RESEARCH CENTRE

Inria Centre at Rennes University

IN PARTNERSHIP WITH:

CNRS, Université de Rennes

2024

ACTIVITY REPORT

Project-Team

DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

IN COLLABORATION WITH: Institut de recherche en informatique et systèmes aléatoires (IRISA)

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Inria

Contents

Project-Team DYLISS

Creation of the Project-Team: 2013 July 01

Keywords

Computer sciences and digital sciences

A3.1.1. – Modeling, representation

A3.1.2. – Data management, quering and storage

A3.1.7. – Open data

A3.1.8. – Big data (production, storage, transfer)

A3.1.11. – Structured data

A3.2.2. – Knowledge extraction, cleaning

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.3.2. – Data mining

A3.4. – Machine learning and statistics

A6.1.3. – Discrete Modeling (multi-agent, people centered)

A7.3.1. – Computational models and calculability

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B1.1.1. – Structural biology

B1.1.4. – Genetics and genomics

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B1.1.11. – Plant Biology

B2.2.3. – Cancer

B2.2.5. – Immune system diseases

1 Team members, visitors, external collaborators

Research Scientists

- Anne Siegel [Team leader, CNRS, Senior Researcher, team leader from Nov 2024, HDR]
- Samuel Blanquart [INRIA, Researcher]
- François Coste [INRIA, Researcher]

Faculty Members

- Emmanuelle Becker [UNIV RENNES, Professor, HDR]
- Catherine Belleannée [UNIV RENNES, Associate Professor]
- Myriam Bontonou [UNIV RENNES, Associate Professor, from Sep 2024, (Enseignante-chercheuse contractuelle)]
- Olivier Dameron [UNIV RENNES, Professor, Team leader until november 2024, HDR]
- Yann Le Cunff [UNIV RENNES, Associate Professor]

PhD Students

- Moana Aulagner [INRIA]
- Cécile Beust [UNIV RENNES]
- Océane Carpentier [UNIV RENNES, from Sep 2024]
- Elisa Chenel [UNIV RENNES, from Oct 2024]
- Pablo Espana Gutierrez [ENS RENNES]
- Juliette Francis [UNIV RENNES, from Oct 2024]
- Pauline Giraud [UNIV RENNES, from Nov 2024]
- Ulysse Le Clanche [UNIV RENNES, from Oct 2024]
- Corentin Lucas [INRIA]
- Baptiste Ruiz [INRIA, until Sep 2024]
- Kerian Thuillier [CNRS, until Sep 2024]
- Yael Tirlet [UNIV RENNES]

Technical Staff

- Pauline Giraud [CNRS, Engineer, until Oct 2024]
- Jeanne Got [CNRS, Engineer]
- Victor Mataigne [CNRS, Engineer]
- Noé Robert [CNRS, Engineer]

Interns and Apprentices

- Elisa Chenel [UNIV RENNES, Intern, from Feb 2024 until Jul 2024]
- Eoghan Chevé [ENS RENNES, Intern, from Jun 2024 until Jul 2024]
- Juliette Francis [UNIV RENNES, Intern, from Feb 2024 until Aug 2024]
- Hugo Mingarelli [UNIV RENNES, Intern, from Feb 2024 until Jul 2024]
- Noryah Safla [UNIV RENNES, Intern, from Apr 2024 until Jun 2024]

Administrative Assistant

- Marie Le Roic [INRIA]

External Collaborators

- Matthieu Bougueon [CNRS, until Sep 2024, Post-doc (DI-ENS)]
- François Moreews [INRAE, Engineer staff, 20% time dedicated to the team]
- Nathalie Théret [INSERM, Senior Researcher (IRSET), HDR]

2 Overall objectives

Bioinformatics context: from life data science to functional information about biological systems and unconventional species. Sequence analysis and systems biology both consist in the interpretation of biological information at the molecular level, that concern mainly intra-cellular compounds. Analyzing genome-level information is the main issue of **sequence analysis**. The ultimate goal here is to build a full catalogue of bio-products together with their functions, and to provide efficient methods to characterize such bio-products in genomic sequences. In regards, contextual physiological information includes all cell events that can be observed when a perturbation is performed over a living system. Analyzing contextual physiological information is the main issue of **systems biology**.

For a long time, computational methods developed within sequence analysis and dynamical modeling had few interplay. However, the emergence and the democratization of new sequencing technologies (NGS, metagenomics) provides information to link systems with genomic sequences. In this research area, the Dyliss team focuses on linking genomic sequence analysis and systems biology. **Our main applicative goal in biology is to characterize groups of genetic actors that control the phenotypic response of species when challenged by their environment**. **Our main computational goals are to develop methods for analyzing the dynamical response of a biological system, modeling and classifying families of gene products with sensitive and expressive languages, and identifying the main actors of a biological system within static interaction maps**. We first formalize and integrate in a set of logical or grammatical constraints both generic knowledge information (literature-based regulatory pathways, diversity of molecular functions, DNA patterns associated with molecular mechanisms) and species-specific information (physiological response to perturbations, sequencing...). We then rely on symbolic methods (Semantic Web technologies for data integration, querying as well as for reasoning with bio-ontologies, solving combinatorial optimization problems, formal classification) to compute the main features of the space of admissible models.

Computational challenges. The main challenges we face are **data incompleteness and heterogeneity, leading to non-identifiability**. Indeed, we have observed that the biological systems that we consider cannot be uniquely identifiable. Indeed, "omics" technologies have allowed the number of measured compounds in a system to increase tremendously. However, it appears that the theoretical number of different experimental measurements required to integrate these compounds in a single discriminative model has increased exponentially with respect to the number of measured compounds. Therefore, according to the current state of knowledge, there is no possibility to explain the data with a single model.

Our rationale is that biological systems will still remain non-identifiable for a very long time. In this context, we favor **the construction and the study of a space of feasible models or hypotheses**, including known constraints and facts on a living system, rather than searching for a single discriminative optimized model. We develop methods allowing a precise and exhaustive investigation of this space of hypotheses. With this strategy, we are in the position of developing experimental strategies to progressively shrink the space of hypotheses and increase the understanding of the system.

Bioinformatics challenges. Our objectives in computer sciences are developed within the team in order to fit with three main bioinformatics challenges (1) data-science and knowledge-science for life sciences (see Section 3.2); (2) understanding metabolism (see Section 3.3); (3) characterizing regulatory and signaling phenotypes (see Section 3.4).

Implementing methods in software and platforms. Seven platforms have been developed in the team during the last five years: Askomics, AuReMe, FinGoc, Caspo, Cadbiom, Logol and Protomata. They aim at guiding the user to progressively reduce the space of models (families of sequences of genes or proteins, families of keys actors involved in a system response or dynamical models) which are compatible with both the knowledge and experimental observations. Most of our platforms are developed with the support of the GenOuest resource and data center hosted in the IRISA laboratory, including their computer facilities [More info]

3 Research program

3.1 Context: Computer science perspective on symbolic artificial intelligence

We develop methods that use an explicit representation of the relationships between heterogeneous data and knowledge in order to construct a space of hypotheses. Therefore, our objective in computer science is mainly to develop accurate representations (oriented graphs, Boolean networks, automata, or expressive grammars) to iteratively capture the complexity of a biological system.

Integrating data with querying languages: Semantic web for life sciences The first level of complexity in the data integration process consists in confronting heterogeneous datasets. Both the size and the heretogeneity of life science data make their integration and analysis by domain experts impractical and prone to the streetlight effect (they will pick up the models that best match what they know or what they would like to discover). Our first objective involves the formalization and management of symbolic knowledge, that is, the explicitation of relations occurring in structured data. In this setting, our main goal is to facilitate and optimize the integration of Semantic Web resources with local users data by relying on the implicit data scheme contained in biological data and Semantic Web resources.

Reasoning over structured data with constraint-based logical paradigms Another level of complexity in life science integration is that very few paradigms exist to model the behavior of a complex biological system. This leads biologists to perform and formulate hypotheses in order to interpret their data. Our strategy is to interpret such hypotheses as combinatorial optimization problems, allowing to reduce the family of models compatible with data. To that goal, we collaborate with Potsdam University in order to use and challenge the most recent developments of Answer Set Programming (ASP) [54], a logical paradigm for solving constraint satisfiability and combinatorial optimization issues.

Our goal is therefore to provide scalable and expressive formal models of queries on biological networks with the focus of integrating dynamical information as explicit logical constraints in the modeling process.

Characterizing biological sequences with formal syntactic models Our last goal is to identify and characterize the function of expressed genes such as transcripts, enzymes or isoforms in non-model species biological networks or specific functional features of metagenomic samples. These are insufficiently precise because of the divergence of biological sequences, the complexity of molecular structures and biological processes, and the weak signals characterizing these elements.

Our goal is therefore to develop accurate formal syntactic models (automata, grammars or abstract gene models) that would enable us to represent sequence conservation, sets of short and degenerated patterns, and crossing or distant dependencies. This requires both to determine the classes of formal

syntactic models adequate for handling biological complexity, and to automatically characterize the functional potential embodied in biological sequences with these models.

3.2 Scalable methods to query data heterogenity

Confronted to large and complex data sets (raw data are associated with graphs depicting explicit or implicit links and correlations) almost all scientific fields have been impacted by the *big data issue*, especially genomics and astronomy [66]. In our opinion, life sciences cumulate several features that are very specific and prevent the direct application of big data strategies that proved successful in other domains such as experimental physics: the existence of **several scales of granularity** (from microscopic to macroscopic) and the associated issue of dependency propagation, datasets **incompleteness and uncertainty** (including highly **heterogeneous** responses to a perturbation from one sample to another), and highly fragmented sources of information that **lacks interoperability** [52]. To explore this research field, we use techniques from symbolic data mining (Semantic Web technologies, symbolic clustering, constraint satisfaction, and grammatical modeling) to take into account those life science features in the analysis of biological data.

3.2.1 Research topics

Facilitating data integration and querying The quantity and inner complexity of life science data require semantically-rich analysis methods. A major challenge is then to combine data (from local project as well as from reference databases) and symbolic knowledge seamlessly. Semantic Web technologies (RDF for annotating data, OWL for representing symbolic knowledge, and SPARQL for querying) provide a relevant framework, as demonstrated by the success of Linked (Open) Data [34]. However, life science end users (1) find it difficult to learn the languages for representing and querying Semantic Web data, and consequently (2) miss the possibility they had to interact with their tabulated data (even when doing so was exceedingly slow and tedious). Our first objective in this axis is to develop accurate abstractions of datasets or knowledge repositories to facilitate their exploration with RDF-based technologies.

Scalability of semantic web queries. A bottleneck in data querying is given by the performance of federated SPARQL queries, which must be improved by several orders of magnitude to allow current massive data to be analyzed. In this direction, our research program focuses on the combination of *linked data fragments* [72], query properties and dataset structure for decomposing federated SPARQL queries.

Building and compressing static maps of interacting compounds A final approach to handle heterogeneity is to gather multi-scale data knowledge into a functional static map of biological models that can be analyzed and/or compressed. This requires to link genomics, metabolomics, expression data and protein measurement of several phenotypes into unified frameworks. In this direction, our main goal is to develop families of constraints, inspired by symbolic dynamical systems, to link datasets together. We currently focus on health (personalized medicine) and environmental (role of non-coding regulations, graph compression) datasets.

3.2.2 Associated software tools

AskOmics platform AskOmics is an integration and interrogation software for linked biological data based on semantic web technologies[1]. AskOmics aims at bridging the gap between end user data and the Linked (Open) Data cloud (LOD cloud). It allows heterogeneous bioinformatics data (formatted as tabular files or directly in RDF) to be loaded into a Triple Store system using a user-friendly web interface. It helps end users (1) to take advantage of the information available in the LOD cloud for analyzing their own data, and (2) to contribute back to the linked data by representing their data and the associated metadata in the proper format, as well as by linking them to other resources. An originality is the graphical interface that allows any dataset to be integrated in a local RDF datawarehouse and SPARQL query to be built transparently and iteratively by a non-expert user.
Pax2graphml aims at easily manipulating BioPAX source files as regulated reaction graphs described in graph format. The goal is to be highly flexible and to integrate graphs of regulated reactions from a single

[1] askomics.org

BioPAX source or by combining and filtering BioPAX sources. The output graphs can then be analyzed with additional tools developed in the team, such as KeyRegulatorFinder.

FinGoc-tools The FinGoc tools allow filtering interaction networks with graph-based optimization criteria in order to elucidate the main regulators of an observed phenotype. The main added-value of these tools is the functionality allowing to make explicit the criteria used to highlight the role of the main regulators. (1) The KeyRegulatorFinder package searches key regulators of lists of molecules (like metabolites, enzymes or genes) by taking advantage of knowledge databases in cell metabolism and signaling[2]. (2) The PowerGrasp python package implements graph compression methods oriented toward visualization, and based on power graph analysis[3]. (3) The iggy package enables the repairing of an interaction graph with respect to expression data[4].

3.3 Metabolism: from protein sequences to systems ecology

Our research in bioinformatics in relation with metabolic processes is driven by the need to understand non-model (eukaryote) species. Their metabolism have acquired specific features that we wish to identify with computational methods. To that goal, we combine sequence analysis with metabolic network analysis, with the final goal to understand better the metabolism of communities of organisms.

3.3.1 Research topics

Genomic level: characterizing functions of protein sequences Precise characterization of functional proteins, such as enzymes or transporters, is a key to better understand and predict the actors involved in a metabolic process. In order to improve the precision of functional annotations, we develop machine learning approaches that take a sample of functional sequences as input and infer a model representing their key syntactical characteristics, including dependencies between residues.

