RESEARCH CENTRE

Rennes - Bretagne Atlantique

IN PARTNERSHIP WITH:

CNRS, Université Rennes 1

2020 ACTIVITY REPORT

Project-Team DYLISS

Dynamics, Logics and Inference for biological Systems and Sequences

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

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Biology

Contents

Project-Team DYLISS

Creation of the Team: 2012 January 01, updated into 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, from Sep 2020]
- 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, from Dec 2020]

Post-Doctoral Fellow

- Celia Biane-Fourati [Inria, until Mar 2020]

PhD Students

- Meziane Aite [Insiliance SAS Paris, CIFRE, from Nov 2020]
- Arnaud Belcour [Inria]
- Matthieu Bougueon [INSERM, from Oct 2020]
- Nicolas Buton [Univ de Rennes I, from Oct 2020]
- Mael Conan [Univ de Rennes I]
- Olivier Dennler [INSERM]
- Nicolas Guillaudeux [Univ de Rennes I]
- Camille Juigne [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, from Dec 2020 (was engineer from Feb to Nov)]
- Virgilio Kmetzsch Rosa E Silva [Inria]
- Marine Louarn [Univ de Rennes I, until Aug 2020]
- Hugo Talibart [Univ de Rennes I]
- Pierre Vignet [Univ de Rennes I, until Nov 2020]
- Meline Wery [Univ de Rennes I, until Sep 2020]

Technical Staff

- Xavier Garnier [Inria, Engineer, until Sep 2020]
- Jeanne Got [CNRS, Engineer]
- Leo Milhade [CNRS, Engineer, from Nov 2020]
- Corentin Raphalen [CNRS, Engineer, from Nov 2020]

Interns and Apprentices

- Eve Barre [Univ de Rennes I, from May 2020 until Jul 2020]
- Konogan Bourhy [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, from Feb 2020 until Jul 2020]
- Nicolas Buton [Univ de Rennes I, from Feb 2020 until Aug 2020]
- Adriana Concha Sepulveda [Inria, until Mar 2020]
- Quentin Delhon-Bugard [Univ de Rennes I, from Feb 2020 until Jul 2020]
- Pierre Gueracher [Inria, from Feb 2020 until Jun 2020]
- Maxime Leger [Inria, from May 2020 until Jul 2020]
- Malo Revel [Inria, from May 2020 until Jun 2020]
- Baptiste Ruiz [Inria, from Mar 2020 until Jun 2020]
- Jean Trippier De Lagrange [Inria, from May 2020 until Jul 2020]

Administrative Assistant

- Marie Le Roic [Inria]

Visiting Scientist

- Oumarou Abdou Arbi [Université Dan Dicko Dankoulodo Maradi - Niger, from Feb 2020 until Mar 2020]

External Collaborators

- Sebastien Auber [INSERM, until Jun 2020]
- Yann Le Cunff [Univ de Rennes I, until Nov 2020]
- François Moreews [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]
- Corentin Raphalen [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, from Feb 2020 until Jul 2020]
- Denis Tagu [Institut national de recherche pour l'agriculture, l'alimentation et l'environnement]
- 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 genomics 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. Indded, "omics" technologies have allowed the number of measured compounds in a systems 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 position of developing experimental strategies to progressively shrink the space of hypotheses and gain in 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 developped in the team for the last five years: Askomics, AuReMe, FinGoc, Caspo, Cadbiom, Logol, 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, dynamical models) which are compatible with both 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 https://www.genouest.org/.

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 objectives 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) [47], 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 in non-model species, such as enzymes and isoforms functions in 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, abstract gene models) enabling us to represent sequence conservation, sets of short and degenerated patterns and crossing or distant dependencies. This requires both to determine classes of formal syntactic models allowing to handle 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 [55]. In our opinion, life sciences cumulates 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**[46]. To explore this research field, we use techniques from symbolic data mining (Semantic Web technologies, symbolic clustering, constraint satisfaction and grammatical modelling) to take into account those life science features in the analysis of biological data.

3.2.1 Research topics

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

Scalability of semantic web queries. A bottleneck in data querying is given by the performance of federated SPARQL queries, which must be improved by several orders of magnitude to allow current massive data to be analyzed. In this direction, our research program focuses on the combination of *linked data fragments* [56], 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 functional static map of biological models that can

be analyzed and/or compressed. This requires to linking 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 to (1) take advantage of the information readily available in the LOD cloud for analyzing there own data and (2) contribute back to the linked data by representing their data and the associated metadata in the proper format as well as by linking them to other resources. An originality is the graphical interface that allows any dataset to be integrated in a local RDF datawarehouse and SPARQL query to be built transparently and iteratively by a non-expert user.

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 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 researches in bioinformatics in relation with metabolic processes are driven by the understanding of 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 taking a sample of functional sequences as input to infer a grammar representing their key syntactical characteristics, including dependencies between residues. Our first goal is to enable an automatic semi-supervised refinement of enzymes classification [6] by combining the Protomata-Learner [40] framework - which captures local dependencies - with formal concept analysis. More challenging, we are exploring the learn of grammars representing long-distance dependencies such as those exhibited by contacts of amino-acids that are far in the sequence but close in the 3D protein folding.

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

Non-model organisms are associated with often incomplete and poorly annotated sequences, leading to draft networks of their metabolism which largely suffer from incompleteness. In former studies, the team has developed several methods to improve the quality of eukaryotes metabolic networks, by solving several variants of the so-called *Metabolic Network gap-filling problem* with logical programming

[1] https://github.com/askomics/askomics
[2] http://biowic.inria.fr/
[3] http://github.com/aluriak/powergrasp
[4] http://bioasp.github.io/iggy/

approaches [10, 9]. The main drawback of these approaches is that they cannot scale to the reconstruction and comparison of families of metabolic networks. Our main objective is therefore to develop new tools for the comparison of species strains at the metabolic level.

