-lapse confocal laser scanning images [17], based on the inference of a kinetic model of swimmer populations including mechanistic interactions with the host biofilm. The fitted model allowed us to stratify different swimmer populations by their swimming behavior and to provide insights into the adaptive mechanisms used by micro-swimmers.

9   Bilateral contracts and grants with industry

9.1   Bilateral contracts with industry

Participants: Clémence Frioux, Éloïse Guillem.

An Inria tech contract was conducted with the biotech Ysopia Bioscience between mid 2021 and early 2022. This collaboration was carried out by Éloïse Guillem, engineer of the SED department, and Clémence Frioux.

10   Partnerships and cooperations

10.1   International initiatives

10.1.1   Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

Clémence Frioux is the co-leader of the Inria associate team *SymBioDiversity*, co-led with Anne Siegel from Dyliss (Rennes, France) and associated to the Centre for Mathematical Modelling at the University de Chile (Santiago de Chile), the Mathomics team led by Alejandro Maass.

10.1.2   Participation in other International Programs

Internships Chile: visit of Arie Wortsman Zurich, for 3 months.

10.2   International research visitors

10.2.1   Visits of international scientists

Other international visits to the team

Sébastien Raguideau

Status: Post-Doc

Institution of origin: Earlham Institute

Country: United Kingdom

Dates: November 13th to November 16th

Context of the visit: Scientific collaboration

Mobility program/type of mobility: seminar

10.2.2   Visits to international teams

Research stays abroad

Clémence Frioux

Visited institution: Centro de Modelamiento Matematico

Country: Chile

Dates: May 9th to May 21st 2022

Context of the visit: Inria associated team

Mobility program/type of mobility: conference, workshop and scientific collaboration

Clémence Frioux

Visited institution: Quadram Institute Bioscience

Country: United Kingdom

Dates: October 15th to October 22nd 2022

Context of the visit: scientific collaboration

Maxime Lecomte

Visited institution: Centro de Modelamiento Matematico

Country: Chile

Dates: May 9th to May 22nd 2022

Context of the visit: Inria associated team

Mobility program/type of mobility: conference, workshop and scientific collaboration

10.3   National initiatives

10.3.1   Agence Française pour la Biodiversité

Participants: Alain Franc, Jean-Marc Frigerio.

The AFB is a public law agency of the French Ministry of Ecology that supports public policy in the domains of knowledge, preservation, management, and restoration of biodiversity in terrestrial, aquatic, and marine environments. Pleiade is a partner in two AFB projects developed with the former ONEMA: one funded by ONEMA, the second by labex COTE, where BioGeCo/Pleiade is responsible for data analysis, with implementaton of the tools recently developed for scaling MDS. Calculations have been made on CURTA at MCIA and PlaFRIM at INRIA.

10.3.2   PEPR Agroecology and ICT

Participants: David Sherman, Clémence Frioux, Simon Labarthe, Alain Franc, Jean-Marc Frigerio.

MISTIC, *Microbial communities and ICT,* has been selected as a five-year flagship project in the PEPR Agroecology and ICT program of the French Government. MISTIC will develop methodological tools for defining spatio-temporal models of microbial community dynamics in the phyllosphere and rhizosphere of crop plants, with the goal of creating new understanding of the role of these communities in plant adaptation to environmental stresses, including climate change. MISTIC is a partnership between seven Inria and INRAE teams in Bordeaux, Rennes, and Sophia Antipolis. The project formally began in November 2022.

10.4   Regional initiatives

10.4.1   Poppy Rosa ProtoImpact

Participants: David Sherman.

Poppy Rosa ProtoImpact is an educational robots project financed by the Conseil Régional de la Nouvelle Aquitaine and coordinated by the Ligue de l'Enseignement de la Gironde, with Poppy Station and Inria as partners. Poppy Rosa is prototyping a new way of teaching concepts of algorithmics and artificial intelligence to young students, with the goal of training tomorrow's citizens. Our deployment specifically targets rural schools across the Neo-Aquitaine region.

