UMR IRISA

Activity Report 2021

Team DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

Joint team with Inria Rennes – Bretagne Atlantique

D7 – Data and Knowledge Management

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

A3.1.10. – Heterogeneous data

A3.2.3. – Inference

A3.2.4. – Semantic Web

A3.2.5. – Ontologies

A3.2.6. – Linked data

A3.3.3. – Big data analysis

A7.2. – Logic in Computer Science

A8.1. – Discrete mathematics, combinatorics

A8.2. – Optimization

A9.1. – Knowledge

A9.2. – Machine learning

A9.7. – AI algorithmics

A9.8. – Reasoning

Other research topics and application domains

B1.1.2. – Molecular and cellular biology

B1.1.7. – Bioinformatics

B1.1.10. – Systems and synthetic biology

B2.2.3. – Cancer

B2.2.5. – Immune system diseases

1 Team members, visitors, external collaborators

Research Scientists

- Samuel Blanquart [Inria, Researcher]
- François Coste [Inria, Researcher]
- Marine Louarn [Univ de Rennes I, Researcher, until August 2021]
- Anne Siegel [CNRS, Senior Researcher, HDR]

Faculty Members

- Olivier Dameron [Team leader, Univ de Rennes I, Professor, HDR]
- Emmanuelle Becker [Univ de Rennes I, Associate Professor]
- Catherine Belleannée [Univ de Rennes I, Associate Professor]
- Yann Le Cunff [Univ de Rennes I, Associate Professor]

PhD Students

- Meziane Aite [Insiliance SAS Paris, CIFRE, until July 2021]
- Arnaud Belcour [Inria]
- Matthieu Bougueon [INSERM]
- Nicolas Buton [Univ de Rennes I]
- Mael Conan [Univ de Rennes I, until February 2021]
- Olivier Dennler [INSERM]
- Nicolas Guillaudeux [Univ de Rennes I]
- Camille Juigné [INRAe]
- Virgilio Kmetzsch Rosa E Silva [Inria]
- Marc Melkonian [CHU Auray, from December 2021]
- Baptiste Ruiz [Inria, from October 2021]
- Hugo Talibart [Univ de Rennes I, until January 2021]
- Kerian Thuillier [CNRS, from October 2021]

Technical Staff

- Mael Conan [CNRS, Engineer, from May 2021 until June 2021]
- Jeanne Got [CNRS, Engineer]
- Leo Milhade [CNRS, Engineer, until July 2021]
- Corentin Raphalen [CNRS, Engineer]
- Hugo Talibart [Univ de Rennes I, Engineer, from March 2021 until April 2021]

Interns and Apprentices

- Lucie Baguet [Inria, from January 2021 until February 2021]
- Eve Barre [Univ de Rennes I, from January 2021 until July 2021]
- Benjamin Blanc [INRAe, from January 2021 until Jul 2021]
- Nancy D'Arminio [Univ de Rennes I, from April 2021 until July 2021]
- Sarah Guinchard [Inria, from April 2021 until September 2021]
- Baptiste Ruiz [Univ de Rennes I, from March 2021 until August 2021]
- Kerian Thuillier [Inria, from February 2021 until July 2021]

Administrative Assistant

- Marie Le Roïc [Inria]

External Collaborators

- François Moreews [INRAe]
- Denis Tagu [INRAe]
- Nathalie Théret [INSERM, 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 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) [57], 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 [68]. 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** [55]. 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 [40]. 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 there 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

[1] askomics.org

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 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 [9, 8]. 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 [54]. 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 [66].

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

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

(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 [46] to 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 [19, 32].

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 and Padmet-utils packages), and the exploration of reconstructed metabolic networks with wikis (wiki-export package, see: aureme.genouest.org/wiki.html). It also generates outputs to explore the resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae [64], extremophile bacteria [49] 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). 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

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

to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [42]. 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 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 [67].

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 [10], multi-layer reactions in interaction graphs [43], and multi-layer information in large-scale Petri nets [38]. 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 software Cell ASP Optimizer (Caspo) 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 [10]. 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 [38]. 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.

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

Marine Biology: seaweed enzymes and metabolism & sea-urchin cell-cycle. 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 enzyme families based on local and partial alignments allowed the seaweed *E. Siliculosus* metabolism to be deciphered [64, 50]. 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 [63]. We also participated in the study of the genome of *Ectocarpus subulatus*, a highly stress-tolerant algal strain [53]. Finally, AuReMe has been used to analyze the metabolic capacity of several strains of cyanobacteria, with results integrated in the Cyanorak database [56] and to characterize synergistic effects of the *synechococcus* strain WH7803 [59].

- **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. [39].

- **Algal-bacteria interactions** We reconstructed the metabolic network of a symbiot bacterium *Ca. P. ectocarpi* [52] 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 [65].

  These works paved the way to the study of host-microbial interactions, as shown in [47] 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

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 [44] and in Salmon pathogenicity [49]. 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 [62].

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) [43] and on semantic-based analysis of metabolic networks [41] was very valuable for interpreting the differences of gene expression in pork meat [60] and figure out the main gene-regulators of the response of porks to several diets [58].

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 [38, 48]. 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 cadbiom.genouest.org.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 develop 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 [51] [35].

- *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 [45]. 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 [29]. 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.

5.2 Impact of research results

Through our ongoing collaborations with INSERM, Rennes' Hospital and IPL NeuroMarkers, 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 the original motivation is to reduce fertilizers or pesticides.

6 Highlights of the year

The team has consolidated its methods and results for the description of metabolic cooperation within microbial consortia, and we are involved in several projects in the field. Regarding methological support, we are involved in the DeepInpact consortium, aiming at describing interactions between crops, soil microbiota and pathogenic organisms, and in the Holo2Plant ERC project, aiming at understanding the selective pressures on a crops-microbiota-pathogenics system.

7 New software and platforms

In 2022, the main software tools of Dyliss in the different scientific axes were enriched with new functions:

- **Integration of heterogeneous data**. The AskOmics suite was enriched with new functionalities allowing to add disjunction to queries, and to have federated queries spanning both local and remote enpoints.
- **Modeling the metabolism of large-scale species and bacteria communities** The Aureme suite was enriched in order to scale the analysis of complete families of genomes. It encompasses AuCoMe (uniformed reconstruction of metabolic networks from annotated genomes), metage2metabo (analysis of synthetic communities), mpwt (online use of the Pathway Tools environment), emapper2GBK (automatic production of genomes compatible with the pathway tools and AuReMe suite)
- **Analysis of regulations in BioPAX knowledge repositories** Pax2graphml allows to interpret BioPAX biological networks as regulated graphs, and Cadbiom allows to identify upstream regulators in such networks.
- **Protein characterisation** protomata, ppsuite and Transformer Framework for Protein Characterization participate to improve protein functions characterisation.

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).

News of the Year: 2021: (1) release 4.3.1, (2) update documentation, (3) add support for date datatype, (4) improve support for CURIEs, (5) add support for negation (still limited), (6) improve UI for dataset management, (7) add support for semantic expansion to superclasses in queries

URL: https://askomics.org/

Authors: Charles Bettembourg, Xavier Garnier, Anthony Bretaudeau, Fabrice Legeai, Olivier Dameron, Olivier Filangi, Yvanne Chaussin, Mateo Boudet

Contact: Olivier Dameron

Partners: Université de Rennes 1, CNRS, INRA

7.1.2 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.