Consortium level: exploring the diversity of community consortia A new emerging field is system ecology, which aims at building predictive models of species interactions within an ecosystem for deciphering cooperative and competitive relationships between species [45]. 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. A second very challenging focus is the prediction of transporters families by obtaining refined characterization of transporters, which are quite unexplored apart from specific databases [53].

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 sequences alignments (said partial and local), learn automata and search for new family members in sequence databases. By enabling to model 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 [40] 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 [26].

AuReMe workspace is designed for tractable reconstruction of metabolic networks[6] 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-classical 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: `http://aureme.genouest.org/wiki.html`). It also generated outputs to explore resulting networks with Askomics. It has been used for reconstructing metabolic networks of micro and macro-algae[51], extremophile bacteria[42] and communities of organisms[4].

Mpwt is a Python package for running Pathway Tools[7] on multiple genomes using multiprocessing. Pathway Tools is a comprehensive systems biology software system that is associated with the BioCyc database collection[8]. Pathway Tools is very used for reconstructing metabolic networks.

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 involves 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. A particular attention is paid to the characterization of products of genes (such as isoform) and of perturbations to identify discriminant signature of pathologies.

[5] `http://protomata-learner.genouest.org`
[6] `http://aureme.genouest.org/`
[7] `http://bioinformatics.ai.sri.com/ptools/`
[8] `https://biocyc.org/`

3.4.1 Research topics

Genomic level: characterizing gene structure with grammatical languages and conservation information The subject here is to accurately represent gene structure, including intron/exon structure, for predicting the products of genes, such as isoform transcripts, and comparing the expression potential of a eukaryotic gene according to its context (e.g. tissue) or according to the species. Our approach consists in designing grammatical and comparative-genomics based models for gene structures able to detect heterogeneous functional sites (splicing sites, regulatory binding sites...), functional regions (exons, promotors...) and global constraints (translation into proteins) [36]. Accurate gene models are defined by identifying general constraints shaping gene families and their structures conserved over evolution. Syntactic elements controlling gene expression (transcription factor binding sites controlling transcription; enhancers and 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[54].

System level: extracting causal signatures of complex phenotypes with systems biology frameworks The main challenge we address is to set up a generic formalism to model inter-layer interactions in large-scale biological networks. To that goal, we have developed several types of abstractions: multi-experiments framework to learn and control signaling networks [11], multi-layer reactions in interaction graphs [37], and multi-layer information in large-scale Petri nets [33]. 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 modelling and matching[9]. 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)[10]. The software handles inherent experimental noise by enumerating all different logical networks which are compatible with a set of experimental observations [11]. The main advantage is that it enables a complete study of logical network without requiring any linear constraint programs.

Cadbiom package aims at building and analyzing the asynchronous dynamics of enriched logical networks[11]. It is based on Guarded transition semantic and allows synchronization events to be investigated in large-scale biological networks [33].

For instance, it was designed to allow controler of phenotypes in large-scale knowledge databases (PID) to be curated and analyzed [5].

4 Application domains

In terms of transfer and societal impact, we consider that our role is to develop fruitful collaborations with laboratories of biology 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.

[9] http://logol.genouest.org/
[10] http://bioasp.github.io/caspo/
[11] http://cadbiom.genouest.org/

Marine Biology: seaweed enzymes and metabolism & sea-urchin cell-cycle. Our main field of study is **marine biology**, as it is a transversal field covering challenges in integrative biology, dynamical systems and sequence analysis. 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 metabolism *E. Siliculosus* to be deciphered [51, 43]. The study of the *HAD* superfamily of proteins thanks to partial local alignments, produced by Protomata tools, allows sub-families to be deciphered and classified, and 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. As a further study, we reconstructed the metabolic network of a symbiot bacterium *Ca. P. ectocarpi* [44] 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 suggest that the bacterium *Ca. P. ectocarpi* is able to provide both β-alanine and vitamin B5 to the seaweed via the phosphopantothenate biosynthesis pathway [52].

Micro-biology: elucidating the functioning of extremophile consortiums of bacteria. In this application field, our main issue is the understanding of bacteria living in extreme environments, mainly in collaboration with the group of bioinformatics at Universidad de Chile (co-funded by the Center of Mathematical Modeling, the Center of Regulation Genomics and Inria-Chile). In order to elucidate the main characteristics of these bacteria, our integrative methods were developed to identify the main groups of regulators for their specific response in their living environment. The integrative biology tools Meneco, Lombarde and Shogen have been designed in this context. In particular, genome-scale metabolic network been recently reconstructed and studied with the Meneco and Shogen approaches, especially on bacteria involved in biomining processes [38] and in Salmon pathogenicity [42].

Agriculture and environmental sciences: upstream controllers of cow, pork and pea-aphid metabolism and regulation. In this application field, our goal is to propose methods to identify regulators of very complex phenotypes related to environmental issues. Our work on the identification of upstream regulators within large-scale knowledge databases (prototype KeyRegulatorFinder) [37] and on semantic-based analysis of metabolic networks [35] was very valuable for interpreting differences of gene expression in pork meat [49] and figure out the main gene-regulators of the response of porks to several diets [48]. In addition, constraints-based programming also allows us to decipher regulators of reproduction for the pea aphid, an insect that is a pest on plants. In terms of biological output of the network studies on the pea aphid microRNAs, we have identified one new microRNA (apmir-3019, not present in any known species other than the pea aphid) who has more than 900 putative mRNA targets.