11   Dissemination

11.1   Promoting scientific activities

11.1.1   Scientific events: organisation

Member of the organizing committees

- Co-organisation of a mini-symposium at JOBIM 2022, the French national conference of bioinformatics. [Clémence Frioux]

11.1.2   Scientific events: selection

Member of the conference program committees

- 20th International Conference on Computational Methods in Systems Biology (CMSB 2022) [Clémence Frioux]

- Journées GDR maths bio santé 2022 [Simon Labarthe]

Reviewer

- 20th International Conference on Computational Methods in Systems Biology (CMSB 2022) [Clémence Frioux]

11.1.3   Journal

Member of the editorial boards

- BMC Evolutionary Biology [Alain Franc]

Reviewer - reviewing activities

- Nature Communication [Clémence Frioux]

- Microbiome [Clémence Frioux]

- iScience [Clémence Frioux]

- BMC Bioinformatics [Clémence Frioux]

- mSystems [Simon Labarthe]

11.1.4   Invited talks

- Workshop SymBioDiversity, CMM, U. de Chile (Santiago de Chile, May 2022) - *Characterizing the human gut microbiome through the lens of metagenomics and metabolic models* [Clémence Frioux]

- Workshop SymBioDiversity, CMM, U. de Chile (Santiago de Chile, May 2022) - *Characterizing microbial interactions in controlled and natural microbial communities* [Maxime Lecomte]

- Journées Scientifiques Inria Chile (Santiago de Chile, May 2022) - *Symbiodiversity associated team* [Clémence Frioux]

- Bordeaux Bioinformatics Center (Bordeaux, France, January 2022) - *Logic modelling of metabolism: from individual networks to communities* [Clémence Frioux]

- INRIA-DFKI Workshop on AI (Bordeaux, France, October 2022) – *Opportunities for learning and explaining microbiomes in agroecology* [David Sherman]

- Journées Française de la nutrition (JFN, Toulouse, France, November 2022) - *Omic-based model of fiber degradation.* [Simon Labarthe]

- Breizh CarnoTech (Rennes, France, November 2022) – *MISTIC: Microbial communities and TIC* [David Sherman]

11.1.5   Leadership within the scientific community

- David Sherman is on the steering committee of Biosena, a regional research network of the New Aquitaine region dedicated to Biodiversity and Ecosystemic Services. Biosena associates actors from the academic and socio-economic sectors, with the goal of contributing to the understanding and preservation of biodiversity and to the improvement of ecosystemic services. Biosena contributes to this goal through research, knowledge dissemination, outreach, and skill transfer in the form of Research Action, in keeping with the recommendations of Ecobiose.

- David Sherman is member of the board (membre du Conseil d'administration) and secretary of the Mobsya Association, Lausanne. Mobsya develops and commercializes the Thymio educational robot, geared towards K-12.

- David Sherman is member of the board (membre du Conseil d'Administration) and lead advisor for software of the Poppy Station Association. Poppy Station develops open-hardware open-source humanoid robots for research and education.

11.1.6   Scientific expertise

Recruitment committees

- Junior researcher selection committee of the Inria Centre at the University of Bordeaux [Clémence Frioux]

- Assistant professor selection committee Univ. Côte d'Azur [Clémence Frioux]

11.1.7   Research administration

Local responsabilites

- Co-creation of and participation in a gender equality and diversity working group in the Inria Centre at the University of Bordeaux [Clémence Frioux]

11.2   Teaching - Supervision - Juries

11.2.1   Teaching

- Master – ENSTBB Bordeaux INP - Bioinformatics [Clémence Frioux]

- Master – ENSEIRB Bordeaux INP - Research algorithms [Clémence Frioux]

- Graduate School – Doctoral School Computer Science, Mathematics, Univ. Bordeaux - Answer Set Programming [Clémence Frioux]

11.2.2   Supervision

- PhD of Maxime Lecomte (2020-2023) - *Approches hybrides en modélisation logique et numérique du métabolisme des écosystèmes microbiens* - David Sherman (Director), Hélène Falentin (director, INRAE STLO), Clémence Frioux (advisor)

- PhD of Chabname Ghassemi Nedjad (2022-2025) - *Combinatorial optimisation problems for reverse ecology* - Clémence Frioux (co-director), Loïc Paulevé (CNRS, LaBRI, co-director).