News of the Year: (1) Improvements of the pipeline and its continuous integration (2) Release of version 1.5.0 (3) Development of m2m-analysis subpipeline

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

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.3 CADBIOM

Name: Computer Aided Design of Biological Models

Keywords: Health, Biology, Biotechnology, Bioinformatics, Systems Biology

Functional Description: The Cadbiom software provides a formal framework to help the modeling of biological systems such as cell signaling network with Guarder Transition Semantics. It allows synchronization events to be investigated in biological networks among large-scale network in order to extract signature of controllers of a phenotype. Three modules are composing Cadbiom. 1) The Cadbiom graphical interface is useful to build and study moderate size models. It provides exploration, simulation and checking. For large-scale models, Cadbiom also allows to focus on specific nodes of interest. 2) The Cadbiom API allows a model to be loaded, performing static analysis and checking temporal properties on a finite horizon in the future or in the past. 3) Exploring large-scale knowledge repositories, since the translations of the large-scale PID repository (about 10,000 curated interactions) have been translated into the Cadbiom formalism.

News of the Year: 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.

URL: http://cadbiom.genouest.org

Contact: Anne Siegel

Participants: Geoffroy Andrieux, Michel Le Borgne, Nathalie Theret, Nolwenn Le Meur, Pierre Vignet, Anne Siegel

7.1.4 pax2graphml

Name: pax2graphml - Large-scale Regulation Network in Python using BIOPAX and Graphml

Keyword: Bioinformatics

Functional Description: PAX2GRAPHML is an open source python library that allows to easily manipulate BioPAX source files as regulated reaction graphs described in .graphml format. PAX2GRAPHML is highly flexible and allows generating graphs of regulated reactions from a single BioPAX source or by combining and filtering BioPAX sources. Supporting the graph exchange format .graphml, the large-scale graphs produced from one or more data sources can be further analyzed with PAX2GRAPHML or standard python and R graph libraries.

News of the Year: The code of Pax2graphml has been refactored and extended for including new reaction graph manipulation features. We have also recoded the RDF import module. New compatible datasets have been generated from 17 BIOPAX data sources. A landing page and a demo jupyter notebook and documentation have been created.

The article "PAX2GRAPHML: a Python library for large-scale regulation network analysis using BIOPAX" was published in Bioinformatics (https://hal.archives-ouvertes.fr/hal-03265223)

URL: https://pax2graphml.genouest.org/

Publication: hal-03265223

Contact: François Moreews

Partner: INRAE

7.1.5 Protomata

Keywords: Proteins, Machine learning, Pattern discovery, Grammatical Inference, Bioinformatics

Scientific Description: Inference of automata modelling protein sequences by partial local alignment

Functional Description: This tool is a grammatical inference framework suitable for learning the specific signature of a functional protein family from unaligned sequences by partial and local multiple alignment and automata modelling. It performs a syntactic characterization of proteins by identification of conservation blocks on sequence subsets and modelling of their succession. Possible fields of application are new members discovery or study (for instance, for site-directed mutagenesis) of, possibly non-homologous, functional families and subfamilies such as enzymatic, signalling or transporting proteins.

Given a sample of sequences belonging to a structural or functional family of proteins, Protomata-Learner infers an automaton characterizing the family by partial local alignment of the sequences. Automata are graphical models representing a (potentially infinite) set of sequences. Able to express alternative local dependencies between the positions, automata offer a finer level of expressivity than classical sequence patterns (such as PSSM, Profile HMM, or Prosite Patterns) and can model more than homologous sequences. They are well suited to get new insights into a family or to search for new family members in the sequence data banks, especially when approaches based on classical multiple sequence alignments are insufficient.

The three main modules integrated in the Protomata-learner workflow are available as well as stand-alone programs: 1) paloma builds partial local multiple alignments, 2) protobuild infers

automata from these alignements and 3) protomatch and protoalign scans, parses and aligns new sequences with learnt automata. The suite is completed by tools to handle or visualize data and can be used online by the biologists via a web interface on Genouest Platform.

News of the Year: Implementation of a new and faster version of paloma in modern C++ relying on a new definition of partial local alignments.

URL: http://tools.genouest.org/tools/protomata/

Contact: François Coste

Participant: François Coste

Partners: Université de Rennes 1, CNRS, Inria

7.1.6 PPsuite

Keywords: Proteins, Sequence alignment, Bioinformatics, Machine learning, Homology search

Scientific Description: Comparison of protein sequences using coevolutionary dependencies between residues.

Functional Description: This suite contains the following tools : - MakePotts infers a Potts model from a sequence or a multiple sequence alignment - PPalign aligns Potts models and corresponding sequences - VizPotts allows to visualize inferred Potts models and VizContacts allows to visualize inferred couplings with respect to actual contacts in a 3D protein structure.

News of the Year: The workflow have been extended to enable modeling position-specific insertion and deletion costs. The rescaling of the models in MakePotts has been rewritten in C++ increasing the speed of the program tenfold and a mean field approach (mfDCA) has been integrated as an option in MakePotts for the inference of Potts models. The exploration and the optimisation of the hyper parameters of the method rely now on the Optuna framework which provides better analysis tools.

URL: https://www-dyliss.irisa.fr/ppalign/

Publications: hal-02862213, hal-02402646, hal-03264248

Authors: Hugo Talibart, François Coste

Contact: François Coste

7.1.7 Transformer Framework for Protein Characterization

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. 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.

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

Contact: Nicolas Buton

Participants: Nicolas Buton, Yann Le Cunff, François Coste

7.1.8 Emapper2GBK

Keywords: Bioinformatics, Metabolic networks, Functional annotation

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

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

Publication: hal-02395024

Contact: Clemence Frioux

Participants: Clemence Frioux, Arnaud Belcour, Anne Siegel

7.1.9 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

Contact: Anne Siegel

7.1.10 mpwt

Keywords: Metabolic networks, Multi-processor

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

Publication: hal-02395024

Contact: Anne Siegel

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

8 New results

8.1 Scalable methods to query data heterogeneity

Participants Emmanuelle Becker, Olivier Dameron, Francois Moreews, Anne Siegel.

PAX2GRAPHML: a Python library for large-scale regulation network analysis using BIOPAX [*F. Morrews*] [18].

- The concept of regulated reactions, which allows connecting regulatory, signaling and metabolic levels, has been used to easily manipulate BioPAX source files as regulated reaction graphs. Biochemical reactions and regulatory interactions are homogeneously described by regulated reactions involving substrates, products, activators and inhibitors as elements.

Converting disease maps into heavyweight ontologies [*O. Dameron*] [15].

- In the context of our participation to the IPL NeuroMarker, we designed the Disease Map Ontology (DMO), an ontological upper model based on systems biology terms. We then applied DMO to Alzheimer's disease (AD). Specifically, we used it to drive the conversion of AlzPathway, a disease map devoted to AD, into a formal ontology called Alzheimer DMO.

Pharmaco-epidemiological queries over administrative healthcare databases [*O. Dameron*] [23, 27].

- Chronicles are a relevant formalism for representing complex temporal queries over healthcare patient trajectories while retaining acceptable performances. However, they lack a proper semantic support for handling generalisation. Conversely, Semantic Web techniques adequately handle generalization and can represent temporal constraints, but the latter remain a performance bottleneck. We proposed an hybrid approach combining chronicles and Semantic Web queries and demonstrated its capacity to detect patients having venous thromboembolism disease in the French medico-administrative database [23].

- Generating synthetic data for administrative healthcare databases allows to perform research on healtcare data without compromising patients privacy. We proposed a probabilistic relational model fitted on publicly available datasets that generates synthetic versions of the national database of French insured patients and mimic statistical distributions but do not hold sensitive personal data [27].