Health: deciphering pathways involved in the TGF-β signalling network. TGF-β is a multifunctional cytokine that regulates mammalian development, differentiation, and homeostasis with both benefical anti-tumor effect [39] and pro-tumor effect [50]. Deciphering protumor versus antitumor signaling 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 interleukine signaling pathways [33, 41].

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, IPL NeuroMarkers and Insiliance, 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

6.1 Members

Yann Le Cunff (associate professor, Univ. Rennes 1) joined the team.

6.2 Collaborations

The Inria International Lab SymBioDiversity with Universidad de Chile has been accepted (cf. Section 10). The project aims at developing methods combining data-mining, reasoning and mathematical modeling to efficiently analyze massive data about microbial biodiversity in extreme environment and identify families of species which characterize environmental niches.

We initiated a new collaboration with Insiliance as a follow-up of the collaboration initiated with Theranexus in 2019. This led to Méziane Aite's CIFRE PhD that started in november. The project aims at identifying new combinations of molecules in a context of drug repositionning.

6.3 Dissemination

AskOmics[12] has been integrated into the French Bioinformatics Institue's (IFB) catalogue of applications and can now be deployed in all the IFB clouds[13].

7 New software and platforms

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: 2020: (1) release 4.4.1, (2) extensive documentation, (3) creation of tutorials, (4) complete redesign of the website, (4) support for disjunction, (5) registration to the Institut Français de Bioinformatique (IFB) catalogue (https://biosphere.france-bioinformatique.fr/catalogue/appliance/166/), (5) support for LDAP authentication

URL: https://github.com/askomics/askomics

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

Contacts: Olivier Dameron, Anthony Bretaudeau, Fabrice Legeai

Partners: Université de Rennes 1, CNRS, INRA

[12] https://askomics.org/
[13] https://biosphere.france-bioinformatique.fr/catalogue/appliance/166/

7.1.2 AuReMe

Name: Automatic Reconstruction of Metabolic networks

Keywords: Workflow, Bioinformatics, Metabolic networks, Omic data, Toolbox, Data management

Functional Description: AuReMe enables the reconstruction of metabolic networks from different sources based on sequence annotation, orthology, gap-filling and manual curation. The metabolic network is exported as a local wiki allowing to trace back all the steps and sources of the reconstruction. It is highly relevant for the study of non-model organisms, or the comparison of metabolic networks for different strains or a single organism.

Five modules are composing AuReMe: 1) The Model-management PADmet module allows manipulating and traceing all metabolic data via a local database. 2) The meneco python package allows the gaps of a metabolic network to be filled by using a topological approach that implements a logical programming approach to solve a combinatorial problem 3) The shogen python package allows genome and metabolic network to be aligned in order to identify genome units which contain a large density of genes coding for enzymes, it also implements a logical programming approach. 4) The manual curation assistance PADmet module allows the reported metabolic networks and their metadata to be curated. 5) The Wiki-export PADmet module enables the export of the metabolic network and its functional genomic unit as a local wiki platform allowing a user-friendly investigation.

Release Contributions: - Reworking padmet and padmet-utils to allow full-python workflow in the future - Adding new script padmet-utils/exploration/prot2genome with exonerate - Fixing minor errors

News of the Year: (1) Pantograph replaced by OrthoFinder (2) Create a readthedocs for AuReMe, padmet and padmet-utils (3) Reworking padmet and padmet-utils to allow full python workflow (4) Adding new script padmet-utils/exploration/prot2genome with exonerate (5) Modify template data structure (6) Fixing errors

URL: http://aureme.genouest.org/

Publication: hal-01807842

Authors: Jeanne Cambefort, Anne Siegel, Alejandro Maass, Marie Chevallier, Meziane Aite, Guillaume Collet, Nicolas Loira, Sylvain Prigent

Contacts: Jeanne Cambefort, Meziane Aite, Marie Chevallier

Participants: Marie Chevallier, Meziane Aite, Guillaume Collet, Nicolas Loira, Sylvain Prigent, Jeanne Cambefort, Anne Siegel, Alejandro Maass

Partner: University of Chile

7.1.3 Metage2Metabo

Keywords: Metabolic networks, Microbiota, Metagenomics, Workflow

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: First release.

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

Publication: hal-02395024

Authors: Clémence Frioux, Arnaud Belcour, Anne Siegel

Contacts: Arnaud Belcour, Clémence Frioux, Anne Siegel

Participants: Clémence Frioux, Arnaud Belcour, Anne Siegel

7.1.4 Pathmodel

Keywords: ASP - Answer Set Programming, Metabolic networks, Metabolic Pathway Drift, Bioinformatics, Systems Biology, Metabolomics

Scientific Description: This tool is a prototype of the Metabolic Pathway Drift concept. This concept states that metabolic pathways undergo substantial turnover. The reactions involved in a pathway can change between species (change in reaction order or replacement of an enzyme by another one). Another goal of this tool is to linked genomics and metabolomics data. To implement this concept, Pathmodel uses the Answer Set Programming language. The input are the reactants and products involved in the pathway, known reactions occurring between these molecules, known m/z ratio, known domains shard by these molecules, an initial molecule and a goal molecule. Using these data, Pathmodel will infer reactions between molecules to reach the goal molecule using the known reactions. The result consist of potential alternative pathways for the studied organism.

Functional Description: A metabolic pathway is a series of biochemical reactions. These reactions modify metabolites in order to synthesize a new metabolite or to produce energy. One difficulty when dealing with pathways in non-model organism is their incomplete conservation during evolution. To deal with this problem, we developed a prototype inferring new biochemical reactions using reactions and metabolites from known metabolic pathways and metabolomics data. This method produce alternative pathway that could occurred in the species of interest.