System level: enriching and comparing metabolic networks for non-model organisms

Non-model organisms often lack both complete and reliable annotated sequences, which cause the draft networks of their metabolism to largely suffer from incompleteness. In former studies, the team has developed several methods to improve the quality of eukaryotic metabolic networks, by solving several variants of the so-called *Metabolic Network gap-filling problem* with logical programming approaches [10, 9]. The main drawback of these approaches is that they cannot scale to the reconstruction and comparison of families of metabolic networks. Our main objective is therefore to develop new tools for the comparison of species strains at the metabolic level.

Consortium level: exploring the diversity of community consortia The newly emerging field of system ecology aims at building predictive models of species interactions within an ecosystem, with the goal of deciphering cooperative and competitive relationships between species [51]. This field raises two new issues: (1) uncertainty on the species present in the ecosystem and (2) uncertainty about the global objective governing an ecosystem. To address these challenges, our first research focus is the inference of metabolic exchanges and relationships for transporter identification, based on our expertise in metabolic network gap-filling. The second challenging focus is the prediction of transporters families via refined characterization of transporters, which are quite unexplored apart from specific databases [64].

3.3.2 Associated software tools

Protomata[5] is a machine learning suite for the inference of automata characterizing (functional) families of proteins at the sequence level. It provides programs to build a new kind of sequence alignments (characterized as partial and local), learn automata, and search for new family members in sequence databases. By enabling to model local dependencies between positions, automata are more expressive than classical tools (PSSMs, Profile HMMs, or Prosite Patterns) and are well suited to predict new family members with a high specificity. This suite is for instance embedded in the cyanolase database [41] to

[2] biowic.inria.fr/
[3] github.com/aluriak/powergrasp
[4] bioasp.github.io/iggy/
[5] protomata-learner.genouest.org

automate its updade and was used for refining the classification of HAD enzymes [6] or identify shared conservations in the core proteome of extracellular vesicles produced by human and animal *S. aureus* strains [69].

PPSuite[6] is one of the first frameworks taking into account coevolutionary dependencies between residues for the comparison of protein sequences. It proposes a complete workflow enabling to infer direct couplings between the positions of a sequence of interest by a Potts model with the help of the sequence close homologs and to score the similarity of the sequences by alignment of the inferred Potts models, as well as tools to visualize the models and their alignments [68, 67].

AuReMe and AuCoMe workspaces is designed for tractable reconstruction of metabolic networks[7]. The toolbox allows for the Automatic Reconstruction of Metabolic networks based on the combination of multiple heterogeneous data and knowledge sources [1]. The main added values are the inclusion of graph-based tools relevant for the study of non-model organisms (Meneco and Menetools packages), the possibility to trace the reconstruction and curation procedures (Padmet package), and the exploration of reconstructed metabolic networks with wikis (wiki-export package, see: aureme.genouest.org/wiki.html) [33]. It also generates outputs to explore the resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae [62], extremophile bacteria [44] and communities of organisms [4].

Mpwt, emmapper2gbk is a Python package for running Pathway Tools[8] on multiple genomes using multiprocessing. Pathway Tools is a comprehensive systems biology software system that is associated with the BioCyc database collection[9]. Pathway Tools is frequently used for reconstructing metabolic networks. In order to allow the output of the eggnoggmapper annotation tool to be used by Mpwt, we also developed emmaper2gbk to create relevant genome files.

Metage2metabo is a Python tool to perform graph-based metabolic analysis starting from annotated genomes (reference genomes or metagenome-assembled genomes) [31]. It uses Mpwt to reconstruct metabolic networks for a large number of genomes. The obtained metabolic networks are then analyzed individually and collectively in order to get the added value of metabolic cooperation in microbiota over individual metabolism and to identify and screen interesting organisms among all.

3.4 Regulation and signaling: detecting complex and discriminant signatures of phenotypes

On the contrary to metabolic networks, regulatory and signaling processes in biological systems involve agents interacting at different granularity levels (from genes, non-coding RNAs to protein complexes) and different time-scales. Our focus is on the reconstruction of large-scale networks involving multiple scales processes, from which controllers can be extracted with symbolic dynamical systems methods. Particular attention is paid to the characterization of products of genes (such as isoform) and of perturbations to identify discriminant signature of pathologies.

3.4.1 Research topics

Genomic level: characterizing gene structure with grammatical languages and conservation information The goal here is to accurately represent gene structure, including intron/exon structure, for predicting the products of genes, such as isoform transcripts, and comparing the expression potential of a eukaryotic gene according to its context (e.g. tissue) or according to the species. Our approach consists in designing grammatical and comparative-genomics based models for gene structures able to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [36]. Accurate gene models are defined by identifying general constraints shaping gene families and their structures conserved over evolution. Syntactic elements controlling gene expression (transcription factor binding sites controlling transcription; enhancers and

[6] www-dyliss.irisa.fr/ppalign/
[7] aureme.genouest.org/
[8] bioinformatics.ai.sri.com/ptools/
[9] biocyc.org

silencers controlling splicing events...), i.e. short, degenerated and overlapping functional sequences, are modeled by relying on the high capability of SVG grammars to deal with structure and ambiguity [65].

System level: extracting causal signatures of complex phenotypes with systems biology frameworks Our main challenge is to set up a generic formalism to model inter-layer interactions in large-scale biological networks. To that goal, we have developed several types of abstractions: multi-experiments framework to learn and control signaling networks [11], multi-layer reactions in interaction graphs [37], and multi-layer information in large-scale Petri nets [30]. Our main issues are to scale these approaches to standardized large-scale repositories by relying on the interoperable Linked Open Data (LOD) resources and to enrich them with ad-hoc regulations extracted from sequence-based analysis. This will allow us to characterize changes in system attractors induced by mutations and how they may be included in pathology signatures.

3.4.2 Associated software tools

Logol software is designed for complex pattern modeling and matching[10]. It is a swiss-army-knife for pattern matching on DNA/RNA/Protein sequences, based on expressive patterns which consist in a complex combination of motifs (such as degenerated strings) and structures (such as imperfect stem-loop ou repeats) [2]. Logol key features are the possibilities (i) to divide a pattern description into several sub-patterns, (ii) to model long range dependencies, and (iii) to enable the use of ambiguous models or to permit the inclusion of negative conditions in a pattern definition. Therefore, Logol encompasses most of the features of specialized tools (Vmatch, Patmatch, Cutadapt, HMM) and enables interplays between several classes of patterns (motifs and structures), including stem-loop identification in CRISPR.

Caspo Cell ASP Optimizer (Caspo) software constitutes a pipeline for automated reasoning on logical signaling networks (learning, classifying, designing experimental perturbations, identifying controllers, take time-series into account)[11]. The software handles inherent experimental noise by enumerating all different logical networks which are compatible with a set of experimental observations [11]. The main advantage is that it enables a complete study of logical network without requiring any linear constraint programs.

Cadbiom package aims at building and analyzing the asynchronous dynamics of enriched logical networks[12]. It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [30]. For example, it allowed to analyze controler of phenotypes in a large-scale knowledge database (PID) [5].

Recently, we have significantly refactored Cadbiom package towards a framework that allows the identification of causal regulators in large-scale models, formalized in the BioPAX language and automatically interpreted as guarded transitions. The Cadbiom framework was applied to the BioPAX version of two ressources (PID, KEGG) of the PathwayCommons database and to the Atlas of Cancer Signalling Network (ACSN). As a case-study, it was used to characterize the causal signatures of markers of the epithelial-mesenchymal transition.

4 Application domains

In terms of transfer and societal impact, we consider that our role is to develop fruitful collaborations with biology laboratories in order to consolidate their studies by a smart use of our tools and prototypes and to generate new biological hypotheses to be tested experimentally.

Marine Biology: seaweed enzymes and metabolism An important field of study is **marine biology**, as it is a transversal field covering challenges in integrative biology, dynamical systems and sequence analysis.

- **Protein functions in seaweed metabolism** Several years ago, our methods based on combinatorial optimization for the reconstruction of genome-scale metabolic networks and on classification of

[10] logol.genouest.org/
[11] bioasp.github.io/caspo/
[12] cadbiom.genouest.org

enzyme families based on local and partial alignments allowed the seaweed *E. siliculosus* metabolism to be deciphered [62, 45]. The study of the *HAD* superfamily of proteins thanks to partial local alignments produced by Protomata tools, allowed sub-families to be deciphered and classified. Additionally, the metabolic map reconstructed with Meneco enabled the reannotation of 56 genes within the *E. siliculosus* genome. These approaches also shed light on evolution of metabolic processes.

- **Elucidating algal metabolism thanks to large-scale metabolic network reconstructions** More recently, the tools developed by Dyliss (based on the AuReMe toolbox) allowed us to participate in the reconstruction of a metabolic network for the brown algae *Saccharina japonica* and *Cladosiphon okamuranus* in order to identify these species specificities on the synthesis of carotenoids biosynthesis [60]. We also participated in the study of the genome of *Ectocarpus subulatus*, a highly stress-tolerant algal strain [50]. Finally, AuReMe has been used to analyze the metabolic capacity of several strains of cyanobacteria, with results integrated in the Cyanorak database [53] and to characterize synergistic effects of the *synechococcus* strain WH7803 [56].

- **Metabolic pathway drift theory** Genome annotations can contribute to understanding algal metabolism. The tool PathModel was developed to add support for biochemical reactions and metabolite structures to the theory of metabolic pathway drift with an approach combining chemoinformatics knowledge reasoning and modeling. This approach was applied to the study of the red alga *Chondrus crispus*, which allowed to show that even for metabolic pathways supposed to be conserved between species (sterols, mycrosporins synthesis), we can see an important turnover in the order of reactions appearing in a metabolic pathway. This work lays the foundations for the concept of "metabolic drift" analogous to the same concept in genomics. [32].

- **Algal-bacteria interactions** We reconstructed the metabolic network of a symbiot bacterium *Ca. P. ectocarpi* [49] and used this reconstructed network to decipher interactions within the algal-bacteria holobiont, revealing several candidates metabolic pathways for algal-bacterial interactions. Similarily, our analyses suggested that the bacterium *Ca. P. ectocarpi* is able to provide both beta-alanine and vitamin B5 to the seaweed via the phosphopantothenate biosynthesis pathway [63].

  These works paved the way to the study of host-microbial interactions, as shown in [42] where we evidenced the role of tools such as miscoto and metage2metabo to predict synthetic communities allowing to restore algal metabolic pathways. To validate these approaches experimentally, we worked with S. Dittami, researcher at the Roscoff biological station. We applied these methods on a set of about fifteen cultivable bacteria identified on the wall membrane of *Ectocarpus siliculosus*. Our approaches predicted that three bacteria were necessary to facilitate the growth of this alga in an axenic medium. The experiments were carried out, and indeed allowed the alga to grow in an axenic medium. This is therefore a proof of concept of the relevance of our approaches. More recently, the study of the freshwater strain of *Ectocarpus subulatus* evidenced the role of metabolism in adaptation, paving the way to biotechnological applications [58].

Microbiology: elucidating the functioning of extremophile consortiums of bacteria. Our main issue is the understanding of bacteria living in extreme environments. The context is mainly a collaboration with the group of bioinformatics at Universidad de Chile (co-funded by the Center of Mathematical Modeling, the Center of Regulation Genomics and Inria-Chile). In order to elucidate the main characteristics of these bacteria, our integrative methods were developed to identify the main groups of regulators for their specific response in their living environment. The integrative biology tools Meneco, Lombarde and Shogen have been designed in this context. In particular, genome-scale metabolic network been recently reconstructed and studied with the Meneco and Shogen approaches, especially on bacteria involved in biomining processes [38] and in Salmon pathogenicity [44]. We have also studied the specificities of two Microbacterium strains, CGR1 and CGR2, isolated in different soils of the Atacama Desert in Chile, showing significant differences on the connectivity of metabolite production in relation to pH tolerance and CO2 production [59].

Agriculture and environmental sciences: upstream controllers of cow, pork and pea-aphid metabolism and regulation. Our goal is to propose methods to identify regulators of complex phenotypes related to

environmental issues. Our work on the identification of upstream regulators within large-scale knowledge databases (tool KeyRegulatorFinder) [37] and on semantic-based analysis of metabolic networks [35] was very valuable for interpreting the differences of gene expression in pork meat [57] and figure out the main gene-regulators of the response of porks to several diets [55]. Our expertise in microbiota analysis is also currently being applied to rumen microbial genomics [61].

Health: Dynamics of microenvironment in chronic liver diseases We develop methods and models to understand the dynamics of the microenvironment in order to propose evolutionary markers and effective therapeutic targets. The matrix microenvironment is the major regulator of events related to fibrosis-cirrhosis-cancer progression and Hepatic Stellate Cells (HSC) are the main actors of microenvironment remodeling. At molecular level, the transforming growth factor TGF-β plays a central role by promoting HSC activation, extracellular matrix remodeling and epithelial-mesenchymal transition. In that context we have developed three programs :

- *TGF-β signaling networks.* TGF-β is a multifunctional cytokine that binds to specific receptors and induce numerous signaling pathways depending on the context. Deciphering TGF-β signaling networks requires to take into account a system-wide view and develop predictive models for therapeutic benefit. For that purpose we developed Cadbiom and identified gene networks associated with innate immune response to viral infection that combine TGF-β and interleukin signaling pathways [30, 43]. More recently we have very significantly refactored Cadbiom package towards a framework that allows the identification of causal regulators in large-scale models, formalized in the BioPAX language and automatically interpreted as guarded transitions[13].The Cadbiom framework was applied to the BioPAX version of two resources (PID,KEGG) of the Pathway Commons database and to the Atlas of Cancer Signalling Network (ACSN). As a case-study, it was used to characterize the causal signatures of markers of the epithelial-mesenchymal transition.