8.2 Metabolism: from protein sequences to systems ecology

Participants Arnaud Belcour, Benjamin Blanc, Samuel Blanquart, Mael Conan, François Coste, Jeanne Got, Anne Siegel, Hugo Talibart, Nathalie Théret.

Detection of genomic recombinations by partial local alignment [*B. Blanc, F. Coste*] [33].

- In collaboration with Marie-Agnès Petit (Phage team, MICALIS, Inrae), we investigated how `paloma` (the partial local multiple sequence alignment tool from Protomata suite) could help studying recombination in proteins from 32 phages, of which some have already been recombined according to the literature. Classical multiple sequence alignment are not suitable for this task. In contrast, the generated partial local alignments allowed to find recombined regions in 8 phages described by the past in 3 phages, and the presence of 4 conserved sequences between these 8 phages around the recombined region which could be recombination fingerprints. [33]

Modeling proteins with crossing dependencies [*F. Coste, H. Talibart*] [19, 32]

- Motivated by their success on contact prediction, we proposed to use Potts models to represent proteins with direct couplings between positions — in addition to positional composition — and compare them by aligning optimally these models thanks to an Integer Linear Programming formulation of the problem. We worked on the inference of robust and more canonical Potts models. We assessed the approach with respect to a non-redundant set of reference pairwise sequence alignments with low sequence identity, showing that Potts models representing proteins can be aligned in reasonable time and that considering couplings can improve significantly the alignments with respect to other methods [19, 32].

Large-scale eukaryotic metabolic network and design of microbial communities [*A. Siegel, A. Belcour, S. Blanquart, J. Got, N. Théret, M. Conan*] [20, 14, 13, 30, 26, 25, 24, 37, 16].

- *Metabolic data analysis enhanced by large-scale metabolic network reconstruction* We used our tools for the reconstruction and analysis of large-scale metabolic networks to provide insights on Ulva compressa, a green tide-forming species, from transcriptome-wide gene expression profiles [20]. We also benefited from the availability of genome data and gas chromatography-mass spectrometry (GC-MS) sterol profiling using a database of internal standards to build such a model of sterol biosynthesis in brown algae [14]. Our results demonstrate that integrative approaches can already be used to infer experimentally testable models, which will be useful to further investigate the biological roles of those newly identified algal pathways.

- *Metabolic pathway inference from non genomic data* We developed a modeling approach in order to predict all the possible metabolite derivatives of a xenobiotic. Our approach relies on the construction of an enriched and annotated map of derivative metabolites from an input metabolite. The pipeline assembles reaction prediction tools (SyGMa), sites of metabolism prediction tools (Way2Drug, SOMP and Fame 3), a tool to estimate the ability of a xenobotics to form DNA adducts (XenoSite Reactivity V1), and a filtering procedure based on Bayesian framework. The method was applied to determine enzyme profiles associated with the maximization of DNA adducts formation derived from each HAA [13, 30]

- *Design of synthetic microbiota* We presented the tool Metage2Metabo (microbiota-scale metabolic complementarity for the identification of key species) in several conferences [26, 25, 24, 37]. Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines.

- *Impact of genome annotations procedures on the design of synthetic microbiomes* [16] As there are multiple annotation pipelines available, the question arises to what extent differences in annotation pipelines impact outcomes of genome-scale metabolic network reconstructions. We compared five commonly used pipelines (Prokka, MaGe, IMG, DFAST, RAST) from predicted annotation features to the metabolic network-based analysis of symbiotic communities (biochemical reactions, producible compounds, and selection of minimal complementary bacterial communities). The consortia generated yielded similar predicted producible compounds and could therefore be considered functionally interchangeable.

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

Participants Emmanuelle Becker, Catherine Belleannée, Samuel Blanquart, Mathieu Bougueon, François Coste, Olivier Dennler, Samuel Blanquart, Olivier Dameron, Nicolas Guillaudeux, Virgilio Kmetzsch, Anne Siegel, Kérian Thuillier, Nathalie Théret.

Learning Boolean controls in regulated metabolic networks: a case-study [*A. Siegel, K. Thuillier*] [22]

- Many techniques have been developed to infer Boolean regulations from a prior knowledge network and experimental data. Existing methods are able to reverse-engineer Boolean regulations for transcriptional and signaling networks, but they fail to infer regulations that control metabolic networks. We provided a formalization of the inference of regulations for metabolic networks as a satisfiability problem with two levels of quantifiers, and introduces a method based on Answer Set Programming to solve this problem on a small-scale example.

Functional signature for ADAMTS [*C. Belleannée, S. Blanquart, F. Coste, O. Dennler, N. Théret*] [35].

- Hepatic Stellate Cells produce a wide variety of molecules involved in ECM remodeling, such as adamalysins (hal-03215892). 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 develop 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.

Creation of predictive functional signaling networks [*M. Bougueon, N. Théret*] [29, 21].

- *The rule-based model approach. A Kappa model for hepatic stellate cells activation by TGFB1* [29, 21] Kappai is a site graph rewriting language. It offers a rule-centric approach, inspired from chemistry, where interaction rules locally modify the state of a system that is defined as a graph of components, connected or not. In this case study, the components will be occurrences of hepatic stellate cells in different states, and occurrences of the protein TGFB1. The protein TGFB1 induces different behaviors of hepatic stellate cells thereby contributing either to tissue repair or to fibrosis. Better understanding the overall behavior of the mechanisms that are involved in these processes is a key issue to identify markers and therapeutic targets likely to promote the resolution of fibrosis at the expense of its progression.

Evidence of a microRNA signature for frontotemporal lobar degeneration and amyotrophic lateral sclerosis [*E. Becker, V. Kmetzsch*] [61].

- In the context of our participation in the IPL NeuroMarker project, a joint study with Institut du Cerveau (Inserm/CNRS/Sorbonne Université) at the Pitié-Salpêtrière hospital and the Aramis team (Inria Paris) evidenced a signature of four plasma microRNAs in presymptomatic and symptomatic subjects with frontotemporal dementia and amyotrophic lateral sclerosis associated with a C9orf72 mutation[13]. The four microRNAs' expression level allows to discriminate patients, presymptomatic or healthy individuals. The study was conducted by Virgilio Kmetzsch in his PhD supervized by Olivier Colliot (Aramis) and Emmanuelle Becker (Dyliss). Future steps will study how combining this signature with medical imaging can refine the classification or can result in a score for characterizing the disease progression.

Characterizing gene structure with grammatical languages and conservation information [*C. Belleannée, S. Blanquart, O. Dameron, N. Guillaudeux*] [31]

- Based on syntactic models and graph formalisms, we compared splicing structures of 2167 triplets of orthologous genes shared in human, mouse and dog. This resulted in the prediction of 6861 new coding transcripts (*i.e.* putative proteins) on these species, mainly for dog, an emergent model species. Every predicted transcript shares an identical exonic structure with a coding transcript already known in another species, hence defining them as orthologs. Additionnaly, we identified a set 253 gene triplets with strictly conserved exonic structures in human, mouse and dog, and so expressing the same proteome (*i.e.* the same isoform coding transcripts). These genes express a total of 879 groups of orthologous isoforms, such that in each group, the same splicing structure is shared in each three species gene. Although these genes express a same proteome, we showed that the expressed transcriptomes may be different, due to the gene's propensity to express distinct alternatively transcribed mRNAs encoding the same protein.