- PhD of Mohamed Anwar Abouabdallah (2020-2023) - *Approche Tenseur-Train pour l'inférence dans les modèles à blocs stochastiques, application à la caractérisation de la biodiversité* - Alain Franc, Olivier Coulaud and Nathalie Peyrard

- Master's internship of Arie Wortsman Zurich - *Metamodelling of genome-scale metabolic networks.* 3 months – co-supervised by Simon Labarthe and Clémence Frioux

- Master's internship of Quentin Chapuzet - *Temporal logic in Answer Set Programming for the modelling of microbial communities.* 6 months – co-supervised by Maxime Lecomte and David Sherman

- Master's internship of Chabname Ghassemi Nedjad - *Explainable models of minimal microbial community assembly.* 6 months – supervised by Clémence Frioux

- Apprenticeship of Léonard Brindel - *Platform for metagenome assembly* – co-supervised by Clémence Frioux and Franck Salin

11.2.3 Juries

PhD defense juries

- Antoine Régimbeau - Univ. Nantes [Clémence Frioux *examiner*]

- Déborah Boyenval - Univ. Côte d'Azur - [Clémence Frioux *examiner*]

PhD thesis committee

- Mathieu Bolteau - Univ. Nantes [Clémence Frioux]

- Coralie Rousseau - Sorbonne Univ. [Clémence Frioux]

- Maxime Mahout - Univ. Paris-Saclay [Clémence Frioux]

11.3 Popularization

11.3.1 Education

- Hosting of 5 students in the *troisième* grade for a week of observation

11.3.2 Interventions

- Chiche ! Un ou une scientifique, une classe – 5 classes de seconde, 4 classes de troisième. [Clémence Frioux]

- Cordées de la réussite *1 scientifique, 1 classe : Chiche!* [Clémence Frioux]

- Cordées de la réussite *Qui se ressemble s'assemble* [David Sherman]

- Fête de la Science : Circuit Scientifique Bordelais – cinq ateliers *Qui se ressemble s'assemble* [David Sherman]

12 Scientific production

12.1 Major publications

[1] P. Almeida, C. Gonçalves, S. Teixeira, D. Libkind, M. Bontrager, I. Masneu-Pomarède, W. Albertin, P. Durrens, D. J. Sherman, P. Marullo, C. Todd Hittinger, P. Gonçalves and J. P. Sampaio. 'A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum.' In: *Nature Communications* 5 (2014), p. 4044. DOI: 10.1038/ncomms5044. URL: https://hal.inria.fr/hal-01002466.

[2] R. Assar, M. A. Montecino, A. Maass and D. J. Sherman. 'Modeling acclimatization by hybrid systems: Condition changes alter biological system behavior models'. In: *BioSystems* 121 (June 2014), pp. 43–53. DOI: 10.1016/j.biosystems.2014.05.007. URL: https://hal.inria.fr/hal-01002987.

[3] M. Bahram, T. Netherway, C. Frioux, P. Ferretti, L. P. Coelho, S. Geisen, P. Bork and F. Hildebrand. 'Metagenomic assessment of the global distribution of bacteria and fungi'. In: *Environmental Microbiology* (13th Nov. 2020). DOI: 10.1111/1462-2920.15314. URL: https://hal.inria.fr/hal-03033570.

[4]   A. Belcour, C. Frioux, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species'. In: *eLife* 9 (29th Dec. 2020). DOI: 10.1101/803056. URL: https://hal.inria.fr/hal-02395024.

[5]   B. Burgunter-Delamare, H. Kleinjan, C. Frioux, E. Fremy, M. Wagner, E. Corre, A. Le Salver, C. Leroux, C. Leblanc, C. Boyen, A. Siegel and S. Dittami. 'Metabolic Complementarity Between a Brown Alga and Associated Cultivable Bacteria Provide Indications of Beneficial Interactions'. In: *Frontiers in Marine Science* 7 (21st Feb. 2020), pp. 1–11. DOI: 10.3389/fmars.2020.00085. URL: https://hal.inria.fr/hal-02866101.