- *Functional signature for ADAMTS.* Hepatic Stellate Cells produce a wide variety of molecules involved in ECM remodeling, such as adamalysins [70]. However, the limitations of discovering new functions of these proteins stem from the experimental approaches that are difficult to implement due to their structure and biochemical features. In that context we developed an original framework combining the identification of small modules in conserved regions independent of known domains and the concepts of phylogenomics (association of conservation and phenotype gained concurrently during evolution). The resulting evolutionary model of motif signatures and protein-protein interaction signatures of the ADAMTS family is validated by data from literature and provides biologists with many new potential functional motifs [46], [48], [47].

- *Dynamic model of hepatic stellate cells.* To characterize the dynamics of HSC activation upon TGFB1 stimulation, we developed a model using Kappa, a site graph rewriting language and its static analyzer Kasa [40]. We previously demonstrated the advantages of Kappa language for modeling TGF-β signaling and extracellular matrix [71]. Unlike previous model based on a population of interacting proteins, we now develop an original Kappa model based on a population of cells interacting with TGF-β [39]. The model recapitulates the dynamics of activation of HSC towards myofibroblast states and the reversion processes. Current work aims to identify the regulators of the repair likely to promote the resolution of fibrosis at the expense of its progression.

5 Social and environmental responsibility

5.1 Footprint of research activities

Dyliss research activities have low environmental footprints. Most of our software solution run on off-the-shelf computers and are not computationally intensive. Indirectly, the analyses and predictions we make intend to reduce the need for long, costly technically or ethically difficult biological experiments.

[13] cadbiom.genouest.org

5.2 Impact of research results

Through our ongoing collaborations with INSERM and Rennes' Hospital, Dyliss research activities have a social impact on human health. Our collaborations with INRAe have a direct impact on vegetal and animal health, and an indirect impact in environment as these projects original motivation is to reduce fertilizers or pesticides.

6 Highlights of the year

6.1 Awards

Camille Juigné (PhD student from 2020 to 2023) won the Open Science Award[14] (*Prix de la science ouverte*) for her work on "Integration and analysis of heterogeneous biological data through multilayer graph exploitation to gain deeper insights into feed efficiency variations in growing pigs". She was co-supervized by E. Becker (DYLISS) and F. Gondret (INRAe Pegase).

6.2 Workshop organization

François Moreews co-organized the WemSemPilot[15] workshop (02–03 July 2024, Rennes), a national workshop and hackathon on interoperability and Semantic Web technology and their application to agriculture and life sciences. The event gathered 30 participants from academia or industry.

7 New software, platforms, open data

7.1 New software

7.1.1 AskOmics

Name: Convert tabulated data into RDF and create SPARQL queries intuitively and "on the fly".

Keywords: RDF, SPARQL, Querying, Graph, LOD - Linked open data

Functional Description: AskOmics aims at bridging the gap between end user data and the Linked (Open) Data cloud. It allows heterogeneous bioinformatics data (formatted as tabular files) to be loaded in a RDF triplestore and then be transparently and interactively queried. AskOmics is made of three software blocks: (1) a web interface for data import, allowing the creation of a local triplestore from user's datasheets and standard data, (2) an interactive web interface allowing "à la carte" query-building, (3) a server performing interactions with local and distant triplestores (queries execution, management of users parameters).

URL: https://askomics.org/

Contact: Olivier Dameron

Partners: Université de Rennes 1, CNRS, INRA

7.1.2 Regulus

Keywords: Bioinformatics, Semantic Web

Functional Description: Regulus, software which computes TF-gene regulation relations (TF for transcriptional regulatory) and qualify them as activation or inhibition. Its main principles are to (1) take into account regulatory factors (TFs and regions) activities, (2) discretize the activities into patterns, (3) produce signed circuits inferred by testing biological constraints likelihood and (4) be easly reusable and applicable to many datasets.

[14] https://www.enseignementsup-recherche.gouv.fr/fr/remise-des-premiers-prix-science-ouverte-de-la-these-97810
[15] https://pegase.pages.mia.inra.fr/wspilot/index_en.html

News of the Year: Regulus is published.

URL: https://gitlab.com/teamDyliss/regulus

Publication: hal-04443527

Contact: Olivier Dameron

Participants: Marine Louarn, Guillaume Collet, Eve Barre, Thierry Fest, Olivier Dameron, Anne Siegel, Fabrice Chatonnet

Partners: Université de Rennes 1, INSERM, University of British Columbia, Université de Lyon

7.1.3 Merrin

Name: MEtabolic Regulation Rule INference from time series data

Keywords: Bioinformatics, Rule-based programming, ASP - Answer Set Programming

Functional Description: Merrin is Python3 software to infer regulatory rules of a regulated metabolic network (RMN) from possibly incomplete and noisy observed time series. Merrin is an extension of Caspo (https://github.com/bioasp/caspo) focused on the inference of regulations from indirect measures on metabolism. Merrin takes as input a metabolic network in SBML format, a set of regulatory proteins P, a set of observed time series in CSV format, and a Prior Knowledge Network (PKN) in a text file format. Merrin generates a CSV file describing the inferred regulatory networks.

URL: https://github.com/bioasp/merrin

Publication: hal-03701755

Contact: Anne Siegel

Participants: Kerian Thuillier, Caroline Baroukh, Alexander Bockmayr, Ludovic Cottret, Loic Paulevé, Anne Siegel

Partners: Université de Rennes 1, Université de Bordeaux, Université de Toulouse, Freie Universität Berlin

7.1.4 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 keytstone species in the production of these compounds are identified.

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

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.5 AuCoMe

Name: Automatic Comparison of Metabolisms

Keywords: Bioinformatics, Workflow, Metabolic networks, Omic data, Data analysis

Functional Description: AuCoMe is a Python package that aims at reconstructing homogeneous metabolic networks and pan-metabolism starting from genomes with heterogeneous levels of annotations. Four steps are composing AuCoMe. 1) It automatically infers annotated genomes from draft metabolic networks thanks to Pathway Tools and MPWT. 2) The Gene-Protein-Reaction (GPR) associations previously obtained are propagated to protein orthogroups in using Orthofinder and, an additional robustness criteria. 3) AuCoMe checking the presence of supplementary GPR associations by finding missing annotation in all genomes. In this step, the tools BlastP, TblastN and, Exonerate are called. 4) It adding spontaneous reactions to metabolic pathways that were completed by the previous steps. AuCoMe generates several outputs to facilitate the analysis of results: tabuled files, SBML files, PADMET files, supervenn and a dendogram of reactions.

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

Publication: hal-03778267

Contact: Anne Siegel

Participants: Arnaud Belcour, Jeanne Got, Meziane Aite, Ludovic Delage, Jonas Collen, Clemence Frioux, Catherine Leblanc, Simon M. Dittami, Samuel Blanquart, Gabriel V. Markov, Anne Siegel

7.1.6 EsMeCaTa

Name: EsMeCaTa

Keywords: Metabolism, Taxonomies, Evolution, Metagenomics

Functional Description: Sequencing methods allowed to detect organisms in environment samples. It is possible to identify the groups in which the organisms belong. But to infer metabolic networks and estimate metabolic interactions between organisms, new methods are needed. That's why we developed EsMeCaTa. This method takes as input taxonomic affiliations showing which organisms are present in the environment and then query UniProt to extract proteomes associated with these organisms. Then, the proteins are clustered into groups to detect if a cluster is present in multiple proteomes. It is possible to extract clusters which contain proteins present in multiple proteomes. Their functional annotations are then retrieved on UniProt, which allows to estimate the metabolic capacities of the taxon.

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

Publication: hal-03697249

Contact: Samuel Blanquart

7.1.7 SPARTA

Name: Shifting Paradigms to Annotation Representation from Taxonomy to identify Archetypes

Keywords: Bioinformatics, Microbiota, Classification

Functional Description: SPARTA can be launched following two steps. The first can be called with the "sparta esmecata'" command. This command takes as input a taxonomic abundance table and launches a run of the EsMeCaTa pipeline. The second part of the pipeline can be called with the "sparta classification" command. It takes as input a file containing the labels associated with each sample in the dataset, a functional and taxonomic description of the samples, a description of each taxon's affiliation, and a table indicating the occurrence of functions in each organism. This last step involves the training of 20 successive Random Forest classifiers.

URL: https://github.com/baptisteruiz/SPARTA

Publication: hal-04816962

Contact: Yann Le Cunff

7.1.8 Transformer Framework for Protein Characterization - EnzBert

Keywords: Deep learning, Transformer, Functional annotation, Proteins, Biological sequences

Scientific Description: A generic framework for the specialization of a pre-trained transformer protein language model for classification or regression tasks.

Functional Description: Given examples of annotated sequences, this tool allows to train and analyse resulting models with respect to evaluation metrics (accuracy, correlation) plots and the importance of the residues for the inference. The process is fully automated and the whole operation can be done by modifying a JSON configuration file and providing a JSON data set. No code skills are thus required. This framework enabled training the EnzBert model which predicts the Enzyme Commission number of protein sequences.

News of the Year: A new Enzbert model was trained and tested with the addition of two enzyme families of interest to the ABIE team at the Station Biologique de Roscoff.

URL: https://gitlab.inria.fr/nbuton/tfpc

Publications: tel-04347632, hal-04382475

Contact: Nicolas Buton

Participants: Nicolas Buton, François Coste, Yann Le Cunff, Noe Robert

7.1.9 EnzBert-GO

Keywords: Proteins, Biological sequences, Functional annotation, Deep learning, Ontologies

Scientific Description: Code for learning and using BERT deep neural architectures for the prediction of multi-level and multi-class functional enzymatic GO annotations of protein sequences.

Functional Description: Prediction of the functional enzymatic GO annotations of protein sequences

News of the Year: EnzBert-GO is a fork of code originally developed by Nicolas Buton. It implements multi-class and multi-level enzymatic function prediction to address Enzbert's main practical limitations and utilize Gene Ontology's hierarchical annotations.

URL: https://gitlab.inria.fr/fcoste/enzbert-go

Contact: François Coste

Participant: François Coste

7.1.10 FUSE-PhyloTree

Name: FUnctions and SEquence conservations on a Phylogenetic Tree

Keywords: Bioinformatics, Biological sequences, Sequence alignment, Phylogenomics, Proteins

Scientific Description: FUSE-PhyloTree is dedicated to estimate the sequence regions which are potentially associated to functions of interest in a multi-functional protein families, such as paralogous and multi-domain protein families. The method uses state-of -the-art programs to estimate a mapping of both the ancestral functions and the ancestral sequence content at each node in the phylogenetic family tree. It enables the association of functions with local sequence conservations through the inference of their co-appearance along the evolutionary gene tree, and it generates interactive Itol representations allowing to explore the annotated tree.

Functional Description: FUSE-PhyloTree is dedicated to studying multi-functional protein families, such as paralogous and multi-domain protein families. It enables the association of functions with local sequence conservations through the inference and exploration of their ancestral co-appearance within the evolutionary tree of genes.

News of the Year: To make the package public, we simplified the outputs, improved the documentation, and added helper scripts to facilitate its use by non-specialists. Additionally, we now provide a Singularity container alongside the Docker container to enable wider distribution.

URL: https://github.com/OcMalde/fuse-phylotree

Publication: hal-04248728

Contact: François Coste

Participants: Olivier Dennler, Elisa Chenel, François Coste, Samuel Blanquart, Catherine Belleannée, Nathalie Theret

8 New results

8.1 Scalable methods to query data heterogeneity

Participants: Emmanuelle Becker, Cécile Beust, Océane Carpentier, Olivier Dameron, Ulysse Le Clanche, Victor Mataigne, François Moreews, Anne Siegel, Nathalie Théret, Yael Tirlet.

Generic and queryable data integration schema for transcriptomics and epigenomics studies [*E. Becker, O. Dameron, Y. Tirlet*] [17]

- The expansion of multi-omics datasets raises significant challenges for data integration and querying. To overcome these challenges, we developed a generic RDF-based integration schema that connects various types of differential -omics data, epigenomics, and regulatory information [17]. It is designed to be fully or partially populated, providing both flexibility and extensibility while supporting complex queries. We validated the schema by reproducing two recently published studies, one in biomedicine and the other in environmental science, proving its genericity and its ability to integrate data efficiently. Along with the integration schema, we provided a library of SPARQL queries supporting analyses.

BioPAX in 2024: Where we are and where we are heading [*E. Becker, C. Beust, O. Dameron, N. Théret*] [12]

- In systems biology, understanding biological pathways is essential to deciphering complex biological systems. The growing availability of pathway data through online databases has highlighted the need for standardization, addressed by the BioPAX format. While BioPAX is highly expressive, allowing to finely describe biological pathways at the molecular and cellular levels, but the associated intrinsic complexity may be an obstacle to its widespread adoption. We examined how various pathway databases utilize BioPAX for data standardization and identified ways to enhance its effectiveness, including tools and software improvements. Additionally, we introduced an original abstraction concept for BioPAX graphs, enabling targeted analysis of specific graph regions. This approach distinguishes the format's role in representation from the abstraction's role in contextual analysis.