[13] institutducerveau-icm.org/fr/actualite/sla-dft-mutation-gene-c9orf72/

Estimating ancestral phenotypes of halophilic enzymes using phylogenetic inferences [S. Blanquart] [12]

- Ancestral sequence reconstruction approaches aim at synthesizing ancient genes, which are estimated using phylogenetic methods, in order to experimentally measure the product's phenotypes. In such a study, we investigated the adaptation of the ancestral malate dehydrogenase enzymes of extrem halophilic Archea. Applying advanced phylogenetic approaches, we infered and synthesized ancient enzyme sequences. We described the phenotype of a transferred enzyme, the evolutionnary drift phenomenon and a secondary adaptation to alkaliphic lifestyle. The stabilisation of tetrameric assembly by ions appeared to modulate the enzymes adaptation to extremely salted environments [12].

Establishing an inventory in human genome of a transposable element with help of grammatical patterns [A. Antoine-Lorquin, C. Belleannée] [11]

- Transposable elements are repeated DNA sequences that represent 45% of the human genome. They play a critical role in genome organization and its evolution. Among them, MADE1 is a 80 bp element with a special structure, being flanked on both ends by short sequences repeated in inverse orientation. The use of grammatical patterns with our Logol tool [2] contributed to characterize the structural MADE1 variants and to establish an exhaustive inventory of MADE1 elements [11].

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

INSILIANCE: co-supervised PhD

Participants Méziane Aite , Olivier Dameron.

This collaborative project is focused on identifying candidate combinations of repositioned drugs for central nervous system's diseases. It evolved from last year's collaboration with Theranexus. **CIFRE co-supervised Grant: PhD. funding. 2020-2023. The collaboration ended prematurely with INSILIANCE liquidation on June 1st 2021.**

Biofortis Mérieux nutrisciences: internship and data sharing

Participants Yann Le Cunff , Baptiste Ruiz.

This collaborative project involved partners from Rennes Hospital (CHU), the INRAE team NuMeCan and the R&D deptartment of Biofortis Mérieux Nutriscience. It focused on using non-supervised machine learning technics to classify patients' microbiota in the context of ovarian cancer.

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: 2020–2026

Coordinator: Alejandro Maass (amaass@dim.uchile.cl)

Partners:

- Universidad de Chile
- Pleiade (Bordeaux, France)
- Pontificia Universidad Católica (Santiago de Chile, Chile)
- Inria Chile.

Inria contact: Anne Siegel

Summary: SymBioDiversity is an Associate Team between the Inria project team Dyliss located in Rennes, France and Mathomics department of the Center for Methematical Modeling (Universidad de Chile), located in Santiago de Chile, Chile. Through the combination of data mining, reasoning and mathematical modeling, this team aims at developping approaches for the analysis of the microbial diversity in extreme environments as well as characterising the functional landscape of these ecosystems. team.inria.fr/symbiodiversity/

10.2 National initiatives

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

Participants Samuel Blanquart , Arnaud Belcour , Olivier Dameron , Jeanne Got , 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. TODO 2021–2024. Dyliss grant: 120k€.

IDEALG (ANR/PIA-Biotechnology and Bioresource)

Participants Arnaud Belcour , François Coste , Jeanne Got , Anne Siegel , Hugo Talibart.

The project gathers 18 partners from Station Biologique de Roscoff (coordinator), CNRS, IFREMER, UEB, UBO, UBS, ENSCR, University of Nantes, INRA, AgroCampus, and the industrial field in order to foster biotechnology applications within the seaweed field. Dyliss is co-leader of the WP related to the establishment of a virtual platform for integrating omics studies on seaweed and the integrative analysis of seaweed metabolism. Major objectives are the building of brown algae metabolic maps, metabolic flux analysis and the selection of symbiotic bacteria for brown algae. We will also contribute to the prediction of specific enzymes (sulfatases and haloacid dehalogenase)[14]. 2012–2021. Total grant: 11M€. Dyliss grant: 534k€.

PhenomiR

Participants Emmanuelle Becker , Olivier Dameron , Leo Mihlade , Anne Siegel.

The objective of the PhenomiR project is to propose an innovative solution for non-invasive phenotyping by analysing circulating microRNAs (miRNAs) (present in plasma) or present in biological fluids (coelomic fluid) and identify relevant biomarkers by the integration of omics data at multiple layers and to test to what extent the miRNAs of interest in trout are well conserved in fish genomes that are relatively complete. The PhenomiR project is carried out on rainbow trout (*Oncorhynchus mykiss*) which is both a major/principal production for the French fish farming industry and also a historical model species for INRAe and the research laboratories involved in the fields of physiology, nutrition, well-being/behaviour and infectiology/immunology. 2019–2022.

10.2.1 Programs funded by Inria

IPL Neuromarkers

Participants Emmanuelle Becker , Olivier Dameron , Virgilio Kmetzsch , Anne Siegel.

This project involves mainly the Inria teams Aramis (coordinator) Dyliss, Genscale and Bonsai. The project aims at identifying the main markers of neurodegenerative pathologies through the production and the integration of imaging and bioinformatics data. Dyliss is in charge of facilitating the interoperability of imaging and bioinformatics data. In 2019 V. Kmetzsch started his PhD (supervized by E. Becker from Dyliss and O. Colliot from Aramis). 2017–2020.

10.3 Regional initiatives

PROLIFIC

Participants Corentin Raphalen , Anne Siegel .

[14] idealg.u-bretagneloire.fr/

The PROLIFIC (PROduits Laitiers et Ingrédients Fermentés Innovants pour des populations Cibles) research project will evaluate the health benefits of fermented dairy products for young children and seniors. The project is led by a consortium of companies grouped within Bba Milk Valley and research teams from Brittany and the Loire Valley. The researchers will study bacteria from a collection of microorganisms (CIRM-BIA) or isolated from maternal milk samples. Using *in silico* (modeling), *in vitro* (cell culture) and *in vivo* (animal models) devices, they will look in particular at their capacity to activate the intestine-brain axis and their potential to participate in the cognitive development of children or in the prevention of neurodegeneration in seniors. They will also study the capacity of these bacteria to stimulate the immune system to prevent the onset of food allergies and inflammatory diseases. 2020–24. Dyliss grant: 100k€.

Pepper (projet Émergence 2021-2022 de l'Alliance Sorbonne Université)

Participants François Coste.

The project Pepper, coordinated by Mathilde Carpentier from ISYEB (Institut de Systématique, Évolution, Biodiversité), aims at proposing a new generation of practical tools based on Potts models for the search and alignment of homologous protein sequences. In continuation of his PhD in Dyliss, Hugo Talibart is working as a postdoc in the Muséum National d'Histoire Naturelle (under the supervision of M. Carpentier and F. Coste) to enhance PPsuite with necessary practical refinements and test its application on viral protein sequences.