[6]   S. M. Dittami, E. Corre, L. Brillet-Guéguen, A. Lipinska, N. Pontoizeau, M. Aite, K. Avia, C. Caron, C. H. Cho, J. Collen, A. Cormier, L. Delage, S. Doubleau, C. Frioux, A. Gobet, I. González-Navarrete, A. Groisillier, C. Herve, D. Jollivet, H. Kleinjan, C. Leblanc, X. Liu, D. Marie, G. V. Markov, A. E. Minoche, M. Monsoor, P. Péricard, M.-M. Perrineau, A. F. Peters, A. Siegel, A. Siméon, C. Trottier, H. S. Yoon, H. Himmelbauer, C. Boyen and T. Tonon. 'The genome of Ectocarpus subulatus – A highly stress-tolerant brown alga'. In: *Marine Genomics* 52 (Jan. 2020), p. 100740. DOI: 10.1016/j.margen.2020.100740. URL: https://hal.inria.fr/hal-02866117.

[7]   C. Frioux, S. Dittami and A. Siegel. 'Using automated reasoning to explore the metabolism of unconventional organisms: a first step to explore host–microbial interactions'. In: *Biochemical Society Transactions* 48.3 (7th May 2020), pp. 901–913. DOI: 10.1042/BST20190667. URL: https://hal.archives-ouvertes.fr/hal-02569935.

[8]   C. Frioux, D. Singh, T. Korcsmaros and F. Hildebrand. 'From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes'. In: *Computational and Structural Biotechnology Journal* (June 2020). DOI: 10.1016/j.csbj.2020.06.028. URL: https://hal.inria.fr/hal-02883309.

[9]   F. Leese, A. Bouchez, K. Abarenkov, F. Altermatt, A. Borja, K. Bruce, T. Ekrem, F. Ćiampor, Z. Ćiampor, F. Costa, S. Duarte, V. Elbrecht, D. Fontaneto, A. A. Franc, M. Geiger, D. Hering, M. Kahlert, B. Kalamujić Stroil, M. Kelly, E. Keskin, I. Liska, P. Mergen, K. Meissner, J. Pawlowski, L. Penev, Y. Reyjol, A. Rotter, D. Steinke, B. van der Wal, S. S. Vitecek, J. Zimmermann and A. Weigand. 'Why We Need Sustainable Networks Bridging Countries, Disciplines, Cultures and Generations for Aquatic Biomonitoring 2.0: A Perspective Derived From the DNAqua-Net COST Action'. In: *Next Generation Biomonitoring: Part 1*. Vol. 58. Elsevier, 2018, pp. 63–99. URL: https://hal.inria.fr/hal-01984996.

[10] N. D. P. Peyrard, M.-J. Cros, S. De Givry, A. A. Franc, S. S. Robin, R. R. Sabbadin, T. Schiex and M. Vignes. 'Exact or approximate inference in graphical models: why the choice is dictated by the treewidth, and how variable elimination can be exploited'. In: *Australian and New Zealand Journal of Statistics* 61.2 (June 2019). to appear, pp. 89–133. DOI: 10.1111/anzs.12257. URL: https://hal.inria.fr/hal-02433018.

[11] D. J. Sherman, T. Martin, M. Nikolski, C. Cayla, J.-L. Souciet and P. Durrens. 'Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes.' In: *Nucleic Acids Research* 37 (2009), pp. D550–D554. DOI: 10.1093/nar/gkn859. URL: https://hal.inria.fr/inria-00341578.

12.2   Publications of the year

International journals

[12] M. A. Abouabdallah, N. Peyrard and A. Franc. 'Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study'. In: *Molecular Ecology Resources* (2022). DOI: 10.1111/1755-0998.13579. URL: https://hal.inrae.fr/hal-03546609.

[13] L. Darrigade, M. Haghebaert, C. Cherbuy, S. Labarthe and B. Laroche. 'A PDMP model of the epithelial cell turn-over in the intestinal crypt including microbiota-derived regulations'. In: *Journal of Mathematical Biology* 84.7 (June 2022), pp. 1–67. DOI: 10.1007/s00285-022-01766-8. URL: https://hal.science/hal-03379023.

[14]   P. Fournier, L. Pellan, D. Barroso-Bergadà, D. Bohan, T. Candresse, F. Delmotte, M.-C. Dufour, V. Lauvergeat, C. Le Marrec, A. Marais, G. Martins, I. Masneuf-Pomarède, P. Rey, D. J. Sherman, P. This, C. Frioux, S. Labarthe and C. Vacher. 'The functional microbiome of grapevine throughout plant evolutionary history and lifetime'. In: *Advances in Ecological Research*. Advances in Ecological Research 67 (2022), pp. 27–99. DOI: 10.1016/bs.aecr.2022.09.001. URL: https://hal.inrae.fr/hal-03858395.