BioPAX-Explorer: a Python Object-Oriented framework for overcoming the complexity of querying biological networks [*F. Moreews, E.Becker, O. Dameron*] [26]

- While public databanks increasingly provide datasets in BioPAX format, their use remains below potential. Users may encounter challenges in harnessing the data due to the BioPAX intricately detailed underlying model. Moreover, extracting data demands specific technical skills, posing a barrier for many potential users. To address these obstacles, we developped BioPAX-Explorer. This tool is designed to facilitate the adoption and usage of BioPAX for extracting data or build algorithms and models, within the Python community. BioPAX-Explorer is a Python package that provides an object-oriented data model automatically generated from the BioPAX OWL specification. Moreover, it offers expressive query capabilities that shield users from BioPAX inner complexity.

Phenotypes extraction from text: analysis and perspective in the LLM era [*O. Dameron*] [19]

- Collecting the relevant list of patient phenotypes can significantly improve the final diagnosis. As textual clinical reports are the richest source of phenotypes information, their automatic extraction is a critical task. The main challenges of this Information Extraction task are to identify precisely the text spans related to a phenotype and to link them unequivocally to referenced entities from a source such as the Human Phenotype Ontology (HPO). Recently, Language Models (LMs) have been the most successful approach for extracting phenotypes from clinical reports. Solutions such as PhenoBERT, relying on BERT or GPT, have shown promising results when applied to datasets built on the hypothesis that most phenotypes are explicitly mentioned in the text. However, this assumption is not always true in medical genetics. We conducted an in-depth analysis of PhenoBERT, one of the best existing solutions, to evaluate the proportion of phenotypes predicted with simple string-matching. We demonstrated how utilizing and incorporating large language models (LLMs) for span detection step can improve performance especially with implicit phenotypes. In addition, this experiment revealed that the annotations of existing dataset are not exhaustive, and that LLM can identify relevant spans missed by human labelers.

8.2 Metabolism: from protein sequences to systems ecology

Participants: Moana Aulagner, Catherine Belleannée, Samuel Blanquart, Myriam Bontonou, Matthieu Bouguéon, Elisa Chenel, Eoghan Chevé, François Coste, Pablo Espana Gutierrez, Pauline Giraud, Jeanne Got, Yann Le Cunff, Victor Mataigne, Hugo Mingarelli, Noé Robert, Baptiste Ruiz, Anne Siegel, Nathalie Théret, Yael Tirlet.

Interpretable functional classification of microbiomes and detection of hidden cumulative effects [*S. Blanquart, Y. Le Cunff, B. Ruiz, A. Siegel*] [16, 21]

- We have developed SPARTA, a computational pipeline for integrating the functional annotation of the gut microbiota into an automatic classification process and facilitating downstream interpretation of its results. The process takes as input taxonomic composition data and links each component to its functional annotations through interrogation of the UniProt database with the EsMeCaTa tool. Both profiles, microbial and functional, are used to train Random Forest classifiers

to discern unhealthy from control samples. SPARTA ensures full reproducibility and exploration of inherent variability by extending state-of-the-art methods in three dimensions: increased number of trained random forests, selection of important variables with an iterative process, repetition of full selection process from different seeds. This approach's main contribution stems from its interpretability: through repetition, it also outputs a robust subset of discriminant variables. These selections were shown to be more consistent than those obtained by a state-of-the-art method, and their contents were validated through a manual bibliographic research. The interconnections between selected taxa and functional annotations were also analyzed and revealed that important annotations emerge from the cumulated influence of non-selected taxa.

Metabolism studies providing insights on evolutionary processes [*S. Blanquart, J. Got, P. Giraud, A. Siegel*] [14, 18]

- *Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems* Brown seaweeds are keystone species of coastal ecosystems, often forming extensive underwater forests, and are under considerable threat from climate change. In this study, analysis of multiple genomes has provided insights across the entire evolutionary history of this lineage, from initial emergence, through later diversification of the brown algal orders, down to microevolutionary events at the genus level [14].

- *Metabolic drift in plant and algal lipid biosynthesis pathways* Metabolic pathway drift has been formulated as a general principle to help in the interpretation of comparative analyses between biosynthesis pathways, such as done in former work with the pathmodel method developed by the team. Indeed, such analyses often indicate substantial differences, even in widespread pathways that are sometimes believed to be conserved. In this paper, we checked how much this interpretation fits to empirical data gathered in the field of plant and algal biosynthesis pathways. After examining several examples representative of the diversity of lipid biosynthesis pathways, we investigate the question of how much biotic interactions are responsible for shaping this metabolic plasticity. We end up introducing some model systems that may be promising for further exploration of this question [18].

Analysis of cross-feeding interactions [*J. Got, P. Giraud, Y. Le Cunff, A. Siegel*] [27, 24]

- In collaboration with Inria team Pleiade, we pushed forward the analysis of microbial interactions in different biological contexts: metagenome-scale metabolic modelling for the characterization of cross-feeding interactions in Microcystis-associated microbial communities, in the context of freshwater cyanobacterial blooms [27], and modeling the emergent metabolic potential of soil microbiomes in Atacama landscapes [24].

Evolutionary study of conserved modules in the Fibulin multi-domain protein family [*C. Belleannée, S. Blanquart, F. Coste, E. Chenel, N. Théret*] [28]

- Fibulins are an eight-member family of multidomain extracellular matrix proteins. Proteomic analysis by IRSET's DYMEC2 team has highlighted the association between three members of the fibulin family and the severity of fibrosis in chronic liver disease. We used and tested the tool FUSE-PhyloTree, developped during the thesis of O. Dennler, to study the fibulin family. Reconstruction of the evolutionary history of small conserved regions (module) and of the evolutionary history of annotations enabled the identification of four ancestral genes showing module/annotation co-appearance.

Amino acid conservation score based on sequence distances [*E. Chevé, F. Coste, P. Espana Gutierrez*] [29]

- Classical models of protein families, such as profile hidden Markov models (pHMM), assign a score to amino acids for each column of the alignment which tells how conserved they are in this column. Such score can be interpreted as expressing how important for the function each amino acids are at a given position. We proposed a new approach that takes distances between pairs of sequences into account for the computation of the conservation score of amino acids at each position.

8.3 Regulation and signaling: detecting complex and discriminant signatures of phenotypes

Participants: Emmanuelle Becker, Catherine Belleannée, Samuel Blanquart, Myriam Bontonou, Mathieu Bougueon, Olivier Dameron, Juliette Francis, Yann Le Cunff, Corentin Lucas, Noryah Safla, Anne Siegel, Nathalie Théret, Kérian Thuillier.

CEGAR-based approach for solving combinatorial optimization modulo quantified linear arithmetics problems [*A. Siegel, K. Thuillier*] [20, 23].

- The synthesis of multi-scale models of biological networks has recently been associated with the resolution of optimization problems mixing Boolean logic and universally quantified linear constraints (OPT+qLP), which can be benchmarked on real-world models. We introduce a CounterExample Guided Abstraction Refinement (CEGAR) to solve such problems efficiently. Our CEGAR exploits monotone properties inherent to linear optimization in order to generalize counterexamples of Boolean relaxations. We implemented our approach by extending Answer Set Programming (ASP) solver Clingo with a quantified linear constraints propagator. Our prototype enables exploiting independence of sub-formulas to further exploit the generalization of counterexamples. We evaluate the impact of refinement and partitioning on two sets of OPT+qLP problems inspired by system biology. Additionally, we conducted a comparison with the state-of-the-art ASP solver Clingo[lpx] that handles non-quantified linear constraints, showing the advantage of our CEGAR approach for solving large problems.

A rule-based multiscale model of hepatic stellate cell plasticity: Critical role of the inactivation loop in fibrosis progression [*M. Bougueon, A. Siegel, N. Théret*] [13].

- Chronic liver diseases are associated with the development of fibrosis which is characterized by an abnormal deposition of extracellular matrix (ECM) leading to severe liver dysfunction. Hepatic stellate cells (HSCs) are key players in liver fibrosis driving ECM remodeling. However numerous biological processes are involved including HSC activation, proliferation, differentiation and inactivation and novel computational modeling is necessary to integrate such complex dynamics. We used the Kappa graph rewriting language to develop the first rule-based model describing the HSCs dynamics during liver fibrosis and its reversion. Simulation analyses enabled us to demonstrate the critical role of the HSC inactivation loop in the development of liver fibrosis, and to identify inactivated HSCs as potential new markers of fibrosis progression.

Regulus infers signed regulatory relations from few samples' information using discretization and likelihood constraints [*O. Dameron, A. Siegel*] [15]

- Transcriptional regulation is performed by transcription factors (TF) binding to DNA in context-dependent regulatory regions and determines the activation or inhibition of gene expression. Current methods of transcriptional regulatory circuits inference, based on one or all of TF, regions and genes activity measurements require a large number of samples for ranking the candidate TF-gene regulation relations and rarely predict whether they are activations or inhibitions. We proposed the Regulus method allowing transcriptional regulatory circuits can be inferred from fewer samples by (1) fully integrating information on TF binding, gene expression and regulatory regions accessibility, (2) reducing data complexity and (3) using biology-based likelihood constraints to determine the global consistency between a candidate TF-gene relation and patterns of genes expressions and region activations, as well as qualify regulations as activations or inhibitions. Regulus provides signed relations consistent with public databases and, when applied to biological data, identifies both known and potential new regulators.

9 Bilateral contracts and grants with industry

9.1 Bilateral Grants with Industry

BeCycle

Participants: Jeanne Got, Noé Robert, Anne Siegel.

In the context of the Grand Défi "Ferment du futur", this private-public project aims at scanning thousands of bacterial genomes to identify the best consortium of strains capable of producing metabolites of interest. Duration: 2024-2026, total of the grant 400k€.

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

SymBioDiversity

Title: Symbolic and numerical mining and exploration of functional biodiversity

Duration: 2020 -> 2024

Coordinator: Alejandro Maass (University of Chile), Clémence Frioux (Pleaide Team), Anne Siegel (Dyliss Team)

Partners:

- Universidad de Chile (Chili)

Inria contact: Anne Siegel

Summary: The project aims at developing methods combining data-mining, reasoning and mathematical modeling to efficiently analyze massive data about microbial biodiversity in extreme environment and identify families of species which characterize environmental niches. The partnership combines Inria Team Dyliss (systems biology, reasoning), Pléiade (systems biology, biodiversity), the chilean Center of Mathematical Modeling (modeling of ecosystems), Inria Chile (data mining, transfer) and chilean biologist partners experts in biodiversity (universidad catholica).

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

- Alejandro Maass, senior researcher, University of Chile, October 2024.
- Sebastian Mendoza, junior researcher, University of Chile, October 2024.

10.2.2 Visits to international teams

Research stays abroad

- Anne Siegel visited University of Chile in January 2024 for one week.
- Moana Aulagner visited University of Chile in November and December 2024.

10.3 European initiatives

10.3.1 Other european programs/initiatives

ERC HoloE2Plant, Exploring the Holobiont concept through a Plant Evolutionary Experiment study

Participants: Moana Aulagner, Samuel Blanquart, Anne Siegel.

Exploring the Holobiont concept through a Plant Experimental Evolution study. In her ERC project, Claudia Bartoli aims at validating the holobiont concept, highlighting how the interactions with its microbiota influence a species evolution. The study will apply to a host/pathogen system, Brassica rapa / Rhizoctonia solani, associated with bacterial and fungal synthetic communities. Examining nine plant generations in an experimental-evolution apparatus should reveal the molecular outcomes of the applied selective pressures. 2022-2027, total of the grant 1500k€.

10.4 National initiatives

PEPR Digital agro-ecology : HOLOBIONT

Participants: Juliette Francis, Yann Le Cunff.

Animals and their microbiota form a composite organism, called a holobiont, which can be considered the ultimate unit on which evolution and selection act. Host genes and the environment influence the colonization, development, and function of the various microbiota, which in turn help shape the host's phenotypes. The phenotypes of the holobiont thus result from the combined action of the host genes and those of its microbiota, and their determinism can be explored by implementing hologenetic approaches capable of considering host genomes and metagenomes jointly. The overall objective of this PEPR is to develop integrative hologenetic approaches for animal breeding, using state-of-the-art technologies to generate, process and analyze genetic and genomic datasets of the host and its microbiota as well as the phenotypes and environmental parameters in which the holobionts evolve. To this end, the project aims to develop methods for the analysis of new-generation phenotyping data of the holobiont (mainly high-throughput and continuous), for their modeling and for the analysis of their interrelationships with the microbiota data. Juliette Francis's Ph.D, co-supervised by Yann Le Cunff (Dyliss) and Mahendra Mariadassou (INRAe, MaIAGe), focuses on co-analyzing genomic data, microbiota data and metabolomic data to efficiently predict a phenotype of interest (food intake efficiency in this case). 2024–2028. Dyliss grant : 178k€.

PEPR digital health : M4DI

Participants: Emmanuelle Becker, Océane Carpentier, Yann Le Cunff.

The main objective of the Methods and Models for Multimodal and Multiscale Data Integration (M4DI) project is to develop innovative methodological frameworks for the integration of biomedical datasets. In particular, the team is involved in designing robust machine learning approaches enhanced by prior knowledge. In particular, Océane Carpentier's Ph.D, co-supervised by Emmanuelle Becker, Yann Le Cunff (Dyliss), Nicolas Jay and Aurélie Bannay (LORIA, Nancy) is dedicated to exploiting the ontology structure of the Gene Ontology database in machine learning algorithms. One key application will be carried out on a local cohort of Crohn's patients with the CHU of Rennes. 2024–2028. DYLISS grant: 169k€.

PEPR Digital health : ShareFAIR

Participants: Olivier Dameron, Ulysse Le Clanche, Yann Le Cunff.