11 Dissemination

Participants Emmanuelle Becker, Catherine Belleannée, Samuel Blanquart, Francois Coste, Olivier Dameron, Yann Le Cunff, Anne Siegel, Nathalie Théret.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

- Jobim 2022, Rennes [C. Belleannée, F. Coste]

Chair of conference program committees

- Jobim 2022, Rennes [E. Becker]

Member of the conference program committees

- ICGI (International Conference on Grammatical Inference), 2020/21 [F. Coste]
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database ECML/PKDD [O. Dameron]
- OnUCAI-KR2021 (Ontology Uses and Contribution to Artificial Intelligence) [O. Dameron]
- Journée Santé et IA 2021, workshop organized by AFIA and AIM [O. Dameron]
- Workshop on Answer Set Programming and Other Computing Paradigms 2021 [A. Siegel]
- ISMB/ECCB 2021 (Intelligent Systems for Molecular Biology and European Conference on Computational Biology 2021) [A. Siegel]

11.1.2 Journal

Member of the editorial boards

- Editor of a special issue in grammatical inference of Machine Learning journal [F. Coste]

Reviewer - reviewing activities

- Briefings in Bioinformatics [O. Dameron]
- Journal of Biomedical Semantics [O. Dameron]
- BioSystems [Y. Le Cunff]

11.1.3 Invited talks

- "Ontologies and Semantic Web in Life Sciences" CATI SICPA INRAe (May 20th 2021, O. Dameron)
- "Semantic Web hands-on tutorial" CATI SICPA INRAe (July 07th 2021, O. Dameron)
- "Modeling biological systems for unconventional organisms: from dynamical systems to automated reasoning" GDR Intelligence Artificielle (May 2021, A. Siegel)
- "A quoi peut servir un GT égalité femmes-hommes ?", Seminar of the LIP6 laboratory (June 2021, A. Siegel)
- "Metabolic network models : from formats to workflows", [BC]2 workshop -Toward a common framework for annotated, accessible, reproducible and interoperable computational models in biology, Basel (September 2021, A. Siegel)
- "Building genome scale metabolic models: good (or at least not too bad) practices ?", Workshop sur la Modélisation du métabolisme, Bordeaux (November 2021, A. Siegel)
- "Harnessing prior knowledge to improve AI algorithms for patients' representation and classification", Knowledge Summit 3 : IA & Santé (November 2021, Y. le Cunff)

11.1.4 Scientific expertise

Recruitment committees

- Associate professor LRU "computer science", Agrocampus Rennes [O. Dameron]
- Associate professor "bioinformatics", Univ. Bordeaux - poste 590 [O. Dameron]
- Engineer, CNRS [A. Siegel]

National scientific boards

- INRAE scientific board of the MIA department [A. Siegel]
- Programme Prioritaire de Recherche "Autonomie" [A. Siegel]
- Groupement de Recherche MaMoVi "MAthématiques de la MOdélisation du VIvant" [A. Siegel]
- ModCov19 coordination committee [A. Siegel]
- Animation of the Systems Biology working group of national infrastructure GDR IM and GDR BIM [A. Siegel].
- Board of directors of the French Society for biology of the extracellular matrix [N. Théret].

Project evaluation

- Cofund AI4theSciences, Paris Science Lettre [A. Siegel]

Local responsibilities

- Social committee of Univ. Rennes 1 [C. Belleannée]
- Emergency aid commission of Univ. Rennes 1 & Rennes 2 [C. Belleannée]
- Organisation of the bioinformatics teams (Dyliss, GenOuest and GenScale as well as members of other bioinformatics teams in Rennes; 138 members for the mailing list) weekly seminars [S. Blanquart]
- Scientific Advisory Board of the GenOuest platform [O. Dameron]
- Member of the Inria Rennes center council [J. Got]
- Member of the Biology department council [Y. Le Cunff]
- Scientific Advisory Board of Biogenouest [N. Théret]
- Delegate to research integrity at the University of Rennes 1 [N. Théret]
- Organisation of the monthly seminar "Data and Knowledge management" department of Irisa [A. Siegel]

11.1.5 Research administration

Institutional boards for the recruitment and evaluation of researchers

- National Council of Universities (Conseil National des Universités - CNU), section 27, since Dec 2019 [F. Coste]

National responsibilties

- Bioinformatics Scientific Advisor at CNRS (INS2I), until september 2021 [A. Siegel]
- Deputy Scientific Directory (CNRS, INS2I), in charge of interdisciplinarity between numerical sciences and other disciplines, gender equality in computer sciences, groupements de recherches (GDR), since september 2021 [A. Siegel]

Local responsibilties

- Head of the "Data and Knowledge Managment" Department (6 teams) of the IRISA lab, until octobrer 2021 [A. Siegel]
- Gender equality commission, IRISA & Inria Rennes, until september 2021 [A. Siegel, coordinator]
- CUMI (Commission des utilisateurs des moyens informatiques) of Inria Rennes [F. Coste]

11.2 Teaching - Supervision - Juries

11.2.1 Teaching tracks responsibilities

- Coordination of the doctoral school "Biology and Health" of University of Bretagne Loire, Rennes [N. Théret]
- Coordination of the master degree "Bioinformatics", Univ. Rennes [E. Becker, O. Dameron]
- Organization of the open day of the UFR of computer science and electronics, Univ. Rennes (journée portes ouvertes Istic) [C. Belleannée]

11.2.2 Course responsibilities

- "Method", Master 2 in Computer Sciences, Univ. Rennes 1 [E. Becker]
- "Statistiques appliquées", 3rd year in Fundamental Computer Sciences, ENS Rennes [E. Becker]
- "Introduction to computational ecology", Master 2 in Ecology, Univ. Rennes 1 [E. Becker]
- "Object-oriented programming", Master 1 in Bioinformatics, Univ. Rennes 1 [E. Becker]
- "Advanced R for data analysis", Master 1 in Ecology + Master 1 in Bioinformatics, Univ. Rennes 1 [E. Becker]
- "Insertion Professionnelle et tables rondes", Master 1 and Master 2 in Bioinformatics, Univ. Rennes 1 [E. Becker]
- "Atelier de Biostatistiques", 2nd year Biology, Univ Rennes 1 [E. Becker]
- "Internship", Master 1 in Computer Sciences, Univ. Rennes 1 [C. Belleannée]
- "Supervised machine learning", Master 2 in Computer Sciences, Univ Rennes 1 [F. Coste]
- "Imperative programming", Licence 1 informatique, Univ. Rennes 1 [O. Dameron]
- "Complément informatique 1", Licence 1 informatique, Univ. Rennes 1 [O. Dameron]
- "Atelier bioinformatique", Licence 2 informatique, Univ. Rennes 1 [O. Dameron]
- "Semantic Web and bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Internship", Master 2 in bioinformatics, Univ. Rennes 1 [O. Dameron]
- "Integrative and Systems biology", Master 2 in bioinformatics, Univ. Rennes 1 [A. Siegel]
- "Micro-environnement Cellulaire normal & pathologique", Master 2 Biologie cellulaire et Moléculaire, Univ. Rennes 1 [N. Théret]
- "Machine Learning", Master 1 in Bioinformatics [Y. le Cunff]
- "Modeling dynamic systems", Licence 2, Biology [Y. Le Cunff]
- "Simulating Biological Systems", Master 2 in Bioinformatics [Y. Le Cunff]
- "Simulation and biology interfaces", Master 1 in Biology [Y. Le Cunff]
- "Applied interdisciplinarity", Master 2 in Biology [Y. Le Cunff]

11.2.3 Teaching

- Licence : E. Becker, "Atelier de Biostatistiques", 34h, 2nd year in Biology, Univ. Rennes 1, France
- Licence : E. Becker, "Statistiques Appliquées", 20h, 3rd year in Fundamental Computer Sciences, ENS Rennes, France
- Master : E. Becker, "Object oriented programming", 56h, Master 1 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Advanced R for data analysis", 36h, Master 1 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Introduction to computational ecology", 34h, Master 2 in Ecology, Univ. Rennes 1, France
- Master : E. Becker, "Method", 15h, Master 2 in Computer Sciences, Univ. Rennes 1, France