[15]   H. Kleinjan, C. Frioux, G. Califano, M. Aite, E. Fremy, E. Karimi, E. Corre, T. Wichard, A. Siegel, C. Boyen and S. M. Dittami. 'Insights into the potential for mutualistic and harmful host-microbe interactions affecting brown alga freshwater acclimation'. In: *Molecular Ecology* (2022). DOI: 10.1111/mec.16766. URL: https://hal.science/hal-03868898.

[16]   C. Muller, B. Serrano-Solano, Y. Sun, C. Tischer and J.-K. Hériché. 'The Image Data Explorer: Interactive exploration of image-derived data'. In: *PLoS ONE* 17.9 (15th Sept. 2022), e0273698. DOI: 10.1371/journal.pone.0273698. URL: https://hal.inria.fr/hal-03830356.

[17]   G. Ravel, M. Bergmann, A. Trubuil, J. Deschamps, R. Briandet and S. Labarthe. 'Inferring characteristics of bacterial swimming in biofilm matrix from time-lapse confocal laser scanning microscopy'. In: *eLife* 11 (14th June 2022). DOI: 10.7554/eLife.76513. URL: https://hal.inrae.fr/hal-03695580.

Conferences without proceedings

[18]   W. Cao, M. Lecomte, S. Le Fur, J. J. Aubert, M.-B. Maillard, A. Nicolas, S.-M. Deutsch, S. Parayre, F. Boissel, A. Leduc, A. Kerjouh, M. Harel-Oger, G. Garric, C. Frioux, D. J. Sherman, S. Labarthe, A. Thierry and H. Falentin. 'Metatranscriptomics and metabolic modeling to identify bacterial metabolic interactions during the manufacture of a model pressed cheese.' In: FOODMICRO 2022. Athènes, Greece, 20th Sept. 2022. URL: https://hal.inrae.fr/hal-03781287.

[19]   W. Cao, M. Lecomte, S. Le Fur, J. J. Aubert, M.-B. Maillard, A. Nicolas, S.-M. Deutsch, S. Parayre, F. Boissel, A. Leduc, A. Kerjouh, M. Harel-Oger, G. Garric, C. Frioux, D. J. Sherman, S. Labarthe, A. Thierry and H. Falentin. 'Métatranscriptomique et modèle métabolique pour identifier la succession des métabolismes bactériens en interaction lors de la fabrication d'un fromage modèle à pâte pressée.' In: 23ème édition du colloque du Club des Bactéries Lactiques. Rennes, France, 13th June 2022. URL: https://hal.inrae.fr/hal-03694338.

[20]   C. Ghassemi Nedjad, C. Frioux and L. Paulevé. 'A logic programming exploration of metabolic networks to explain microbial community models'. In: Journée GT Bioss 2022. Nantes, France, 17th Nov. 2022. URL: https://hal.inria.fr/hal-03910483.

[21]   M. Iannacito, O. Coulaud and A. Franc. 'Extension of Correspondence Analysis to multiway datasets through HOSVD: a geometric framework'. In: MDS 2022 - SIAM Conference on Mathematics of Data Science. San Diego / Hybrid, United States, 26th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03826894.

[22]   M. Lecomte, S. Labarthe, D. J. Sherman, H. Falentin and C. Frioux. 'Characterizing microbial interactions in controlled and natural microbial communities'. In: Workshop SymBioDiversity. Santiago, Chile, 16th May 2022. URL: https://hal.inrae.fr/hal-03857848.

[23]   M. Lecomte, D. J. Sherman, H. Falentin and C. Frioux. 'Reasoning approaches for the characterization of cooperation and competition in large-scale microbial communities'. In: Journée GT Bioss 2022. Nantes, France, 17th Nov. 2022. URL: https://hal.inrae.fr/hal-03857781.

Scientific book chapters

[24]   C. Frioux and A. Siegel. 'Problèmes d'optimisation combinatoire pour l'étude du métabolisme'. In: *Approches symboliques de la modélisation et de l'analyse des systèmes biologiques*. Encyclopédie Sciences. ISTE éditions, July 2022. DOI: 10.51926/ISTE.9029.ch2. URL: https://hal.inria.fr/hal-03885249.