The increasing availability of life science data offers unprecedented opportunities for healthcare research, it has the potential to revolutionize the way we understand and treat diseases, as it allows researchers to identify trends and patterns that may not have been apparent with smaller data sets. However, exploiting this potential requires innovative solutions for the annotation of biomedical and clinical datasets and extraction of provenance. Challenges thus include standardization and annotation for datasets and protocols, extracting protocols from text and datasets, and synthesizing them into interoperable, yet shareable protocols. ShareFAIR will provide (i) standards to uniformly annotate datasets and protocols with ontologies/common vocabularies and provenance to trace their origin, (ii) an interoperable framework to index, design and annotate reliable and shareable analysis protocols, (iii) approaches to extract new protocols, based on the literature, learned from biomedical and clinical datasets, and from international data challenges in neuroimaging. Dyliss contribution consists in designing a semi-automated dataset FAIRification method that will extend low-level metadata by higher level descriptions inferred from the workflow specification and execution. These descriptions will provide a summary focusing on the “what" rather than the “how", that will be instrumental to workflow recommendation as well as improved reusability of data analysis results. To this end, we will leverage domain-specific knowledge associated to biomedical datasets, as well as fine-grained workflow execution provenance traces so that data analysis results can be more easily understood, explained and shared, in line with critical open and reproducible sciences initiatives. The PhD of Ulysse Le Clanche is co-supervized by Olivier Dameron at Dyliss and Alban Gaignard at Institut du Thorax, INSERM and Univ. Nantes. 2023–2026. Dyliss grant: 185k€.

DeepImpact : Deciphering plant-microbiome interactions to enhance crop defense to bioagressors

Participants: Samuel Blanquart, Olivier Dameron, Jeanne Got, Victor Mataigne, Pauline Giraud, Anne Siegel.

DEEP IMPACT is a multidisciplinary consortium-based project that aims at combining ecology, biology, plant genetics and mathematics to identify, characterize and validate the microbial communities, plant communities and abiotic factors (including agricultural managements) explaining variation in *Brassica napus* and *Triticum aestivum* resistance to several pests. For this, we will start from an *in situ* approach by characterizing 100 fields (50 for each crop species) for both habitat (climatic and edaphic variables) and biotic (microbiota, virome, weed communities, pest attacks and pathobiota prevalence) features. Information from this broad characterization will be integrated into sparse and correlative statistical models to describe the relative part of the variance explained by both habitat and biotic features and correlated with a reduction of pest’s attacks. This analysis will allow us to identify a combination of microbial species and soils, correlated with an increase of crop’s resistance to pests. These microbial consortia will be isolated by taking advantages of newly developed culturomics methods and characterized by both whole genome sequencing and biochemical assays. Synthetic Consortia (SynComs) will be reconstructed to test their efficacy on a broad range of pests attacking both crops. 2021–2026. Dyliss grant: 176k€.

SEABIOZ : Potential microbial origins of the biostimulant properties of extracts from a brown algae holobinte

Participants: Samuel Blanquart, Olivier Dameron, Jeanne Got, Anne Siegel.

For sustainable agriculture, new bio-based solutions include biocontrol and the use of plant biostimulants such as aqueous seaweed extracts. The most widely exploited biomass for biostimulant production is the brown seaweed *Ascophyllum nodosum* and its commercial extracts, including products from the Roullier Group, have demonstrated their ability to improve plant growth and mitigate certain abiotic and biotic stresses. A unique feature of the alga is its mutualistic association with the fungal endophyte *Mycophycias ascophylli* and other microbes constituting an holobiont. Many questions remain as to the nature and origin of the active compounds in algal extracts. Are these bioactive metabolites produced by the host or by its microbiota? The main objective of SEABIOZ is to answer these questions by combining a multi-omics approach and systems biology. 2021–2024. Dyliss grant: 120k€.

ENDOVIRE (ANR)

Participants: Emmanuelle Becker, Olivier Dameron, Yael Tirlet.

The whole ANR project gathers 4 partners : the BIPAA platform (INRAe), the DGIMI laboratory, the BF2I laboratory and the Dyliss team of IRISA. The project is focused about the understanding of how genes of a endogeneized viral genome in a parasitoid wasp are the activated and regulated. The available data produced by the consortium will cover genomics, epigenomics, pathways, regulation and orthology. We will contribute to identify the key actors involved in the activation of parasitoids genes, to propose a data and knowledge integration framework for the data of the global project, and to develop integrative data analysis methods for elucidating the mechanism involving the key actors identified in the first point. It will consist in proposing a library of queries (which contains a reasoning part), and further to propose regulation mechanisms based on heterogeneous -omics data across interacting organisms. To tackle the different challenges, our appoach will be based on (1) adequate statistical analysis workflows or methods, (2) Semantic Web technologies and AskOmics developed within the team, (3) knowledge-guided traversal strategies across multiplex graphs. 2023–2025. Total grant: 630k€. Dyliss grant: 176k€.

10.4.1 Programs funded by Inria

Défi Inria OmicFinder

Participants: Olivier Dameron.

Coordinator: Pierre Peterlongo

Duration: 48 months (May 2023 - May 2027)

Partners: Inria teams: Dyliss, Zenith, Taran.

External partners are CEA-GenoScope, Elixir, Pasteur Institute, Inria Challenge OceanIA, CEA-CNRGH, and Mediterranean Institute of Oceanography.

Description: The project aims to provide a novel global search engine making it possible to query nucleotidic sequences against the vast amount of publicly available genomic data. The central algorithmic idea of a genomic search engine is to index and query small exact words (hundreds of billions over millions of datasets), as well as the associated metadata.

website: project.inria.fr/omicfinder

11 Dissemination

Participants: Emmanuelle Becker, Catherine Belleannée, Cécile Beust, Samuel Blanquart, Myriam Bontonou, Océane Carpentier, François Coste, Olivier Dameron, Pablo Espana Gutierrez, Juliette Francis, Jeanne Got, Yann Le Cunff, Corentin Lucas, Victor Mataigne, François Moreews, Anne Siegel, Yael Tirlet, Nathalie Théret.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- Colloque "jumeaux numériques", January 2024, Paris, CNRS [16] [Anne Siegel]

Member of the organizing committees

- WemSemPilot [17] workshop (02–03 July 2024, Rennes), a national workshop and hackathon on interoperability and Semantic Web technology and their application to agriculture and life sciences. [François Moreews]
- Meeting algometabionte, Roscoff, october 2024, 30 participants [Anne Siegel]

11.1.2 Scientific events: selection

Member of the conference program committees

- SWAT4HCLS 2024, Leiden Netherlands (international conference Semantic Web Applications and Tools for Health Care and Life Science) [Olivier Dameron]
- European Conference on Computational Biology (ECCB 2024, Turku Finland) [Olivier Dameron]
- Jobim 2024 (Journées Ouvertes Biologie Informatique Mathématiques), France [Olivier Dameron; Anne Siegel]
- Journée Santé et IA (plateforme IA, La Rochelle) [Olivier Dameron]
- ISMB-2024 (International Symposium on Molecular Biology) [Anne Siegel]

Reviewer

- SWAT4HCLS 2024 [Olivier Dameron]
- ECCB 2024 [Olivier Dameron]
- Jobim 2024 [Olivier Dameron]
- Santé et IA [Olivier Dameron]
- ISMB 2024 [Yann Le Cunff]

11.1.3 Journal

Member of the editorial boards

- Journal of Biomedical Semantics [Olivier Dameron]

[16] https://miti.cnrs.fr/evenement-scientifique/colloque-jumeaux-numeriques/
[17] https://pegase.pages.mia.inra.fr/wspilot/index_en.html

Reviewer - reviewing activities

- PLOS Computational Biology [Samuel Blanquart]
- Bioinformatics advances [Yann Le Cunff]
- ISME Journal [Anne Siegel]
- Matrix Biology [Nathalie Théret]

11.1.4 Invited talks

- Journées Algométabionte, 3rd March 2024, Roscoff. *Some perspectives on the enzymatic annotation of the dark proteome using deep protein language models* [François Coste]
- Journées CNRS MERIT: Usage de l'IA pour l'annotation fonctionnelle et structurale des génomes, 7th Nov 2024, Paris. *Enzymatic annotation of protein sequences with a deep language model* [François Coste]
- École d'été interdisciplinaire en numérique de la santé (EINS 2024), 03–07th June 2024 Univ. Sherbrooke Canada. *Intégration et interrogation avancées de données et de connaissances grâce au Web Sémantique* [Olivier Dameron]
- École thématique sur la Bioinformatique Intégrative, Institut Français de Bioinformatique (ETBII, 25th–29th March 2024, Fréjus) : animateur du thème *Web Sémantique* [Olivier Dameron]
- Congreso futuro 2024, organized by the Chilean congress at santiago de Chile, january 2024 *Symbiosis and metabolism* [Anne Siegel]
- Metabolic club workshop, university of chile, january 2024 *Study and Comparisons of metabolic maps : adapt to the triptych of knowledge – data – scaling up* [Anne Siegel]
- 5 minutes du centre Henri Lebesgue*Symbiose, biologie des systèmes, et discrétisation de systèmes dynamiques*, février 2024 [Anne Siegel]

11.1.5 Leadership within the scientific community

National responsibilities

- Deputy Scientific Directory (CNRS, INS2I), in charge of interdisciplinarity between numerical sciences and other disciplines, gender equality in computer sciences, groupements de recherches (GDR) [Anne Siegel]
- Mediator and Member of the steering committee of the programme LORIER: The Organization for Ethical and Responsible Research at Inserm (https://lorier.inserm.fr/en/) [Nathalie Théret]

Local responsibilities

- Head of the Master degree "Bioinformatics" [Emmanuelle Becker]
- Scientific Advisory Board of the BioGenOuest network (37 platforms) [Emmanuelle Becker]
- Reponsible of the 2nd and 3rd years of the "Yes-si" ISTN licence program, Univ. Rennes, France [Catherine Belleannée]
- In charge of the "Open Day and student fair" for Istic, Univ. Rennes, France [Catherine Belleannée]
- Referent teacher, 15h, L1 informatique, Univ. Rennes, France [Catherine Belleannée]
- Organisation of the bioinformatics teams (Dyliss, GenOuest and GenScale as well as members of other bioinformatics teams in Rennes) weekly seminars [Samuel Blanquart]

- Chargé de mission "Numérique et Environnement" for Inria centre at Rennes University [Samuel Blanquart]
- Chargé de mission "Biologie et Santé Numériques" for Inria centre at Rennes University [François Coste]
- Scientific Advisory Board of the GenOuest platform [Olivier Dameron]
- Scientific Director of the GenOuest platform [Yann Le Cunff]
- Responsability of the IRISA laboratory "Health-biology" cross-cutting axis [Yann Le Cunff]
- Head of the double-diploma Licence degree "Life Science, Maths and Artificial Intelligence" [Yann Le Cunff]
- Member of the Parcoursup jury for the Life Science Licence, Univ. Rennes, France [Yann Le Cunff]

11.1.6 Scientific expertise

Evaluation of international projects

- COST action 2024 [Anne Siegel]

Evaluation of national projects

- ELIXIR-FR [Olivier Dameron]

11.1.7 Research administration

Institutional boards for the recruitment and evaluation of researchers

- Junior professor selection committee University of Strasbourg [Emmanuelle Becker]
- Junior professor selection committee University of Poitiers [Emmanuelle Becker]
- Associate professor selection committee University of Marseille [Emmanuelle Becker]
- Associate professor selection committee University of Brest [Emmanuelle Becker]
- Research Engineer selection committee INRAe IGEPP [Emmanuelle Becker]
- Junior professor selection committee Istic, University of Rennes [Catherine Belleannée]
- Junior professor in Biology selection committee, CNRS [Anne Siegel]
- Junior professor in Compuer Science selection committee, CNRS [Anne Siegel]

Scientific councils

- Scientific referent (for CNRS) of the PEPR exploratoire Molecularxiv [Anne Siegel]
- Comité de pilotage of the Mission for Interdisciplinarity (MITI) at CNRS [Anne Siegel]
- Scientific advisory Board of the LPHI lab [Anne Siegel]

Local responsibilities

- Member of the social committee of Univ. Rennes [Catherine Belleannée]
- Member of the emergency aid commission of Univ. Rennes and Rennes 2 [Catherine Belleannée]
- Member of CUMI (Commission des utilisateurs des moyens informatiques) of Inria Rennes [François Coste]
- Member of the Inria Rennes center council [Jeanne Got]
- Member of the Biology department council [Yann Le Cunff]