- Master : E. Becker, "Insertion Professionnelle et tables rondes", 6h, Master 1 and Master 2 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Systems Biology : biological netorks", 27h, Master 2 in Bioinformatics, Univ. Rennes 1, France
- Master : E. Becker, "Introduction to Bioinformatics", 3h, Master MEEF Biology, Univ. Rennes 1, France.
- Licence: C. Belleannée, Langages formels, 20h, L3 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Projet professionnel et communication, 16h, L1 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Enseignant référent, 20h, L1 informatique, Univ. Rennes 1, France.
- Licence: C. Belleannée, Spécialité informatique : Functional and immutable programming , 42h, L1 informatique, Univ. Rennes 1, France
- Master: C. Belleannée, Algorithmique du texte et bioinformatique, 10h, M1 informatique, Univ. Rennes 1, France
- Master: C. Belleannée, Programmation logique et contraintes, 32h, M1 informatique, Univ. Rennes 1, France
- Master: F. Coste, Supervised machine learning, 10h, M2 Science Informatique, Univ. Rennes, France
- Licence: O. Dameron, "Programmation 1", 40h, Licence 1 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Complément informatique", 24h, Licence 1 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Atelier bioinformatique", 24h, Licence 2 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Databases", 24h, Licence 2 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Programmation", 54h, Licence 3 miage, Univ. Rennes 1, France
- Master: O. Dameron, "Semantic Web", 20h, Master 1 miage, Univ. Rennes 1, France
- Master: O. Dameron, "Veille technologique", 2h, Master 2 miage, Univ. Rennes 1, France
- Master: O. Dameron, 2h, "Internship", Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 20h, "Semantic Web and bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 18h, "Internship", Master 2 in bioinformatics, Univ. Rennes 1, France
- Licence: N. Guillaudeux, Projet professionnel et communication, 16h, 1st year Computer Science, Univ. Rennes 1, France
- Licence: N. Guillaudeux, "TPs Python", 36h, 1st year in Biology, Univ. Rennes 1, France
- Licence: M. Louarn, Introduction à la BioInformatique, 6h, L2 Informatique, Univ. Rennes 1, France.
- Licence: M. Louarn, Informatique, 10h, L1 Physique Chimie, Univ. Rennes 1, France.
- Master: M. Louarn, Informatique Médicale Avancée, 2h, M1 Médecine, Univ. Rennes 1, France.
- Master: M. Louarn, Object-oriented programming, 25h, M2 bioinformatique et génomique, Univ. Rennes 1, France.

- Master: M. Louarn, Jury de stage, 6h, M1 bioinformatique et génomique, Univ. Rennes 1, France.
- Master: A. Siegel, Integrative and Systems biology, Master 2 in bioinformatics, Univ. Rennes 1.
- Licence : Y. Le Cunff "Modélisation des phénomènes du vivant", 30h, L2 Biologie, Univ. Rennes 1, France
- Master: Y. Le Cunff, "Apprentissage statistique", 110h, Master 1 in Bioinfortmatics Univ. Rennes 1, France
- Master: Y. Le Cunff, "Biologie aux interfaces", 25h, Master 1 in Biology, Univ. Rennes 1, France
- Master: Y. Le Cunff,"Simulating dynamic systems in biology", Master 2 in bioinformatics, 20h, Univ. Rennes 1, France
- Master: Y. Le Cunff, "Applied Interdisciplinarity", 20h, Master 2 in biology, Univ. Rennes 1, France
- PhD program: Y. Le Cunff,"Introduction to Machine Learning", 20h, FdV PhD Program, Sorbonne Paris Université, Paris, France

11.2.4 Supervision

- PhD: Hugo Talibart, *Comparison of homologous protein sequences using direct coupling information by pairwise Potts model alignments*, defended February 24th 2021, supervised by F. Coste and J. Nicolas (GenScale). [32]
- PhD: Maël Conan, *Approche prédictive pour évaluer la génotoxicité des contaminants de l'environnement*, defended 23rd March 2021, supervised by A. Siegel and S. Langouët. [30]
- PhD: Nicolas Guillaudeux, *Comparer des structures de gènes pour la prédiction de transcrits alternatifs codants chez l'humain, la souris et le chien*, defended December 16th 2021, supervised by O. Dameron, S. Blanquart and C. Belleannée. [31]
- PhD in progress: Johanne Bakalara, *Temporal models of care sequences for the exploration of medico-administrative data*, started in Oct. 2018, supervised by T. Guyet (Lacodam), E. Oger (Repères) and O. Dameron.
- PhD in progress: Arnaud Belcour, *Inferring Model metabolisms for bacterial ecosystems reduction*, started in Oct. 2019, supervised by A. Siegel and S. Blanquart.
- PhD in progress: Matthieu Bouguéon, *Modélisation prédictive pour le ciblage thérapeutique du TGF-beta dans les pathologies chroniques hépatiques*, started in Oct. 2020, supervised by N. Théret and A. Siegel.
- PhD in progress: Nicolas Buton, *Deep learning for proteins functional annotation : novel architectures and interpretability methods*, started in Oct. 2020, supervised by F. Coste, Y. Le Cunff and O. Dameron.
- PhD in progress: Olivier Dennler, *Modular functional characterization of ADAMTL and ADAMTSL protein families*, started in Oct. 2019, supervised by N. Theret, F. Coste, S. Blanquart and C. Belleannée.
- PhD in progress: Camille Juigné, *Analyse des données biologiques hétérogènes par exploitation de graphes multicouches pour comprendre et prédire les variations d'efficacité alimentaire chez le porc*, started in Dec. 2020, supervised by E. Becker and F. Gondret (INRAE Pegase).
- PhD in progress: Virgilio Kmetzsch *Multi-modal analysis of neuroimaging and transcriptomics data in genetically-induced fronto-temporal dementia*, started in Oct. 2019, supervised by E. Becker and O. Colliot (INRIA Aramis, ICM Paris)
- PhD in progress: Marc Melkonian, *Intégration de données et de connaissances pour l'analyse fine de l'interactome*, started in Dec. 2021, supervised by E. Becker and G. Rabut (IGDR).

- PhD in progress: Baptiste Ruiz *Algorithmes d'apprentissage automatique appliqués au microbiote : Intégration de connaissances a priori pour de meilleures prédictions de phénotype*, started in Oct. 2021, supervised by Y. Le Cunff and A. Siegel.
- PhD in progress: Kerian Thuillier, *Inférence de règles booléennes contrôlant des modèles hybrides de systèmes biologiques multi-échelles*, started in Oct. 2021, supervised by A. Siegel and L. Paulevé (LABRI)
- PhD interrupted in 2021: Méziane Aite. *Identification de nouvelles combinaisons thérapeutiques dans les indications neurologiques*, started in Nov. 2020, supervised by O. Dameron and V. Lafon (Insiliance).
- PhD interrupted in 2021: Pierre Vignet, *Identification et conception expérimentale de nouveaux agents thérapeutiques à partir d'un modèle informatique des réseaux d'influence du TGF-beta dans les pathologies hépatiques chroniques*, started in Dec. 2018, supervised by N. Théret and A. Siegel.
- M2 Internship: Ève Barré, *Analyse de réseaux de régulation de la transcription et priorisation des facteurs de transcription, introduction de variants non codants*, Jan. – Jul. 2021, co-supervised by M. Louarn and O. Dameron
- M2 Internship: Benjamin Blanc, *Detection of genomic and proteic recombinations in phages by partial local alignment*, Jan. – Jul. 2021, co-supervised by F. Coste
- M2 Internship: Nancy d'Arminio (Erasmus+ exchange with Salerno Univ., Italy) *Extraction of yeast ubiquitin ligases -protein substrates relations from the litterature*, Apr. – Jul. 2021, supervised by E. Becker
- M2 Internship: Sarah Guinchard, *Characterizing MADE2, an ancient miniature transposable element, in the human genome*, Apr. – Sept. 2021, supervised by C. Belleannée
- M2 Internship: Baptiste Ruiz, *Intégration de données hétérogènes pour la caractérisation de patientes atteintes du cancer de l'ovaire : Microbiotes, données cliniques et habitudes alimentaires*, Mar. – Aug. 2021, supervised by Y. Le Cunff
- M2 Internship: Kerian Thuillier, *Inferring boolean rules controling hybrid models inspired by systems biology*, Feb. – Jul. 2021, supervised by A. Siegel
- M2 Internship: Leo Maury, "Machine learning applied to characterize cell death in liver", Apr.-Jun. 2021, co-supervised by Y. Le Cunff and J. Le Seyec (IRSET)
- M1 Research Project: Malo Revel, Luca Paparazzo, *Learning substitutable languages*, Sept. 2020 – Jun. 2021, supervised by F. Coste.