Release Contributions: Fix an issue with test data.

News of the Year: (1) Add a container in Singularity Hub (https://singularity-hub.org/collections/3758). (2) Rewrite data files (sterol and MAA). (3) Add creation of pictures of new molecules from MZ. (4) Add new output files to ease understanding of PathModel output. (5) Rewrite the Readme.

URL: `https://github.com/pathmodel`

Publication: hal-01943880

Contacts: Arnaud Belcour, Gabriel Markov, Anne Siegel

Participants: Arnaud Belcour, Jacques Nicolas, Gabriel Markov, Anne Siegel

Partner: Station Biologique de Roscoff

7.1.5 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: - Comprehensive command line to run the calculations and analyze the generated results. - Module designed to produce models through the interpretation of various databases or ontologies, formalized according to the BioPAX standard. - Update of the site and the documentation.

We recently developed a framework that integrates an updated version of the CADBIOM core software and visualization tools. We provided a command line interface allowing users to translate the interactions between biomolecules described in data sources in BioPAX format into the formalism based on the guarded transitions used by CADBIOM. The command line also makes it easy to search for scenarios based on the constraints of a model, to compare scenarios with each other and to visualize interaction graphs facilitating the biologist's expertise (Vignet et al 2019 JOBIM)

URL: http://cadbiom.genouest.org

Authors: Michel Le Borgne, Geoffroy Andrieux, Nolwenn Le Meur, Nathalie Theret, Pierre Vignet

Contact: Anne Siegel

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

7.1.6 PPsuite

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

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

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

Publications: hal-03134517, hal-02862213, hal-02402646

Authors: Hugo Talibart, François Coste

Contacts: Hugo Talibart, François Coste

7.1.7 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. Supported by 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.

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

Contacts: François Moreews, Emmanuelle Becker, Anne Siegel

Partner: INRAE

8 New results

8.1 Scalable methods to query data heterogeneity

Participants Olivier Dameron, Marine Louarn, Méline Wéry.

Increasing life science resources re-usability using Semantic Web technologies [*M. Louarn*] [29].

- Our work was focused on assessing to what extent Semantic Web technologies also facilitate reproducibility and reuse of life sciences studies involving pipelines that compute associations between entities according to intermediary relations and dependencies.

Multi-omic data integration for identifying causal pathologic signatures [*M. Wéry*] [30].

- We designed a transomic model in order to structure all the omic data, using semantic Web (RDF). This model is based on a patient-centric approach. SPARQL queries allow the identification of expression Individually-Consistent Trait Loci (eICTLs).

Converting disease maps into heavyweight ontologies [*O. Dameron*] [23].

- 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.

8.2 Metabolism: from protein sequences to systems ecology

Participants Méziane Aite, Arnaud Belcour, Samuel Blanquart, Mael Conan, François Coste, Clémence Frioux, Jeanne Got, Anne Siegel, Hugo Talibart, Nathalie Théret.

Modelling proteins by partial local multiple sequence alignment [*F. Coste*] [26]

- In collaboration with the team of E. Guedon (STLO, Inrae), we applied paloma (partial local multiple sequence alignment tool from our Protomata Suite) to study shared conservations in the core proteome of extracellular vesicles produced by human and animal *Staphylococcus aureus* strains and better understand how proteins packed in extracellular vesicles produced by S. aureus can mediate the pathogenesis of infection [26].

Modelling proteins with crossing dependencies [*F. Coste, H. Talibart*] [27]

- 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 [27].

Large-scale eukaryotic metabolic network reconstruction [*A. Siegel, M. Aite, A. Belcour, J. Got, N. Théret, M. Conan*] [25, 18, 14, 16, 21, 22].

- *Genome studies based on large-scale metabolic network reconstruction*: We participated to the reconstruction of the metabolic networks of two Microbacterium sp., CGR1 and CGR2, previously isolated from physicochemically contrasting soil sites in the Atacama Desert, showing significant differences in the connectivity of specific metabolites related to pH tolerance and co2 production [25] . We also participated to the deciphering of the genome of Ectocarpus subulatus, A highly stress-tolerant brown alga [18]. Finally, we contributed to the study of cyanobacteria genomes by providing highlights on the metabolic networks of several marine strains, which were integrated in the Cyanorak database [21] and used to analyse synergic effects in the marine synechococcus strain WH7803 [22].

- *Metabolic pathway inference from metabolomic data and appplication to metabolic pathway drift* Inferring genome-scale metabolic networks in emerging model organisms is challenged by incomplete biochemical knowledge and partial conservation of biochemical pathways during evolution. Using an integrative approach combining genomic and metabolomic data in the red algal model Chondrus crispus, we show that, even metabolic pathways considered as conserved, like sterols or mycosporine-like amino acid synthesis pathways, undergo substantial turnover. We present a proof of concept with a methodological approach to formalize the logical reasoning necessary to infer reactions and molecular structures, abstracting molecular transformations based on previous biochemical knowledge. [14]

- *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 [16]

Systems ecology: design of microbial consortia and study of host-microbial interactions [*A. Belcour, S. Blanquart, J. Got, M. Aite, A. Siegel*] [20, 13, 15, 17].

- We participated to the application of our methods to algal-microbial consortia, with good preliminary results, and presented them as an invited conference [13, 15, 17].