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

- Master : Emmanuelle Becker, "R, Data, and Visualisation", 61h, Master 1 in Bioinformatics, Master 1 in Ecology and Environment, Univ. Rennes, France
- Master : Emmanuelle Becker, "Object oriented programming", 39h, Master in Bioinformatics, Univ. Rennes, France
- Master : Emmanuelle Becker, "Method", 15h, Master 2 in Computer Sciences, Univ. Rennes, France
- Master : Emmanuelle Becker, "Biological networks", 31h, Master 2 in Bioinformatics, Univ. Rennes, France
- Master : Emmanuelle Becker, "Introduction to Bioinformatics", 3h, Master MEEF Biology, Univ. Rennes, France
- Master : Emmanuelle Becker, "Manipulate and Visualize Data", 27h, Bioinformatics Minor for Master Students, Univ. Rennes, France
- Licence : Emmanuelle Becker, member of the Parcoursup jury for the Life Science Licence, Univ. Rennes, France
- Licence: Catherine Belleannée, "Formal Languages", 20h, L3 informatique, Univ. Rennes, France
- Licence: Catherine Belleannée, "Projet professionnel et communication", 16h, L1 informatique, Univ. Rennes, France
- Licence: Catherine Belleannée, "Projet professionnel et communication", 12h, L2 informatique, Univ. Rennes, France
- Licence: Catherine Belleannée, Spécialité informatique, "Functional and immutable programming", 44h, L1 mathématiques, Univ. Rennes, France
- Master: Catherine Belleannée, "Answer Set Programming", 15h, M1 informatique, Univ. Rennes, France
- Master: Catherine Belleannée, "Programmation logique et contraintes", 32h, M1 informatique, Univ. Rennes, France
- Licence: Catherine Belleannée, "Outils formels pour l'informatique", 46h, L2 informatique, Univ. Rennes, France
- Licence: Catherine Belleannée, "Fondements mathématiques", 49h, L1 informatique, Univ. Rennes, France
- Licence : Cécile Beust, "Informatique", 16h, Licence 1 PCSTM, Univ. Rennes, France
- Licence : Cécile Beust, "Data : Sciences des données", 24h, Licence 2 ISTN, ISTIC, France
- Master : Cécile Beust, "Apprentissage Statistique", 27h, Master 1 Molecular and Cellular Biology, Univ. Rennes, France
- Master : Cécile Beust, "Data Engineering in Life Sciences", 5h, Master 2 Bioinformatics, Univ. Rennes, France
- Master : Myriam Bontonou, "Optimisation", 10.5h, Master 1 Info (IA), ISTIC, Univ. Rennes, France
- Master : Myriam Bontonou, "Modelisation and Discrete Optimisation", 30h, Master 1 MIAGE, ISTIC, Univ. Rennes, France
- Licence : Myriam Bontonou, "Graph Modeling and Algorithms", 21h, Licence 3 Info and MIAGE, ISTIC, Univ. Rennes, France

- Licence : Myriam Bontonou, "Introduction to AI", 27h, Licence 1, SVE, Univ. Rennes, France
- Licence : Myriam Bontonou, "Databases", 6h, Licence 3 MIAGE, ISTIC, Univ. Rennes, France
- Master : Océane Carpentier, "Data Engineering in Life Sciences", 5h, Master 2 Bioinformatics, Univ. Rennes, France
- Licence: Olivier Dameron, "Programmation 1", 98h, Licence 1 informatique, Univ. Rennes, France
- Licence: Olivier Dameron, "Complément informatique", 24h, Licence 1 informatique, Univ. Rennes, France
- Licence: Olivier Dameron, "Introduction à l'IA", 3h, Licence 1 sciences de la vie et de l'environnement, Univ. Rennes, France
- Licence: Olivier Dameron, "Data analysis and statistics", 24h, Licence 2 informatique, Univ. Rennes, France
- Licence: Olivier Dameron, "Graph Modeling and Algorithms", 21h, Licence 2 informatique, Univ. Rennes, France
- Licence: Olivier Dameron, "Programmation avancée", 36h, Licence 3 miage, Univ. Rennes, France
- Master: Olivier Dameron, "Semantic Web", 10h, Master 1 miage, Univ. Rennes, France
- Master: Olivier Dameron, "Data Engineering in Life Science", 20h, Master 2 in bioinformatics, Univ. Rennes, France
- Master: Olivier Dameron, "Internship", 28h, Master 2 in bioinformatics, Univ. Rennes, France
- Licence: Pablo Espana Gutierrez, "Langages Formels et Calculabilité", 20h, L3SIF, ENS Rennes, France
- Licence: Pablo Espana Gutierrez, "Remise à niveau MPI", 20h, L3SIF, ENS Rennes, France
- Licence: Pablo Espana Gutierrez, "Préparation à l'agrégation", 8h, ENS Rennes, France
- Master : Juliette Francis, "Apprentissage Statistique", 30h, Master 1 in Bioinfortmatics, Univ. Rennes, France
- Licence : Yann Le Cunff "Modélisation des phénomènes du vivant", 30h, L2 Biologie, Univ. Rennes, France
- Master: Yann Le Cunff, "Apprentissage statistique", 110h, Master 1 in Bioinfortmatics Univ. Rennes, France
- Master: Yann Le Cunff, "Biologie aux interfaces", 25h, Master 1 in Biology, Univ. Rennes, France
- Master: Yann Le Cunff,"Simulating dynamic systems in biology", 20h, Master 2 in bioinformatics, Univ. Rennes, France
- Master: Yann Le Cunff, "Applied Interdisciplinarity", 20h, Master 2 in biology, Univ. Rennes, France
- PhD program: Yann Le Cunff, "Introduction to Machine Learning", 20h, FdV PhD Program, Sorbonne Paris Université, Paris, France
- Master : Corentin Lucas, "Apprentissage Statistique", 27h, Master 1 Molecular and Cellular Biology, Univ. Rennes, France
- Master : Corentin Lucas, "Data Engineering in Life Sciences (DEL)", 5h, Master 2 Bioinformatics, Univ. Rennes, France
- Master : Victor Mataigne, "Programming in R", 18h, Master 1 in Natural Resources and Biodiversity, Univ. Rennes, France

- Master : Victor Mataigne, "Introduction to Bioinformatics", 8.5h, Master in Biology and Agrosciences, Institut Agro Rennes Angers, France
- Master : Yael Tirlet, "Short Introduction to R (SIR) / Programming with R (PAR)", 18h, Master 1 in Bioinformatics, Master 1 in Ecology and Environment, Univ. Rennes, France
- Master : Yael Tirlet, "Data Engineering in Life Sciences (DEL)", 5h, Master 2 Bioinformatics, Univ. Rennes, France
- Licence : Yael Tirlet, "Data : Science des Données (DSD)", 24h, Licence 2 Informatique, Istic, Rennes, France
- Licence : Yael Tirlet, "Projet Personnel et Professionnel de l'étudiant (3PE)", 9h, Licence 1 informatique, Istic, Rennes, France

11.2.2 Supervision

PhD thesis

- PhD in progress: Moana Aulagner, Modeling microbiota interactions in plants to build synthetic microbial communities for enhanced biocontrol and biostimulation, started in Oct 2023, supervized by Samuel Blanquart, Anne Siegel and C. Bartoli-Kautski (INRAe IGEPP)
- PhD in progress: Moussa Baddour, Extraction de phénotypes à partir de comptes-rendus médicaux textuels et mise en relation avec le génotype, started in May 2023, supervized by Olivier Dameron, M. De Tayrac (Rennes Hospital), S. Paquelet (b<>com) and T. Labbé (Orange)
- PhD in progress: Cécile Beust, Knowledge-guided rules for generating context-specific views on a knowledge graph: application to biological networks, started in Oct 2023, supervized by Emmanuelle Becker, Olivier Dameron and Nathalie Théret
- PhD in progress: Océane Carpentier, Integrating prior knowledge for a better patient representation, started September 2024, supervized by Emmanuelle Becker, Yann Le Cunff, A. Bannay and N. Jay (LORIA)
- PhD in progress: Elisa Chenel, Study of protein co-evolution to identify interaction regions involved in TGFbeta growth factor activation, started in Oct 2024, supervized by Samuel Blanquart, François Coste and N. Nathalie Théret
- PhD in progress: Pablo Espana Gutierrez, Learning models with explicit dependencies between residues to predict protein functions, started in September 2023, supervized by François Coste and Olivier Dameron
- PhD in progress: Juliette Francis, Intégration de données hétérogènes pour la prédiction de phénotype, started in October 2024, supervized by Yann Le Cunff and M. Mariadasssou (INRAe)
- PhD in progress: Pauline Giraud, Méthodes hybrides pour l'inférence ab-initio de voies métaboliques chez des eucaryotes marins, started in November 2024, supervized by Anne Siegel and G. Markov (CNRS, Station biologique de Roscoff)
- PhD in progress: Ulysse Le Clanche, Knowledge-driven dataset FAIRification: from workflow runs to domain-specific annotations, started in October 2024, supervized by Olivier Dameron and A. Gaignard (CNRS, Institut du Thorax INSERM Nantes)
- PhD in progress: Corentin Lucas, Integration of multi-modal data for longitudinal follow-up of Crohn's disease patients, started in Oct 2023, supervized by Emmanuelle Becker, Yann Le Cunff
- PhD: Baptiste Ruiz, Algorithmes d'apprentissage automatique appliqués au microbiote Intégration de connaissances a priori pour de meilleures prédictions de phénotype, started in Oct 2021, defended in Nov 2024, supervized by Yann Le Cunff, Anne Siegel

- PhD: Kerian Thuillier, Inférence de règles booléennes contrôlant des modèles hybrides de systèmes biologiques multi-échelles, started in Oct 2021, defended in Sep 2024, supervized by Anne Siegel and L. Paulevé (LABRI)
- PhD in progress: Yael Tirlet, Integrative method for multi-omics data analysis with application to the activation and regulation of an endogeneized viral genome in a parasitoid wasp, started in Oct 2023, supervized by Emmanuelle Becker, Olivier Dameron and F. Legeai (INRAe)

Internship

- M2 intership: Elisa Chenel, Recherche de motifs fonctionnels dans une famille de protéines multi-domaines : les Fibulines. Jan-Jul 2024, supervized by Samuel Blanquart and Nathalie Théret.
- L3 internship: Eoghan Chevé, Distance-based amino acid conservation score, Jun-Jul 2024, supervized by Pablo Espana Gutierrez and François Coste
- M2 internship : Juliette Francis, Intégration de données hétérogènes pour la prédiction de phénotype, Jan-Jul 2024, supervized by Yann Le Cunff and M. Mariadasssou (INRAe)
- M2 intership: Hugo Mingarelli, Enrichir MetaCyc avec des connaissances sur les molécules pour l'analyse de réseaux de réactions chimiques. Jan-Jul 2024, supervized by Olivier Dameron et Jeanne Got.
- M1 Internship : Noryah Safla, Prédiction de phénotype de croissance d'algue à partir de données de microbiote, Apr-May 2024, supervized by Yann Le Cunff and S.Dittami (Roscoff)

11.2.3 Doctoral advisory committees (CSID)

- Maria-Mafalda Almeida, Univ. Rennes [Emmanuelle Becker]
- Juan Andrés Cisneros–Jacome, Univ. Rennes [Emmanuelle Becker]
- Maëlys Auffret, Univ. Rennes 2 [Emmanuelle Becker]
- Yvon Awuklu, Univ. Bordeaux [Olivier Dameron]
- Clément Bousquet, Univ. de Rennes [François Coste]
- Simon Brocard, Nantes Université [François Coste]
- Andrea Checcoli, Institut Curie [Anne Siegel]
- Dorian Chenet, Univ. Rennes [Samuel Blanquart]
- Guénolé Dande, Univ. de Rennes [Olivier Dameron]
- Guillaume Doré, Univ. Rennes [Emmanuelle Becker]
- Jin-Mei Gao, Université Paris-Saclay [Emmanuelle Becker]
- Silvia Grosso, INSA Lyon [Yann Le Cunff]
- Mats Kohler–Dijkstra, Univ. Rennes [Emmanuelle Becker]
- Gabriel Mastrilli, Univ. de Rennes [François Coste]
- Meije Mathé, Univ. Toulouse [Olivier Dameron]
- Thibaut Peyric, Univ. Lyon [Yann Le Cunff]
- Quentin Vacher, Univ. Rennes [Emmanuelle Becker]
- Maelle Zonnequin, Sorbonne Université [Anne Siegel]

11.2.4 Juries

Referee of PhD thesis

- Elsa Claude, Université Laval and Université de Bordeaux [Emmanuelle Becker, president, François Coste]
- Louise Dupuis, Sorbonne University [Emmanuelle Becker]
- Danilo Dursoniah, Lille university [Anne Siegel]
- Romane Junker, Paris-Saclay University [Emmanuelle Becker]
- Hasna Maayouf, Université de Haute Alsace, Université de Strasbourg [Nathalie Théret]
- Alix Simon, Strasbourg University [Emmanuelle Becker]
- Kerian Thuillier, Univ. Rennes [Emmanuelle Becker, president]

Member of PhD thesis juries

- Morgane Lallier, Université de Nantes [Nathalie Théret, president]
- Raissa Silva, Université de Montpellier [Yann Le Cunff]

Member of habilitation thesis juries

- Carito Guziolowski, Univ. Nantes [Anne Siegel, president]
- Emeline Roux, Univ. Rennes [Emmanuelle Becker, president]

11.3 Popularization

11.3.1 Participation in Live events

- Invited talk at Conférence Culture Générale Bio-info, Masters 1 et 2 Bioinformatique, 15th Nov 2024, Rennes. *Linguistic modelling of protein sequences* [François Coste]
- Intervention in the research initiation workshop Réunion des Jeunes Mathématiciennes et Informaticiennes (RJMI), 12 Jan - 14 Jan 2024 in Bruz (Rennes). [Pablo Espana Gutierrez]
- Invited talk at the graduation ceremony of the computer science department of ENS Rennes, 17th May 2024 in Bruz (Rennes). *Learning models with explicit dependencies between residues to predict protein functions* [Pablo Espana Gutierrez]
- Intervention and organization of the computer science research initiation week CAMP FACTO, 21 Oct - 25 Oct 2024 in Saint-Chamond [Pablo Espana Gutierrez]
- Invited talk by JeBIF association for Masters 1 and 2 Bioinformatique, 29th Nov 2024 in Rennes. *careers in bioinformatics* [Jeanne Got]
- Portraits de femmes engagées, ministère éducation nationale mars 2024 [Anne Siegel]
- Teaching computer science and the Python language to middle-school girls through creative programming, Action LCLC (elles codent, elles créent), May and Jun 2024, Collège le Landry, Rennes, France [Yael Tirlet]

12 Scientific production

12.1 Major publications

[1] 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 (May 2018). e1006146. DOI: 10.1371/journal.pcbi.1006146. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01807842 (cit. on p. 7).