11.2.5 Juries

- Member of PhD thesis juries (9):
  - N. Guillaudeux, Université de Rennes 1 [C. Belleannée, S. Blanquart, O. Dameron]
  - C. Roussel, Ecole Normale Supérieure Paris [F. Coste, president]
  - N. Romashchenko, Université de Montpellier, déc. 2021 [F. Coste]
  - H. Talibart, Université de Rennes 1 [F. Coste]
  - M. Balluet, Université de Rennes 1 [O. Dameron, president]
  - M. Conan, Université de Rennes 1 [A. Siegel]
  - A. Weber, Sorbonne Université [A. Siegel]
  - A. Desoeuvres, Univ. Montpellier [A. Siegel]
  - V. Mataigne, Univ Rennes [A. Siegel].

11.3 Popularization

11.3.1 Articles and contents

- **Science en Cour[t]s**[15] Many of our current and former PhD students (N. Guillaudeux, O. Dennler, M. Louarn, M.Wéry, L. Bourneuf, H. Talibart, A. Antoine-Lorquin, C. Bettembourg, J. Coquet, V. Delannée, G. Garet, S. Prigent) have been heavily involved in organization of a local Popularization Festival where PhD. students explain their thesis via short movies. The movies are presented to a professional jury composed of artists and scientists, and of high-school students. Previous years films can be viewed on the festival website.

- **Les décodeuses du numérique**. Co-supervision of a comic book gathering the portraits of 12 female computer scientists. The book was sent to all French high schools and is freely available on line (more than 20,000 views in monthes and 5,000 downloads since september 2021). [16] [A. Siegel]

11.3.2 Education

- **General introduction to bioinformatics and presentation of the bioinformatics-related careers**, Lycée du Léon, Landivisiau December 10 2021 (postponed because of covid-related restrictions) [O. Dameron]

- **ESIR (engineer school, Rennes)** table-ronde sur la parité et les métiers du numérique au féminin [A. Siegel]

- **J'peux pas j'ai informatique - Prof** Since 2018, the Commission Égalité Femmes-Hommes has been hosting more than 150 fifth-grade students every year, during a "J'peux pas j'ai informatique" day to raise awareness of the very wide diversity of digital sciences. In 2021, in partnership with the rectorat and the association Femmes & Sciences, IRISA and Inria Rennes Bretagne Atlantique have also offered this training to teachers, so that each and every one of them can make the training and the workshops their own and duplicate them within their own school. [A. Siegel]

- **Rencontre des jeunes mathématiciennes et informaticiennes**, Rennes. [A. Siegel]

- **LCodent L Creent** the program introduces high school girls to programming during workshops led by computer science female phD students. The approach, based on creativity, allows the college girls to appropriate the concepts necessary for the realization of computer programs in 8 sessions carried out in high schools. [C. Juigné]

11.3.3 Interventions

- **Présentation CNU section 27**, Journée doctorants D3, juin 2021 [F. Coste]

- **Journée des 20 ans de la Mission pour la place des femmes du CNRS**. *le Rôle des référentes et référents parité d'un laboratoire* [A. Siegel]

- **Radio France International, Emission "autour de la question"**. *Quelles perspectives les nouveaux métiers du numérique offrent-ils aux femmes?* [17] [A. Siegel]

[15] sciences-en-courts.fr
[16] www.ins2i.cnrs.fr/fr/les-decodeuses-du-numerique
[17] podcast

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.

[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.

[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.

[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.

[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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

12.2 Publications of the year

International journals

[11] A. Antoine-Lorquin, P. Arensburger, A. Arnaoty, S. Asgari, M. Batailler, L. Beauclair, C. Belleannée, N. Buisine, V. Coustham, S. Guyetant, L. Helou, T. Lecomte, B. Pitard, I. Stévant and Y. Bigot. 'Two repeated motifs enriched within some enhancers and origins of replication are bound by SETMAR

isoforms in human colon cells'. In: *Genomics* 113.3 (2021), pp. 1589–1604. DOI: 10.1016/j.ygeno.2021.03.032. URL: https://hal.archives-ouvertes.fr/hal-03385429.

[12] S. Blanquart, M. Groussin, A. Le Roy, G. J. Szöllosi, E. Girard, B. Franzetti, M. Gouy and D. Madern. 'Resurrection of Ancestral Malate Dehydrogenases Reveals the Evolutionary History of Halobacterial Proteins : Deciphering Gene Trajectories and Changes in Biochemical Properties'. In: *Molecular Biology and Evolution* (2021), pp. 1–44. DOI: 10.1093/molbev/msab146. URL: https://hal-cnrs.archives-ouvertes.fr/hal-03231689.

[13] M. Conan, N. Theret, S. Langouet and A. Siegel. 'Constructing xenobiotic maps of metabolism to predict enzymes catalyzing metabolites capable of binding to DNA'. In: *BMC Bioinformatics* 22.1 (21st Sept. 2021), p. 450. DOI: 10.1186/s12859-021-04363-6. URL: https://www.hal.inserm.fr/inserm-03355908.

[14] J. Girard, G. Lanneau, L. Delage, C. Leroux, A. Belcour, J. Got, J. Collén, C. Boyen, A. Siegel, S. M. Dittami, C. Leblanc and G. V. Markov. 'Semi-Quantitative Targeted Gas Chromatography-Mass Spectrometry Profiling Supports a Late Side-Chain Reductase Cycloartenol-to-Cholesterol Biosynthesis Pathway in Brown Algae'. In: *Frontiers in Plant Science* 12 (2021), pp. 1–10. DOI: 10.3389/fpls.2021.648426. URL: https://hal.sorbonne-universite.fr/hal-03222505.

[15] V. Henry, I. Moszer, O. Dameron, L. Vila Xicota, B. Dubois, M.-C. Potier, M. Hofmann-Apitius and O. Colliot. 'Converting disease maps into heavyweight ontologies: general methodology and application to Alzheimer's disease'. In: *Database - The journal of Biological Databases and Curation* (16th Feb. 2021), pp. 1–33. DOI: 10.1093/database/baab004. URL: https://hal.archives-ouvertes.fr/hal-03144306.

[16] E. Karimi, E. Geslain, A. Belcour, C. Frioux, M. Aite, A. Siegel, E. Corre and S. M. Dittami. 'Robustness analysis of metabolic predictions in algal microbial communities based on different annotation pipelines'. In: *PeerJ* 9 (6th May 2021), pp. 1–24. DOI: 10.7717/peerj.11344. URL: https://hal.sorbonne-universite.fr/hal-03223662.