Reports & preprints

[25] E. Agullo, O. Coulaud, A. Denis, M. Faverge, A. Franc, J.-M. Frigerio, N. Furmento, A. Guilbaud, E. Jeannot, R. Peressoni, F. Pruvost and S. Thibault. *Task-based randomized singular value decomposition and multidimensional scaling*. RR-9482. Inria Bordeaux - Sud Ouest; Inrae - BioGeCo, 9th Sept. 2022, p. 37. URL: https://hal.inria.fr/hal-03773985.

[26] A. Belcour, J. Got, M. Aite, L. Delage, J. Collen, C. Frioux, C. Leblanc, S. Dittami, S. Blanquart, G. Markov and A. Siegel. *AuCoMe: inferring and comparing metabolisms across heterogeneous sets of annotated genomes*. 15th Sept. 2022. DOI: 10.1101/2022.06.14.496215. URL: https://hal.science/hal-03778267.

[27] A. Belcour, B. Ruiz, C. Frioux, S. Blanquart and A. Siegel. *EsMeCaTa: Estimating metabolic capabilities from taxonomic affiliations*. 2022. DOI: 10.1101/2022.03.16.484574. URL: https://hal.science/hal-03697249.

[28] A. Franc. *Linear Dimensionality Reduction*. 9488. Inria Bordeaux Sud-Ouest, 23rd Sept. 2022, p. 69. URL: https://hal.inria.fr/hal-03784623.

[29] C. Frioux, S. Huet, S. Labarthe, J. Martinelli, T. Malou, D. J. Sherman, M.-L. Taupin and P. Ugalde-Salas. *Accelerating metabolic models evaluation with statistical metamodels: application to Salmonella infection models*. Feb. 2022. URL: https://hal.inrae.fr/hal-03635862.

[30] S. Labarthe, S. Plancade, S. Raguideau, F. P. Onate, E. Le Chatelier, M. Leclerc and B. Laroche. *Four functional profiles for fibre and mucin metabolism in the human gut microbiome*. 2nd Jan. 2023. URL: https://hal.inrae.fr/hal-03918193.

Other scientific publications

[31] C. Ghassemi Nedjad and C. Frioux. 'Analyzing and modelling functions carried by key species in minimal microbial communities'. In: JOBIM 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03910438.

[32] S. Labarthe, B. Laroche, B. Polizzi, F. Patout and M. Ribot. 'A multi-scale epidemic model of enteric infection with heterogeneous shedding resulting from host-microbiota-pathogen interactions'. In: ECMTB 2022 - 12th European Conference on Mathematical and Theoretical Biology. Heidelberg, Germany, 19th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03834709.

[33] M. Lecomte, D. J. Sherman, H. Falentin and C. Frioux. 'Characterising competition and cooperation potentials in microbial communities using discrete models of metabolism. An exploratory analysis'. In: JOBIM 2022 - JOURNÉES OUVERTES EN BIOLOGIE, INFORMATIQUE ET MATHÉMATIQUES. Rennes, France, 5th July 2022. URL: https://hal.inrae.fr/hal-03839337.

[34] T. Malou, S. Labarthe, B. Laroche, E. Vergu and F. Deslandes. 'Pest detection by inversion of a pheromone dispersion model'. In: ECMTB 2022 - 12th European Conference on Mathematical and Theoretical Biology. Heidelberg, Germany, 19th Sept. 2022. URL: https://hal.archives-ouvertes.fr/hal-03777604.

[35] C. Muller, A. Worstman, P. Ugalde-Salas, C. Frioux and S. Labarthe. 'Modelling the dynamics of Salmonella infection in the gut at the bacterial and host levels'. In: JOBIM 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Rennes, France, 5th July 2022. URL: https://hal.inria.fr/hal-03839374.

[36] C. Muller, A. Wortsman, P. Ugalde-Salas, C. Frioux and S. Labarthe. 'Modelling the dynamics of Salmonella infection in the gut at the bacterial and the host level'. In: Microbial communities: current approaches and open challenges. Cambridge, United Kingdom, 10th Oct. 2022. URL: https://hal.inria.fr/hal-03839405.