[2] C. Belleannée, O. Sallou and J. Nicolas. 'Logol: Expressive Pattern Matching in sequences. Application to Ribosomal Frameshift Modeling'. In: *PRIB2014 - Pattern Recognition in Bioinformatics, 9th IAPR International Conference*. Ed. by M. Comin, L. Kall, E. Marchiori, A. Ngom and J. Rajapakse. Vol. 8626. Lukas KALL. Stockholm, Sweden: Springer International Publishing, Aug. 2014, pp. 34–47. DOI: 10.1007/978-3-319-09192-1_4. URL: https://hal.inria.fr/hal-01059506 (cit. on p. 8).

[3] C. Bettembourg, C. Diot and O. Dameron. 'Optimal Threshold Determination for Interpreting Semantic Similarity and Particularity: Application to the Comparison of Gene Sets and Metabolic Pathways Using GO and ChEBI'. In: *PLoS ONE* (2015), p. 30. DOI: 10.1371/journal.pone.0133579. URL: https://hal.inria.fr/hal-01184934.

[4] P. Bordron, M. Latorre, M.-P. Cortés, M. Gonzales, S. Thiele, A. Siegel, A. Maass and D. Eveillard. 'Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach'. In: *MicrobiologyOpen* 5.1 (2015), pp. 106–117. DOI: 10.1002/mbo3.315. URL: https://hal.inria.fr/hal-01246173 (cit. on p. 7).

[5] J. Coquet, N. Théret, V. Legagneux and O. Dameron. 'Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-β Signaling'. In: *CMSB 2017 - 15th International Conference on Computational Methods in Systems Biology*. Lecture Notes in Computer Sciences. Darmstadt, Sept. 2017, p. 17. URL: https://hal.archives-ouvertes.fr/hal-01559249 (cit. on p. 8).

[6] F. Coste, G. Garet, A. Groisillier, J. Nicolas and T. Tonon. 'Automated Enzyme classification by Formal Concept Analysis'. In: *ICFCA - 12th International Conference on Formal Concept Analysis*. Cluj-Napoca, Romania: Springer, June 2014. URL: https://hal.inria.fr/hal-01063727 (cit. on p. 7).

[7] F. Coste and J. Nicolas. 'Learning local substitutable context-free languages from positive examples in polynomial time and data by reduction'. In: *ICGI 2018 - 14th International Conference on Grammatical Inference*. Vol. 93. Wrocław, Poland, Sept. 2018, pp. 155–168. URL: https://hal.inria.fr/hal-01872266.

[8] 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 (Sept. 2018), pp. i934–i943. DOI: 10.1093/bioinformatics/bty588. URL: https://hal.inria.fr/hal-01871600.

[9] C. Frioux, T. Schaub, S. Schellhorn, A. Siegel and P. Wanko. 'Hybrid Metitebolic Network Completion'. In: *Theory and Practice of Logic Programming* (Nov. 2018), pp. 1–23. URL: https://hal.inria.fr/hal-01936778 (cit. on p. 6).

[10] S. Prigent, C. Frioux, S. M. Dittami, S. Thiele, A. Larhlimi, G. Collet, G. Fabien, 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 (Jan. 2017), p. 32. DOI: 10.1371/journal.pcbi.1005276. URL: https://hal.inria.fr/hal-01449100 (cit. on p. 6).

[11] S. Videla, J. Saez-Rodriguez, C. Guziolowski and A. Siegel. 'caspo: a toolbox for automated reasoning on the response of logical signaling networks families'. In: *Bioinformatics* (2017). DOI: 10.1093/bioinformatics/btw738. URL: https://hal.inria.fr/hal-01426880 (cit. on p. 8).

12.2 Publications of the year

International journals

[12] C. Beust, E. Becker, N. Théret and O. Dameron. 'BioPAX in 2024: Where we are and where we are heading'. In: *Computational and Structural Biotechnology Journal* 23 (4th Nov. 2024), pp. 3999–4010. DOI: 10.1016/j.csbj.2024.10.045. URL: https://inria.hal.science/hal-04801710 (cit. on p. 15).

[13] M. Bouguéon, V. Legagneux, O. Hazard, J. Bomo, A. Siegel, J. Feret and N. Théret. 'A rule-based multiscale model of hepatic stellate cell plasticity: Critical role of the inactivation loop in fibrosis progression'. In: *PLoS Computational Biology* 20.7 (2024), e1011858. DOI: 10.1371/journal.pcbi.1011858. URL: https://hal.science/hal-04689199 (cit. on p. 18).

[14] F. Denoeud, O. Godfroy, C. Cruaud, S. Heesch, Z. Nehr, N. Tadrent, A. Couloux, L. Brillet-Guéguen, L. Delage, D. Mckeown et al. 'Evolutionary genomics of the emergence of brown algae as key components of coastal ecosystems'. In: *Cell* (2024). DOI: 10.1016/j.cell.2024.10.049. URL: https://hal.science/hal-04794231. In press (cit. on p. 17).

[15] M. Louarn, G. Collet, E. Barre, T. Fest, O. Dameron, A. Siegel and F. Chatonnet. 'Regulus infers signed regulatory relations from few samples' information using discretization and likelihood constraints'. In: *PLoS Computational Biology* 20.1 (2024), e1011816. DOI: 10.1371/journal.pcbi.1011816. URL: https://hal.science/hal-04443527 (cit. on p. 18).

[16] B. Ruiz, A. Belcour, S. Blanquart, S. Buffet-Bataillon, I. L. Le Huërou-Luron, A. Siegel and Y. Le Cunff. 'SPARTA: Interpretable functional classification of microbiomes and detection of hidden cumulative effects'. In: *PLoS Computational Biology* 20.11 (2024), e1012577. DOI: 10.1371/journal.pcbi.1012577. URL: https://hal.science/hal-04816962 (cit. on p. 16).

[17] Y. Tirlet, M. Boudet, E. Becker, F. Legeai and O. Dameron. 'Generic and queryable data integration schema for transcriptomics and epigenomics studies'. In: *Computational and Structural Biotechnology Journal* 23 (Dec. 2024), pp. 4232–4241. DOI: 10.1016/j.csbj.2024.11.022. URL: https://inria.hal.science/hal-04818860 (cit. on p. 15).

[18] M. Zonnequin, A. Belcour, L. Delage, A. Siegel, S. Blanquart, C. Leblanc and G. V. Markov. 'Empirical evidence for metabolic drift in plant and algal lipid biosynthesis pathways'. In: *Frontiers in Plant Science* 15 (2024), p. 1339132. DOI: 10.3389/fpls.2024.1339132. URL: https://hal.science/hal-04473577 (cit. on p. 17).

International peer-reviewed conferences

[19] M. Baddour, S. Paquelet, P. Rollier, M. D. Tayrac, O. Dameron and T. Labbé. 'Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era'. In: IS 2024 - 12th IEEE International Conference on Intelligent Systems. Varna, Bulgaria, 2024, pp. 1–8. DOI: 10.1109/IS61756.2024.10705235. URL: https://inria.hal.science/hal-04647016 (cit. on p. 16).

[20] K. Thuillier, A. Siegel and L. Paulevé. 'CEGAR-Based Approach for Solving Combinatorial Optimization Modulo Quantified Linear Arithmetics Problems'. In: AAAI 2024 - The 38th Annual AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, pp. 1–8. URL: https://hal.science/hal-04420454 (cit. on p. 18).

Doctoral dissertations and habilitation theses

[21] B. Ruiz. 'Machine Learning Algorithms in the health sector : integration of functional knowledge to enhance the analysis of gut microbiota data'. INRIA Rennes - Bretagne Atlantique; INRAE, 28th Nov. 2024. URL: https://theses.hal.science/tel-04843542 (cit. on p. 16).

[22] B. Ruiz. 'Machine learning algorithms in the health sector : integration of functional knowledge toenhance the analysis of gut microbiota data'. Université de Rennes, 28th Nov. 2024. URL: https://theses.hal.science/tel-04884557.

[23] K. Thuillier. 'Hybrid satisfiability methods for the inference of boolean regulations controlling metabolic networks'. Université de Rennes, 27th Sept. 2024. URL: https://theses.hal.science/tel-04810903 (cit. on p. 18).

Reports & preprints

[24] C. M. Andreani-Gerard, N. E. Jiménez, R. Palma, C. Muller, P. Hamon-Giraud, Y. Le Cunff, V. Cambiazo, M. González, A. Siegel, C. Frioux and A. Maass. *Modeling the emergent metabolic potential of soil microbiomes in Atacama landscapes*. 2024. DOI: 10.1101/2024.12.23.630026. URL: https://hal.science/hal-04854948 (cit. on p. 17).

[25] A. Belcour, P. Hamon-Giraud, A. Mataigne, B. Ruiz, Y. L. Cunff, J. Got, L. Awhangbo, M. Lebreton, C. Frioux, S. Dittami, P. Dabert, A. Siegel and S. Blanquart. *Estimating consensus proteomes and metabolic functions from taxonomic affiliations*. 2025. DOI: 10.1101/2022.03.16.484574. URL: https://hal.science/hal-03697249.

[26] F. Moreews, J.-B. Bougaud, E. Becker, F. Gondret and O. Dameron. *BioPAX-Explorer: a Python Object-Oriented framework for overcoming the complexity of querying biological networks*. 21st Sept. 2024. DOI: 10.1101/2024.09.18.613626. URL: https://hal.science/hal-04815773 (cit. on p. 16).

Other scientific publications

[27] J. Audemard, S. Halary, G. Markov, J. Got, A. Siegel, M. Lefebvre, J. Leloup, B. Marie, N. Creusot, B. Dieme and C. Frioux. 'Metagenome-scale metabolic modelling for the characterization of cross-feeding interactions in Microcystis-associated microbial communities, in the context of freshwater cyanobacterial blooms'. In: JOBIM 2024 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Toulouse, France, 2024, pp. 1–1. URL: https://inria.hal.science/hal-04654522 (cit. on p. 17).

[28] E. Chenel. 'Searching for functional motifs in a family of multi-domain proteins: Fibulins'. Rennes: Rennes 1, 3rd July 2024, p. 38. URL: https://inria.hal.science/hal-04839500 (cit. on p. 17).

[29] E. Chevé. 'Distance-based amino acid conservation score'. Ens Rennes, 26th July 2024. URL: https://inria.hal.science/hal-04873960 (cit. on p. 17).

12.3 Cited publications

[30] G. Andrieux, M. Le Borgne and N. Théret. 'An integrative modeling framework reveals plasticity of TGF-Beta signaling'. In: *BMC Systems Biology* 8.1 (2014), p. 30. DOI: 10.1186/1752-0509-8-30. URL: http://www.hal.inserm.fr/inserm-00978313 (cit. on pp. 8, 10).

[31] 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 (Dec. 2020). DOI: 10.7554/eLife.61968. URL: https://inria.hal.science/hal-02395024 (cit. on p. 7).

[32] A. Belcour, J. Girard, M. Aite, L. Delage, C. Trottier, C. Marteau, C. J.-J. Leroux, S. M. Dittami, P. Sauleau, E. Corre, J. Nicolas, C. Boyen, C. Leblanc, J. Collén, A. Siegel and G. V. Markov. 'Inferring Biochemical Reactions and Metabolite Structures to Understand Metabolic Pathway Drift'. In: *iScience* 23.2 (Feb. 2020), p. 100849. DOI: 10.1016/j.isci.2020.100849. URL: https://hal.inria.fr/hal-01943880 (cit. on p. 9).

[33] A. Belcour, J. Got, M. Aite, L. Delage, J. Collén, C. Frioux, C. Leblanc, S. M. Dittami, S. Blanquart, G. V. Markov and A. Siegel. 'Inferring and comparing metabolism across heterogeneous sets of annotated genomes using AuCoMe'. In: *Genome Research* 33 (June 2023), pp. 972–987. DOI: 10.1101/gr.277056.122. URL: https://hal.science/hal-04192851 (cit. on p. 7).

[34] T. Berners Lee, W. Hall, J. A. Hendler, K. O'Hara, N. Shadbolt and D. J. Weitzner. 'A Framework for Web Science'. In: *Foundations and Trends in Web Science* 1.1 (2007), pp. 1–130 (cit. on p. 5).

[35] C. Bettembourg, C. Diot and O. Dameron. 'Semantic particularity measure for functional characterization of gene sets using gene ontology'. In: *PLoS ONE* 9.1 (2014). e86525. DOI: 10.1371/journal.pone.0086525. URL: https://hal.inria.fr/hal-00941850 (cit. on p. 10).

[36] S. Blanquart, J.-S. Varré, P. Guertin, A. Perrin, A. Bergeron and K. M. Swenson. 'Assisted transcriptome reconstruction and splicing orthology'. In: *BMC Genomics* 17.10 (Nov. 2016), p. 786. DOI: 10.1186/s12864-016-3103-6. URL: https://doi.org/10.1186/s12864-016-3103-6 (cit. on p. 7).

[37] P. Blavy, F. Gondret, S. Lagarrigue, J. Van Milgen and A. Siegel. 'Using a large-scale knowledge database on reactions and regulations to propose key upstream regulators of various sets of molecules participating in cell metabolism'. In: *BMC Systems Biology* 8.1 (2014), p. 32. DOI: 10.1186/1752-0509-8-32. URL: https://hal.inria.fr/hal-00980499 (cit. on pp. 8, 10).