- *Unifying the study of metabolic networks and host-microbial interactions as a single combinatorial optimisation problem* Systems modelled in the context of molecular and cellular biology are difficult to represent with a single calibrated numerical model. For more complex organisms or non-cultured microorganisms such as those evidenced in microbiomes with metagenomic techniques, flux optimisation techniques may not be applicable to elucidate systems functioning. In this context, we describe how automatic reasoning allows relevant features of an unconventional biological system to be identified despite a lack of data by comparing steady-states of Boolean abstractions of metabolic models and illustrate their complementarity via applications to the metabolic analysis of macro-algae, allowing for the generation of several instances of gap-filling problems [20].

- *Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species* We developed a method called Metage2Metabo (M2M) that simulates the metabolism of the gut microbiota and describes the metabolic relationships between the different microbes. Metage2Metabo analyses the roles of the metabolic genes of a large number of microbe species to establish how they complement each other metabolically. Then, it can calculate the minimum number of species needed to perform a metabolic role of interest within that microbiota, and which key species are needed to perform that role. In the context of the gut microbiota, the predictions of Metage2Metabo could shed lights on the interactions between the host and the microbes and contribute to a better understanding of microbe environments. [13]

- *Metabolic Complementarity Between a Brown Alga and Associated Cultivable Bacteria Provide Indications of Beneficial Interactions* We used the methods underlying the Metage2Metabo approach to start disentangling the complex interactions that may occur between Ectocarpus siliculosus, a genetic and genomic model for brown algae algal, and 10 alga-associated bacteria. The predicted metabolic capacities were used to identify metabolic complementarities between the algal host and the bacteria and predict consortia consisting of a subset of these ten bacteria that would best complement the algal metabolism. Finally, co-culture experiments were set up, demonstrating a significant increase in algal growth in cultures inoculated with the selected consortia [15].

- To explore the pathobiome concept, which represents the interaction between the pathogen, the host plant and the associated biotic microbial community, we contributed to the deciphering of how the soil microbial environment may influence the functions of a pathogen and its pathogenesis, and the molecular response of the plant to the infection, using Brassica napus and Plasmodiophora brassicae. The soil microbial diversity levels had an impact on disease development (symptom levels and pathogen quantity). The functional analysis of gene expressions allowed the identification of pathogen and plant host functions potentially involved in the change of plant disease level [17]

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

Participants Emmanuelle Becker, Olivier Dameron, Maxime Folschette, Virgilio Kmetzsch, Nathalie Théret, Pierre Vignet, Méline Wéry.

Creation of predictive functional signaling networks [*M. Folschette, N. Théret*] [19, 28].

- Integrating genome-wide gene expression patient profiles with regulatory knowledge is a challenging task because of the inherent heterogeneity, noise and incompleteness of biological data. We proposed an automatic pipeline to extract automatically regulatory knowledge from pathway databases and generate novel computational predictions related to the state of expression or activity of biological molecules. We applied it in the context of hepatocellular carcinoma (HCC) progression, and evaluated the precision and the stability of these computational predictions. Our computational model predicted the shifts of expression of 146 initially non-observed biological components. Our predictions were validated at 88% using a larger experimental dataset and cross-validation techniques. [19]

- **Integrative models for TGF-beta signaling and extracellular matrix** The extracellular matrix (ECM) is the most important regulator of cell-cell communication within tissues. In this context, we reviewed how signaling networks associated with the polypeptide transforming growth factor TGF-bet are unique since their activation are controlled by ECM and TGF-β is a major regulator of ECM remodeling in return. [28]

Analysis of miRNA regulatory network to identify a key regulator of liver fibrosis [*N. Théret, O. Dameron*] [12].

- We developed an original approach to analyze the miRNA-dependent regulation of genes associated with the activation of hepatic stellate cells. Analyses of the regulatory network allowed us to identify TIMP3 as a potential regulator of fibrosis. We validated this prediction by using experimental methods.

Evidence of a microRNA signature for frontotemporal lobar degeneration and amyotrophic lateral sclerosis [*E. Becker, V. Kmetzsch*] [24].

- In the context of our participation to the IPL NeuroMarker, 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[14]. 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.

Identification of causal pathologic signature by multi-omic data integration [*M. Wéry*] [30].

- We designed a patient-centric approach based on expression Individually-Consistent Trait Loci (eICTLs). An eICTL is an association between a SNP and a gene where the presence of the SNP impacts the variation of the gene expression. Those elements provide a reduction of omics data dimension and show a more informative contribution than genomic data. By combining the dynamics of biological system with formal concept analysis, we were able to classify automatically the system's stable states. This classification enables the enrichment of biological signature, which characterised a phenotype. Moreover, it allowed the identification of a new hybrid phenotype.

9 Bilateral contracts and grants with industry

9.1 Bilateral contracts with industry

SANOFI: co-supervised PhD

Participants Emmanuelle Becker , Olivier Dameron , Anne Siegel , Méline Wéry.

This collaboration project is focused on the implementation of an integrative analysis framework based on semantic web technologies and reasoning in the framework of systemic lupus erythematosus pathology [30].

CIFRE co-supervised Grant: PhD. funding. 2017-2020

INSILIANCE: co-supervised PhD

Participants Méziane Aite , Olivier Dameron.

This collaboration 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**

10 Partnerships and cooperations

10.1 International initiatives

10.1.1 Inria International Labs

SymBioDiversity

Title: SymBioDiversity

Duration: 2020 - 2022

Coordinator: Anne Siegel

[14] https://institutducerveau-icm.org/fr/actualite/sla-dft-mutation-gene-c9orf72/

Partners:

- Computer Science Department, PLEIAD laboratory, Universidad de Chile (Chile)

Inria contact: Anne Siegel

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

10.2 International research visitors

Visits of international scientists Oumarou Abdou Arbi (Université Dan Dicko Dankoulodo Maradi, Niger) was scheduled to visit Dyliss in February and March 2020. His visit was canceled due to covid-related restrictions and the collaboration continued remotely.