[17] M. Melkonian, C. Juigné, O. Dameron, G. Rabut and E. Becker. 'Towards a reproducible interactome: semantic-based detection of redundancies to unify protein-protein interaction databases'. In: *Bioinformatics* (2022). DOI: 10.1093/bioinformatics/btac013. URL: https://hal.archives-ouvertes.fr/hal-03522989.

[18] F. Moreews, H. Simon, A. Siegel, F. Gondret and E. Becker. 'PAX2GRAPHML: a Python library for large-scale regulation network analysis using BIOPAX'. In: *Bioinformatics* 37.24 (2021), pp. 4889–4891. DOI: 10.1093/bioinformatics/btab441. URL: https://hal.archives-ouvertes.fr/hal-03265223.

[19] H. Talibart and F. Coste. 'PPalign: optimal alignment of Potts models representing proteins with direct coupling information'. In: *BMC Bioinformatics* 22.1 (Dec. 2021), pp. 1–22. DOI: 10.1186/s12859-021-04222-4. URL: https://hal.inria.fr/hal-03264248.

[20] Q. Xing, G. Bi, M. Cao, A. Belcour, M. Aite and Y. Mao. 'Comparative Transcriptome Analysis Provides Insights into Response of Ulva compressa to Fluctuating Salinity Conditions'. In: *Journal of Phycology* 57.4 (Aug. 2021), pp. 1295–1308. DOI: 10.1111/jpy.13167. URL: https://hal.archives-ouvertes.fr/hal-03334031.

International peer-reviewed conferences

[21] M. Bouguéon, P. Boutillier, J. Feret, O. Hazard and N. Theret. 'A Kappa model for hepatic stellate cells activation by TGFB1'. In: CMSB 2021 - 19th International Conference on Computational Methods in Systems Biology. Bordeaux / Virtual, France, 22nd Sept. 2021. URL: https://hal.inria.fr/hal-03545256.

[22] K. Thuillier, C. Baroukh, A. Bockmayr, L. Cottret, L. Paulevé and A. Siegel. 'Learning Boolean controls in regulated metabolic networks: a case-study'. In: CMSB 2021 - 19th International Conference on Computational Methods in Systems Biology. Bordeaux, France: Springer, 2021. URL: https://hal.archives-ouvertes.fr/hal-03207589.

Conferences without proceedings

[23] J. Bakalara, T. Guyet, O. Dameron, A. Happe and E. Oger. 'An extension of chronicles temporal model with taxonomies -Application to epidemiological studies'. In: HEALTHINF 2021 - 14th International Conference on Health Informatics. online, France, 11th Feb. 2021, pp. 1–10. URL: https://hal.archives-ouvertes.fr/hal-03096846.

[24] C. Frioux, A. Belcour, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Metabolic complementarity applied to the screening of microbiota and the identification of key species'. In: JOBIM 2021 - Journées Ouvertes en Biologie, Informatique et Mathématiques. Paris, France, 6th July 2021. URL: https://hal.inria.fr/hal-03440232.

[25] C. Frioux, A. Belcour, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Metabolic complementarity applied to the screening of microbiota and the identification of key species'. In: MPA 2021 - 8th Metabolic Pathway Analysis. Knoxville, TN, United States, 2nd Aug. 2021, pp. 1–27. URL: https://hal.inria.fr/hal-03440221.

[26] C. Frioux, A. Belcour, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Metabolic complementarity applied to the screening of microbiota and the identification of key species'. In: CMSB 2021 - 19th International Conference on Computational Methods in Systems Biology. Bordeaux, France, 22nd Sept. 2021, pp. 1–27. URL: https://hal.inria.fr/hal-03440212.

[27] T. Guyet, T. Allard, J. Bakalara and O. Dameron. 'An open generator of synthetic administrative healthcare databases'. In: *Actes de l'atelier Intelligence Artificielle et Santé (IAS)*. IAS 2021 - Atelier Intelligence Artificielle et Santé. Bordeaux (virtuel), France, 29th June 2021, pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-03326618.

[28] F. Ibrahim, J. Got, A. Siegel, E. Forano and R. Munoz Tamayo. 'Genome-scale network reconstruction of the predominant cellulolytic rumen bacterium Fibrobacter succinogenes S85'. In: 12. International Symposium on Gut Microbiology. En ligne, France, 2021. URL: https://hal.inrae.fr/hal-03402155.

Scientific book chapters

[29] 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*. 2021, pp. 1–76. URL: https://hal.inria.fr/hal-03388100.

Doctoral dissertations and habilitation theses

[30] M. Conan. 'Predictive approach to asses the genotoxicity of environmental contaminants'. Université Rennes 1, 23rd Mar. 2021. URL: https://tel.archives-ouvertes.fr/tel-03334212.

[31] N. Guillaudeux. 'To compare gene structures for prediction of alternative coding transcripts in human,mouse and dog'. Université Rennes 1, 16th Dec. 2021. URL: https://tel.archives-ouvertes.fr/tel-03540882.

[32] H. Talibart. 'Comparison of homologous protein sequences using direct coupling information by pairwise Potts model alignments'. Université Rennes 1, 24th Feb. 2021. URL: https://tel.archives-ouvertes.fr/tel-03376771.

Other scientific publications

[33] B. Blanc. 'Détection de recombinaisons génomiques et protéomiques homologues par alignement multiple local et partiel'. Rennes 1, 15th June 2021. URL: https://hal.inria.fr/hal-03524403.

[34] M. Bouguéon, P. Boutillier, J. Feret, O. Hazard and N. Theret. 'A Kappa model for hepatic stellate cells activation by TGFB1'. In: CompSysBio 2021 - Advanced Lecture Course on Computational Systems Biology. Aussois, France, 14th Nov. 2021. URL: https://hal.inria.fr/hal-03545135.

[35] O. Dennler, S. Blanquart, F. Coste, C. Belleannée and N. Theret. 'Phylogenetic Functional Module Characterization of the ADAMTS / ADAMTS like Protein Family'. In: WABI 2021 - Workshop on Algorithms in Bioinformatics. Chicago (Online), United States, 2nd Aug. 2021. URL: https://hal.archives-ouvertes.fr/hal-03543214.

[36] O. Dennler, S. Blanquart, F. Coste, C. Belleannée and N. Theret. 'Phylogenetic Functional Module Characterization of the ADAMTS / ADAMTS like Protein Family'. In: JOBIM : Journées Ouvertes en Biologie, Informatique & Mathématiques. Paris, France, 6th July 2021. URL: https://hal.archives-ouvertes.fr/hal-03543182.

[37] C. Frioux, A. Belcour, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. 'Assessment of metabolic complementarity in large-scale microbiotas for the identification of key species'. In: IHMC 2021 - 8th International Human Microbiome Consortium Congress. Barcelone, Spain, 27th June 2021, p. 1. URL: https://hal.archives-ouvertes.fr/hal-03438983.

12.3 Cited publications

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] V. Kmetzsch, V. Anquetil, D. Saracino, D. Rinaldi, A. Camuzat, T. Gareau, L. Jornea, S. Forlani, P. Couratier, D. Wallon, F. Pasquier, N. Robil, P. De La Grange, I. Moszer, I. Le Ber, O. Colliot and E. Becker. 'Plasma microRNA signature in presymptomatic and symptomatic subjects with C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis'. In: *Journal of Neurology, Neurosurgery and Psychiatry* 92.5 (Nov. 2020), pp. 485–493. DOI: 10.1136/jnnp-2020-324647. URL: https://hal.inria.fr/hal-03046771.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[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.

[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.

[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.

[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.