[38] P. Bordron, M. Latorre, M.-P. Cortés, M. Gonzales, S. Thiele, A. Siegel, A. Maass and D. Eveillard. 'Putative bacterial interactions from metagenomic knowledge with an integrative systems ecology approach'. In: *MicrobiologyOpen* 5.1 (2015), pp. 106–117. DOI: 10.1002/mbo3.315. URL: https://hal.inria.fr/hal-01246173 (cit. on p. 9).

[39] M. Bouguéon, P. Boutillier, J. Feret, O. Hazard and N. Théret. 'The rule-based model approach. A Kappa model for hepatic stellate cells activation by TGFB1'. In: *Systems Biology Modelling and Analysis: Formal Bioinformatics Methods and Tools*. Ed. by E. D. Maria. Wiley, Nov. 2022, pp. 1–76. URL: https://inria.hal.science/hal-03388100 (cit. on p. 10).

[40] P. Boutillier, F. Camporesi, J. Coquet, J. Feret, K. Q. Lý, N. Théret and P. Vignet. 'KaSa: A Static Analyzer for Kappa'. In: *CMSB 2018 - 16th International Conference on Computational Methods in Systems Biology*. Ed. by M. Češka and D. Šafránek. Vol. 11095. LNCS. Brno, Czech Republic: Springer Verlag, Sept. 2018, pp. 285–291. DOI: 10.1007/978-3-319-99429-1_17. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01888951 (cit. on p. 10).

[41] A. Bretaudeau, F. Coste, F. Humily, L. Garczarek, G. Le Corguillé, C. Six, M. Ratin, O. Collin, W. M. Schluchter and F. Partensky. 'CyanoLyase: a database of phycobilin lyase sequences, motifs and functions'. In: *Nucleic Acids Research* (Nov. 2012), p. 6. DOI: 10.1093/nar/gks1091. URL: https://hal.inria.fr/hal-01094087 (cit. on p. 6).

[42] 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 (Feb. 2020), pp. 1–11. DOI: 10.3389/fmars.2020.00085. URL: https://hal.inria.fr/hal-02866101 (cit. on p. 9).

[43] J. Coquet, N. Théret, V. Legagneux and O. Dameron. 'Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-β Signaling'. In: *CMSB 2017 - 15th International Conference on Computational Methods in Systems Biology*. Lecture Notes in Computer Sciences. Darmstadt, France, Sept. 2017, p. 17. URL: https://hal.archives-ouvertes.fr/hal-01559249 (cit. on p. 10).

[44] M.-P. Cortés, S. N. Mendoza, D. Travisany, A. Gaete, A. Siegel, V. Cambiazo and A. Maass. 'Analysis of Piscirickettsia salmonis Metabolism Using Genome-Scale Reconstruction, Modeling, and Testing'. In: *Frontiers in Microbiology* 8 (Dec. 2017), p. 15. DOI: 10.3389/fmicb.2017.02462. URL: https://hal.inria.fr/hal-01661270 (cit. on pp. 7, 9).

[45] F. Coste, G. Garet, A. Groisillier, J. Nicolas and T. Tonon. 'Automated Enzyme classification by Formal Concept Analysis'. In: *ICFCA - 12th International Conference on Formal Concept Analysis*. Cluj-Napoca, Romania: Springer, June 2014. URL: https://hal.inria.fr/hal-01063727 (cit. on p. 9).

[46] O. Dennler. 'Caractérisation en modules fonctionnels de la famille de protéines ADAMTS / ADAMTSL'. MA thesis. Univ Rennes, June 2019. URL: https://hal.inria.fr/hal-02403084 (cit. on p. 10).

[47] O. Dennler. 'Caractérisation en modules fonctionnels des protéines ADAMTS-TSL, par approches de phylogénies'. Theses. Université Rennes 1, Dec. 2022. URL: https://hal.science/tel-03927428 (cit. on p. 10).

[48] O. Dennler, S. Blanquart, F. Coste, C. Belleannée and N. Theret. *Phylogenetic Functional Module Characterization of the ADAMTS / ADAMTS like Protein Family*. WABI 2021 - Workshop on Algorithms in Bioinformatics. Poster. Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03543214 (cit. on p. 10).

[49] S. M. Dittami, T. Barbeyron, C. Boyen, J. Cambefort, G. Collet, L. Delage, A. Gobet, A. Groisillier, C. Leblanc, G. Michel, D. Scornet, A. Siegel, J. E. Tapia and T. Tonon. 'Genome and metabolic network of "Candidatus Phaeomarinobacter ectocarpi" Ec32, a new candidate genus of Alphaproteobacteria frequently associated with brown algae'. In: *Frontiers in Genetics* 5 (2014), p. 241. DOI: 10.3389/fgene.2014.00241. URL: https://hal.inria.fr/hal-01079739 (cit. on p. 9).

[50] 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 (cit. on p. 9).

[51] K. Faust and J. Raes. 'Microbial interactions: from networks to models'. In: *Nat. Rev. Microbiol.* 10.8 (July 2012), pp. 538–550 (cit. on p. 6).

[52] M. Y. Galperin, D. J. Rigden and X. M. Fernández-Suárez. 'The 2015 Nucleic Acids Research Database Issue and molecular biology database collection'. In: *Nucleic acids research* 43.Database issue (2015), pp. D1–D5 (cit. on p. 5).

[53] L. Garczarek, U. Guyet, H. Doré, G. Farrant, M. Hoebeke, L. Brillet-Guéguen, A. Bisch, M. Ferrieux, J. Siltanen, E. Corre, G. Le Corguillé, M. Ratin, F. Pitt, M. Ostrowski, M. Conan, A. Siegel, K. Labadie, J.-M. Aury, P. Wincker, D. Scanlan and F. Partensky. 'Cyanorak v2.1: a scalable information system dedicated to the visualization and expert curation of marine and brackish picocyanobacteria genomes'. In: *Nucleic Acids Research* 49.D1 (Oct. 2020), pp. D667–D676. DOI: 10.1093/nar/gkaa958. URL: https://hal.archives-ouvertes.fr/hal-02988562 (cit. on p. 9).

[54] M. Gebser, R. Kaminski, B. Kaufmann and T. Schaub. *Answer Set Solving in Practice*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers, 2012 (cit. on p. 4).

[55] F. Gondret, I. Louveau, M. Houee, D. Causeur and A. Siegel. 'Data integration'. In: *Meeting INRA-ISU*. Ames, United States, Mar. 2015, p. 11. URL: https://hal.archives-ouvertes.fr/hal-01210940 (cit. on p. 10).

[56] U. Guyet, N. T. Nguyen, H. Doré, J. Haguait, J. Pittera, M. Conan, M. Ratin, E. Corre, G. Le Corguillé, L. A. Brillet-Guéguen, M. M. Hoebeke, C. Six, C. Steglich, A. Siegel, D. Eveillard, F. Partensky and L. Garczarek. 'Synergic Effects of Temperature and Irradiance on the Physiology of the Marine Synechococcus Strain WH7803'. In: *Frontiers in Microbiology* 11 (July 2020). DOI: 10.3389/fmicb.2020.01707. URL: https://hal.sorbonne-universite.fr/hal-02929424 (cit. on p. 9).

[57] F. Herault, A. Vincent, O. Dameron, P. Le Roy, P. Cherel and M. Damon. 'The longissimus and semimembranosus muscles display marked differences in their gene expression profiles in pig'. In: *PLoS ONE* 9.5 (2014). e96491. DOI: 10.1371/journal.pone.0096491. URL: https://hal.inria.fr/hal-00989635 (cit. on p. 10).

[58] 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* 32.3 (2022), pp. 703–723. DOI: 10.1111/mec.16766. URL: https://hal.science/hal-03868898 (cit. on p. 9).

[59] D. Mandakovic, Á. Cintolesi, J. Maldonado, S. Mendoza, M. Aite, A. Gaete, F. Saitua, M. Allende, V. Cambiazo, A. Siegel, A. Maass, M. Gonzalez and M. Latorre. 'Genome-scale metabolic models of Microbacterium species isolated from a high altitude desert environment'. In: *Scientific Reports* 10.1 (Dec. 2020), pp. 1–12. DOI: 10.1038/s41598-020-62130-8. URL: https://hal.inria.fr/hal-02524471 (cit. on p. 9).

[60] D. Nègre, M. Aite, A. Belcour, C. Frioux, L. Brillet-Guéguen, X. Liu, P. Bordron, O. Godfroy, A. P. Lipinska, C. Leblanc, A. Siegel, S. Dittami, E. Corre and G. V. Markov. 'Genome–Scale Metabolic Networks Shed Light on the Carotenoid Biosynthesis Pathway in the Brown Algae Saccharina japonica and Cladosiphon okamuranus'. In: *Antioxidants* 8.11 (Nov. 2019), p. 564. DOI: 10.3390/antiox8110564. URL: https://hal.inria.fr/hal-02395080 (cit. on p. 9).

[61] M. Popova, I. Fakih, E. Forano, A. Siegel, R. Muñoz-Tamayo and D. Morgavi. 'Rumen microbial genomics: from cells to genes (and back to cells)'. In: *CAB Reviews Perspectives in Agriculture Veterinary Science Nutrition and Natural Resources* 2022 (Aug. 2022). DOI: 10.1079/cabireviews202217025. URL: https://hal.inrae.fr/hal-03929845 (cit. on p. 10).

[62] S. Prigent, G. Collet, S. M. Dittami, L. Delage, F. Ethis de Corny, O. Dameron, D. Eveillard, S. Thiele, J. Cambefort, C. Boyen, A. Siegel and T. Tonon. 'The genome-scale metabolic network of Ectocarpus siliculosus (EctoGEM): a resource to study brown algal physiology and beyond'. In: *Plant Journal* (Sept. 2014), pp. 367–81. DOI: 10.1111/tpj.12627. URL: https://hal.archives-ouvertes.fr/hal-01057153 (cit. on pp. 7, 9).

[63] S. Prigent, C. Frioux, S. M. Dittami, S. Thiele, A. Larhlimi, G. Collet, G. Fabien, 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 (Jan. 2017), p. 32. DOI: 10.1371/journal.pcbi.1005276. URL: https://hal.inria.fr/hal-01449100 (cit. on p. 9).

[64] M. H. Saier, V. S. Reddy, B. V. Tsu, M. S. Ahmed, C. Li and G. Moreno-Hagelsieb. 'The Transporter Classification Database (TCDB): recent advances'. In: *Nucleic Acids Res.* 44.D1 (Jan. 2016), pp. D372–379 (cit. on p. 6).

[65] D. B. Searls. 'String variable grammar: A logic grammar formalism for the biological language of DNA'. In: *The Journal of Logic Programming* 24.1 (1995). Computational Linguistics and Logic Programming, pp. 73–102. DOI: http://dx.doi.org/10.1016/0743-1066(95)00034-H. URL: http://www.sciencedirect.com/science/article/pii/074310669500034H (cit. on p. 8).

[66] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha and G. E. Robinson. 'Big Data: Astronomical or Genomical?' In: *PLoS biology* 13.7 (2015), e1002195 (cit. on p. 5).

[67] H. Talibart. 'Comparison of homologous protein sequences using direct coupling information by pairwise Potts model alignments'. Theses. Université Rennes 1, Feb. 2021. URL: https://theses.hal.science/tel-03376771 (cit. on p. 7).

[68] H. Talibart and F. Coste. 'PPalign: optimal alignment of Potts models representing proteins with direct coupling information'. In: *BMC Bioinformatics* 22.317 (Dec. 2021), pp. 1–22. DOI: 10.1186/s12859-021-04222-4. URL: https://hal.inria.fr/hal-03264248 (cit. on p. 7).

[69] N. R. Tartaglia, A. Nicolas, V. DE REZENDE RODOVALHO, B. S. R. d. Luz, V. Briard-Bion, Z. Krupova, A. Thierry, F. Coste, A. Burel, P. P. Martin, J. Jardin, V. Azevedo, Y. Le Loir and E. Guédon. 'Extracellular vesicles produced by human and animal Staphylococcus aureus strains share a highly conserved core proteome'. In: *Scientific Reports* 10.1 (Apr. 2020), pp. 1–13. DOI: 10.1038/s41598-020-64952-y. URL: https://hal.inrae.fr/hal-02638124 (cit. on p. 7).

[70] N. Theret, F. Bouezzeddine, F. Azar, M. Diab-Assaf and V. Legagneux. 'ADAM and ADAMTS Proteins, New Players in the Regulation of Hepatocellular Carcinoma Microenvironment'. In: *Cancers* 13.7 (2021), p. 1563. DOI: 10.3390/cancers13071563. URL: https://hal.archives-ouvertes.fr/hal-03215892 (cit. on p. 10).

[71] N. Theret, J. Feret, A. Hodgkinson, P. Boutillier, P. Vignet and O. Radulescu. 'Integrative models for TGF-beta signaling and extracellular matrix'. In: *Extracellular Matrix Omics*. Ed. by S. Ricard-Blum. Vol. 7. Biology of Extracellular Matrix. Springer, Dec. 2020, p. 17. DOI: 10.1007/978-3-030-58330-9_10. URL: https://hal.inria.fr/hal-02458073 (cit. on p. 10).

[72] R. Verborgh, M. Vander Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck and P. Colpaert. 'Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web'. In: *Journal of Web Semantics* 37–38 (Mar. 2016), pp. 184–206. DOI: doi:10.1016/j.websem.2016.03.003. URL: http://linkeddatafragments.org/publications/jws2016.pdf (cit. on p. 5).