10.3 European initiatives

Collaborations with major European organizations We initiated an informal collaboration with Lydie Lane and the Calipho group (creators of the neXtProt knowledge base[15]) at SIB Genève. We explored two topics. The first focused on determing whether neXtProt's public SPARQL endpoint can be used as a remote resource allowing AskOmics' users to extend their data with the neXtProt knowledge base and perform semantically-rich federated queries. The second stemmed from the observation that just like most life science resources exposed as SPARQL endpoints, neXtProt data structure is complex, which hampers their adoption by users. However, neXtProt benefits from a large base of example SPARQL queries designed for educational purposes. We explored whether this set of queries can be used to identify modules repeated in several queries, with the hypothesis that these modules provide an abstraction corresponding to a functional view on the neXtProt knowledge base. This work was supported by Maxime Léger's 3rd year BSc internship and his 1st year MSc internship.

10.4 National initiatives

IDEALG (ANR/PIA-Biotechnology and Bioresource)

Participants Méziane Aite , Arnaud Belcour , François Coste , Clémence Frioux , 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)[16]. 2012–21. Total grant: 11M€. Dyliss grant: 534k€.

[15] https://www.nextprot.org/
[16] https://idealg.u-bretagneloire.fr/

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 INRA and the research laboratories involved in the fields of physiology, nutrition, well-being/behaviour and infectiology/immunology. 2019–22.

10.4.1 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). 2017–20.

Askomics (ADT)

Participants Olivier Dameron , Xavier Garnier , Anne Siegel.

AskOmics[17] is a visual SPARQL query interface supporting both intuitive data integration and querying while avoiding the user to face most of the technical difficulties underlying RDF and SPARQL. The underlying motivation is that even though Linked (Open) Data now provide the infrastructure for accessing large corpora of data and knowledge, life science end-users seldom use them, nor contribute back their data to the LOD cloud by lack of technical expertise. AskOmics aims at bridging the gap between end users and the LOD cloud. 2018–2020.

10.5 Regional initiatives

Prolific

Participants Corentin Raphalen , Jeanne Got , Anne Siegel.

Food fermentations, including lactic and propionic fermentations (FDP), are one of the oldest methods of preserving perishable foodstuffs and remain one of the most durable processes (less use of the cold chain and preservatives). The activity of the bacteria during these fermentations radically transforms the raw material and generates compounds of interest whose action and benefit can go far beyond simple conservation. The objective of the PROLIFIC project is to study the potential of bacteria to provide

[17] https://askomics.org/

added health value. To achieve this, the PROLIFIC project was built in close collaboration between the industrialists grouped within BBA (Milk Valley) and 6 Breton and Loire academic partners, bringing complementary skills: microbiology, dairy technology, nutrition, biochemistry, immunology, physiology, neurology, bioinformatics, modelling. It is funded by an industrial and regional program. 2020–2022.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

Member of the organizing committees

- 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]
- Organisation of the monthly "Data and Knowledge management" department of Irisa [A. Siegel]
- Organisation of the "rencontres transdisciplinaires technologies et santé" (Day 1: ADN, Polymeres, et big data; Day 2: Interfaces cerveau-machine) https://technosante.sciencesconf.org [A. Siegel]
- Organisation de la journée Prospective en Science des Données, Intelligence Artificielle et Biologie https://www.france-bioinformatique.fr/prospective-en-science-des-donnees-intelligence-artificielle-et-biologie/ [A. Siegel]

11.1.2 Scientific events: selection

Chair of conference program committees

Member of the conference program committees

- BIOINFORMATICS [A. Siegel]
- CSBio [A. Siegel]
- IJCAI 2020 [A. Siegel]
- BBCC 2020 [E. Becker, O. Dameron]
- WCB@ICML 2020 [A. Siegel]
- IA et santé workshop, organized by AFIA and AIM [O. Dameron]
- Jobim 2020 [F. Coste, A. Siegel]

11.1.3 Journal

Member of the editorial boards

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

Reviewer - reviewing activities

- BMC Bioinformatics [A. Siegel]
- PLoS Computational Biology [A. Siegel]
- IEEE/ACM Transctions on Computational Biology [A. Siegel]
- Briefings in bioinformatics [O. Dameron]
- f1000 Research [O. Dameron]
- Journal of Biomedical Semantics [O. Dameron]
- Journal on Data Semantics [O. Dameron]

11.1.4 Invited talks

- Seminar at the "Microbiome Day", Rennes, *Using metabolic complementarity in microbiome to understand interactions among organisms* [A. Belcour]
- Seminar at the Computer Science Department of ENS Paris. *Des systèmes dynamiques à la biologie des systèmes pour des organismes non conventionnels: place du raisonnement symbolique.* [A. Siegel]
- Seminar at the Swiss Institute of Bioinformatics Geneva, team Calipho *AskOmics for integrating and querying life science data* [X. Garnier and O. Dameron]
- Webinar "Using ontologies to improve animal science research" INRAe *Bio-ontologies for complex life science data integration and reuse* [O. Dameron]

11.1.5 Scientific expertise

Recruitment Committee

- Associate professor "Biochimie et interactions moléculaires", Université de Lyon 1 - poste 4557 [E. Becker]
- Associate professor "Bioinformatique et génomique" Université de Marseille - poste 893 [E. Becker]
- Associate professor "Informatique" Université Paris-Saclay - poste 326 [O. Dameron]