12.3 Cited publications

[37]   M. Aite, M. Chevallier, C. Frioux, C. Trottier, J. Got, M. P. Cortés, S. N. Mendoza, G. Carrier, O. Dameron, N. Guillaudeux, M. Latorre, N. Loira, G. V. Markov, A. Maass and A. Siegel. 'Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models'. In: *PLOS Computational Biology* 14.5 (2018), e1006146. DOI: `10.1371/journal.pcbi.1006146`.

[38]   B. Arnold, R. Corbett-Detig, D. Hartl and K. Bomblies. 'RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling'. In: *Mol. Ecol.* 22.11 (2013), pp. 3179–90.

[39]   M. Bakonyi and C. R. Johnson. 'The Euclidean Distance Matrix Completion Problem'. In: *SIAM J. Matrix Anal. App.* 16.2 (1995), pp. 646–654.

[40]   A. Belcour, C. Frioux, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Metage2Metabo, microbiota-scale metabolic complementarity for the identication of key species'. In: *eLife* 9 (2020), e61968. DOI: `10.7554/elife.61968`.

[41]   M. Bourgin, S. Labarthe, A. Kriaa, M. Lhomme, P. Gérard, P. Lesnik, B. Laroche, E. Maguin and M. M. Rhimi. 'Exploring the Bacterial Impact on Cholesterol Cycle: A Numerical Study'. In: *Frontiers in Microbiology* 11 (2020), p. 1121. DOI: `10.3389/fmicb.2020.01121`. URL: `https://hal.science/hal-02863236`.

[42]   E. J. Candès and B. Recht. 'Exact Matrix Completion via Convex Optimization'. In: *Found. Comput. Math.* 9 (2009), pp. 717–772.

[43]   C. Combes. *Parasitism: The Ecology and Evolution of Intimate Interactions*. University of Chicago Press, 2001.

[44]   'Comparative genomics of protoploid Saccharomycetaceae'. In: *Genome Research* 19 (2009), pp. 1696–1709. DOI: `10.1101/gr.091546.109`. URL: `http://hal.inria.fr/inria-00407511/en/`.

[45]   B. Dujon, D. Sherman, G. Fischer, P. Durrens et al. 'Genome evolution in yeasts'. In: *Nature* 430 (2004), pp. 35–44. DOI: `10.1038/nature02579`. URL: `http://www.nature.com/nature/journal/v430/n6995/abs/nature02579.html`.

[46]   O. Ebenhöh, T. Handorf and R. Heinrich. 'Structural analysis of expanding metabolic networks.' In: *Genome informatics. International Conference on Genome Informatics* 15.1 (2004), pp. 35–45.

[47]   A. Franc. *Tensor Ranks for the Pedestrian for Dimension Reduction and Disentangling Interactions*. Research Report RR-9445. Inrae - BioGeCo ; Inria Bordeaux Sud-Ouest, Dec. 2021. URL: `https://hal.inria.fr/hal-03518107`.

[48]   A. A. Franc, P. Blanchard and O. Coulaud. 'Nonlinear mapping and distance geometry'. In: *Optimization Letters* 14.2 (May 2019), pp. 453–467. DOI: `10.1007/s11590-019-01431-y`. URL: `https://hal.inria.fr/hal-02124882`.

[49]   C. Frioux, E. Fremy, C. Trottier and A. Siegel. 'Scalable and exhaustive screening of metabolic functions carried out by microbial consortia'. In: *Bioinformatics* 34.17 (2018), pp. i934–i943. DOI: `10.1093/bioinformatics/bty588`.

[50]   C. Frioux, T. Schaub, S. Schellhorn, A. Siegel and P. Wanko. 'Hybrid metabolic network completion'. In: *Theory and Practice of Logic Programming* 19.1 (2019), pp. 83–108. DOI: `10.1017/s1471068418000455`.

[51]   P. Gayral, J. Melo-Ferreira and S. Glemin. 'Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap'. In: *PLoS Genetic* 9.4 (2013). e1003457.

[52]   M. Gebser, R. Kaminski, B. Kaufmann and T. Schaub. 'Clingo = ASP + Control: Preliminary Report'. In: *CoRR* abs/1405.3694 (2014).

[53]   M. Gebser, R. Kaminski, A. Konig and T. Schaub. 'Advances in gringo Series 3'. In: *LPNMR*. Vol. 6645. Lecture Notes in Computer Science. Springer, 2011, pp. 345–351.