National scientific boards

- Animation of the Systems Biology working group of national infrastructure GDR IM and GDR BIM `http://bioss-cnrs.fr` [A. Siegel].
- Board of directors of the French Society for biology of the extracellular matrix [N. Théret].
- ModCov19 coordination committee `https://modcov19.math.cnrs.fr` [A. Siegel]
- Scientific Board of the MathNum Department of INRAE [A. Siegel]

Project evaluation

- Banque Publique d'Investissement [A. Siegel]
- ANR Flash Covid [A. Siegel]

Local responsibilities

- Social committee of Univ. Rennes 1 [C. Belleannée]
- Emergency aid commission of Univ. Rennes 1 & Rennes 2 [C. Belleannée]
- Scientific committee of Univ. Rennes 1 school of medicine [O. Dameron].
- Member of the Inria Rennes center council [J. Got]
- Scientific Advisory Board of Biogenouest [N. Théret]
- Delegate to research integrity at the University of Rennes 1 [N. Théret]
- Inria Rennes PhD recruitment (CORDIs) [E. Becker]
- Head of the "Data and Knowledge Management" department (6 teams) of IRISA [A. Siegel]

11.1.6 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) [A. Siegel]

Local responsabilities

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

11.2 Teaching - Supervision - Juries

11.2.1 Teaching track responsibilities

- Coordination of the doctoral school "Biology and Health" of University of Bretagne Loire, Rennes Site 1 [N. Théret]
- Coordination of the master degree "Bioinformatics", Univ. Rennes 1 [E. Becker, O. Dameron]

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]
- "Complément informatique 1", Licence 1 informatique, Univ. Rennes 1 [O. Dameron]
- "Atelier bioinformatique", Licence 2 informatique, Univ. Rennes 1 [O. Dameron]
- "Intégration: Remise à niveau en informatique", Master 1 in bioinformatics, 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]

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: S. Blanquart, Juries of Master 2 in bioinformatics, 10h, Univ. Rennes 1, France
- Master: L. Bourneuf, Projet, 25h, M1 Santé Publique, France.
- Master: F. Coste, Supervised machine learning, 10h, M2 Science Informatique, Univ. Rennes, France
- Licence: O. Dameron, Biostatistiques, 12h, 1st year school of medicine, Univ. Rennes 1, 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", 12h, Licence 2 informatique, Univ. Rennes 1, France
- Licence: O. Dameron, "Programmation", 54h, Licence 3 miage, Univ. Rennes 1, France
- Master: O. Dameron, "Intégration: Remise à niveau en informatique", 18h Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, "Programmation impérative en Python", 98h Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, "Système informatique GNU/Linux", 10h, Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 5h, "Internship", Master 1 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, "Bases de mathématiques et probablité", 9h, Master1 in public health, Univ. Rennes 1, France
- Master: O. Dameron, "Programmation impérative en Python (2)", 3h Master 1 in public health, Univ. Rennes 1, France
- Master: O. Dameron, 20h, "Semantic Web and bio-ontologies", Master 2 in bioinformatics, Univ. Rennes 1, France
- Master: O. Dameron, 15h, "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.

11.2.4 Supervision

- PhD: Marine Louarn, *Intégration de données génomiques massives et hétérogènes, application aux mutations non-codantes dans le lymphome folliculaire*, defended November 26th 2020, supervised by A. Siegel, T. Fest (CHU) and O. Dameron. [29]
- PhD: Méline Wéry, *Methodology development in disease treatment projects.* , defended December 16th 2020, supervised by E. Becker, C. Bettembourg (Sanofi), O. Dameron, and A. Siegel. [30]
- PhD in progress : Hugo Talibart, Learning grammars with long-distance correlations on proteins, started in Nov. 2017, supervised by F. Coste and J. Nicolas.
- PhD in progress: Mael Conan, *Predictive approach to assess the genotoxicity of environmental contaminants during liver fibrosis*, started in Oct. 2017, supervised by S. Langouet and A. Siegel.
- PhD in progress : 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.
- 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: Nicolas Guillaudeux, *Compare gene structures to predict isoform transcripts*, started in Oct. 2018, supervised by O. Dameron, S. Blanquart and C. Belleannée
- 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: Olivier Dennler, *Modular functional characterization of ADAMTL and ADAMTSL protein families*, started in Oct. 2019, supervised by N. Theret, F. Coste and S. Blanquart.
- 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 : Mathieu 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 : 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 : Méziane Aite. *Identification de nouvelles combinaisons thérapeutiques dans les indications neurologiques*, started in Nov. 2020, supervised by O. Dameron and V. Lafon (Insiliance).

11.2.5 Juries

- Referee of PhD thesis (1): K. Guillaumier, University of Malta [F. Coste]
- Member of PhD thesis juries (6): M. Wery, Université de Rennes 1 [E. Becker, O. Dameron, A. Siegel], M. Louarn, Université de Rennes 1 [O. Dameron, A. Siegel], A. Aubert, Université Lyon 1 [N. Théret], R. Gazzotti, Université Côte d'Azur [O. Dameron], C. Dalloux, Université de Rennes 1 [O. Dameron], W. Delage, Université de Rennes 1 [O. Dameron, president]
- Referee of habilitation thesis (0):
- Member of habilitation thesis juries (1): V. Claveau, Université de Rennes 1 [O. Dameron]