[54]   'Génolevures – a novel approach to 'evolutionary genomics''. In: *FEBS Letters (complete issue)* 487.1 (2000). URL: `https://febs.onlinelibrary.wiley.com/toc/18733468/2000/487/1`.

[55] S. Labarthe, B. Polizzi, T. Phan, T. Goudon, M. Ribot and B. Laroche. 'A mathematical model to investigate the key drivers of the biogeography of the colon microbiota.' In: *Journal of Theoretical Biology* 462.7 (2019), pp. 552–581. DOI: 10.1016/j.jtbi.2018.12.009. URL: https://hal.science/hal-01761191.

[56] L. Liberti, C. Lavor, N. Maculan and A. Mucherino. 'Euclidean Distance Geometry and Applications'. In: *SIAM review* 56(1) (2014), pp. 3–69.

[57] M. Lynch. 'Estimation of Nucleotide Diversity, Disequilibrium Coefficients, and Mutation Rates from High-Coverage Genome-Sequencing Projects'. In: *Mol. Biol. Evol.* 25.11 (2008), pp. 2409–19.

[58] J. D. Orth, I. Thiele and B. Ø. Palsson. 'What is flux balance analysis?' In: *Nature Biotechnology* 28.3 (2010), pp. 245–248. DOI: 10.1038/nbt.1614.

[59] S. Prigent, C. Frioux, S. M. Dittami, S. Thiele, A. Larhlimi, G. Collet, F. Gutknecht, J. Got, D. Eveillard, J. Bourdon, F. Plewniak, T. Tonon and A. Siegel. 'Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks'. In: *PLOS Computational Biology* 13.1 (2017), e1005276. DOI: 10.1371/journal.pcbi.1005276.

[60] R. E. Ricklefs. 'A comprehensive framework for global patterns in biodiversity'. In: *Ecology Letters* 7.1 (2004), pp. 1–15. DOI: 10.1046/j.1461-0248.2003.00554.x. URL: http://dx.doi.org/10.1046/j.1461-0248.2003.00554.x.

[61] D. J. Sherman, T. Martin, M. Nikolski, C. Cayla, J.-L. Souciet and P. Durrens. 'Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes'. In: *Nucleic Acids Research* 37.suppl 1 (2009), pp. D550–D554. DOI: 10.1093/nar/gkn859. eprint: http://nar.oxfordjournals.org/content/37/suppl_1/D550.full.pdf+html. URL: http://nar.oxfordjournals.org/content/37/suppl_1/D550.abstract.

[62] D. J. Sherman, P. Durrens, F. Iragne, E. Beyne, M. Nikolski and J.-L. Souciet. 'Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts.' In: *Nucleic Acids Research* 34.Database issue (Jan. 2006), pp. D432–5. DOI: 10.1093/nar/gkj160. URL: https://hal.archives-ouvertes.fr/hal-00118142.

[63] B. de Thoisy, O. Duron, L. Epelboin, L. Musset, P. Quénel, B. Roche, F. Binetruy, S. Briolant, L. Carvalho, A. Chavy, P. Couppié, M. P. Demar, M. Douine, I. Dusfour, S. Gourbière, Y. Epelboin, C. Flamand, A. Franc, M. Ginouvès, E. S. Houël, A. Kocher, A. Lavergne, P. Le Turnier, L. Mathieu, J. Murienne, M. Nacher, R. Schaub, S. Pelleau, G. Prévot, D. Rousset, E. Roux, S. Talaga, P. Thill, S. Tirera and J.-F. Guégan. 'Ecology, evolution, and epidemiology of zoonotic and vector-borne infectious diseases in French Guiana: Transdisciplinarity does matter to tackle new emerging threats'. In: *Infection, Genetics and Evolution* 93 (Sept. 2021), p. 104916. DOI: 10.1016/j.meegid.2021.104916. URL: https://hal-pasteur.archives-ouvertes.fr/pasteur-03261181.

[64] S. Tirera, B. de Thoisy, D. Donato, C. Bouchier, V. Lacoste, A. Franc and A. Lavergne. 'The influence of habitat on viral diversity in neotropical rodent hosts'. In: *Viruses* 13.9 (Aug. 2021), pp. 1–29. DOI: 10.3390/v13091690. URL: https://hal.inrae.fr/hal-03370479.