11.2.6 Interns

- Internship, from May until Jun 2020. Supervised by F. Coste. Student: Malo Revel (Licence 3, ENS Rennes). Subject: Learning k,l-locally substitutable Multiple Context-Free Grammars
- Internship, from Feb until Aug 2020. Supervised by F. Coste and Y. Le Cunff. Student: Nicolas Buton (Master 2, Sorbonne University). Subject: Deep Learning Encoder Representations from Transformers: from Natural Languages to Protein Sequences Processing [31]
- Internship, from Jan until July 2020. Supervized by F. Moreews and E. Becker. Student : Quentin Delhon (Université de Rennes 1). Subject: Graph exploration algorithms to search for key regulators of metabolic pathways
- Internship, from Jan until July 2020. Supervized by E. Becker, O. Dameron and G. Rabut (IGDR). Student : Marc Melkonian (Université de Rennes 1). Subject: Réseaux biologiques : intégration de données d'interactions homogènes et hétérogènes
- Internship, from Jan until July 2020. Supervized by A. Siegel and H. Falentin. Student : Corentin Raphalen (Université de Rennes 1). Subject: Sélection de consortia microbiens optimaux pour la synthèse de vitamines dans les produits laitiers et fermentés
- Internship, from Jan until July 2020. Supervized by A. Siegel. Student : Pierre Gueracher (Université de Rennes 1). Subject: Variabilité des regroupements issus de l'analyse de données à cellule unique : impact du choix des paramètres et des marqueurs de types cellulaires
- Internship, from March until July 2020. Supervized by A. Belcour, S. Blanquart and A. Siegel. Student : Baptiste Ruiz (Agrocampus Paris). Subject: Modélisation de la diversité microbienne mesurée dans des expérimentations de fermenteurs
- Internship, from January until March 2020. Supervized by A. Siegel. Student : Adriana Concha (University of Chile). Subject: Integration of metagenomic datasets of Atacama desert at the metabolic scale.
- Internship, from April until June 2020. Supervized by O. Dameron. Student : Christophe Héligon (Université de Rennes 1). Subject: Integrating gene-phenotype association data using Semantic Web technologies.
- Internship, from May until July 2020. Supervized by O. Dameron. Student : Maxime Léger (Université de Rennes 1). Subject: Identification de modules au sein d'une collection de requêtes SPARQL.

11.3 Popularization

- Science en Cour[t]s (http://sciences-en-courts.fr/) Many of our current and former PhD students (M.Wéry, N. Guillaudeux, L. Bourneuf, A. Antoine-Lorquin, C. Bettembourg, J. Coquet, V. Dellannée, 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 web-site

11.3.1 Education

- Operation DECLIC (Dialogues Entre Chercheurs et Lycéens pour les Intéresser à la Construction des Savoirs). Lycée Descartes, Rennes. [N. Théret, M. Conan, O. Dennler]
- "Elles codent Elles créent" Several PhD students have been involved in creative Python learning sessions for female students in two high-schools. [M. Louarn, M. Wéry, A. Siegel]
- Formation des étudiants au capes informatique *Sensibilisation aux stéreotypes en informatique* [A. Siegel]

11.3.2 Interventions

- Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes de l'ENS Rennes, 18 novembre. *Atelier "déconstruction des stéreotypes"* [A. Siegel]
- Rendez-Vous des Jeunes Mathématiciennes et Informaticiennes de l'ENS, 12 décembre. *Informatique et biologie moléculaire: la place des modèles mathématiques* [A. Siegel]

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

[8] C. Frioux, E. Fremy, C. Trottier and A. Siegel. 'Scalable and exhaustive screening of metabolic functions carried out by microbial consortia'. In: *Bioinformatics* 34.17 (Sept. 2018), pp. i934–i943. DOI: 10.1093/bioinformatics/bty588. URL: https://hal.inria.fr/hal-01871600.

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

[10] S. Prigent, C. Frioux, S. M. Dittami, S. Thiele, A. Larhlimi, G. Collet, G. Fabien, J. Got, D. Eveillard, J. Bourdon, F. Plewniak, T. Tonon and A. Siegel. 'Meneco, a Topology-Based Gap-Filling Tool Applicable to Degraded Genome-Wide Metabolic Networks'. In: *PLoS Computational Biology* 13.1 (Jan. 2017), p. 32. DOI: 10.1371/journal.pcbi.1005276. URL: https://hal.inria.fr/hal-01449100.

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

12.2 Publications of the year

International journals

[12] F. Azar, K. Courtet, B. Dekky, D. Bonnier, O. Dameron, A. Colige, V. Legagneux and N. Théret. 'Integration of miRNA-regulatory networks in hepatic stellate cells identifies TIMP3 as a key factor in chronic liver disease'. In: *Liver International* 40.8 (Aug. 2020), pp. 2021–2033. DOI: 10.1111/liv.14476. URL: https://hal.archives-ouvertes.fr/hal-02549948.

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

[14] 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 (21st Feb. 2020), p. 100849. DOI: 10.1016/j.isci.2020.100849. URL: https://hal.inria.fr/hal-01943880.

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

[16] M. Conan, S. Langouët, A. Siegel and N. Théret. 'Innovative Approach to Predict DNA Adduct Formation of Environnemental Contaminants'. In: *Environmental and Molecular Mutagenesis* 61.S1 (2020), p. 58. DOI: 10.1002/em.22405. URL: https://hal.archives-ouvertes.fr/hal-02964170.

[17] S. Daval, K. Gazengel, A. Belcour, J. Linglin, A.-Y. Guillerm-Erckelboudt, A. Sarniguet, M. M. Manzanares-Dauleux, L. Lebreton and C. Mougel. 'Soil microbiota influences clubroot disease by modulating Plasmodiophora brassicae and Brassica napus transcriptomes'. In: *Microbial Biotechnology* 13.5 (2020), pp. 1648–1672. DOI: 10.1111/1751-7915.13634. URL: https://hal.inrae.fr/hal-02624824.

[18] 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* (Jan. 2020), pp. 1–24. DOI: 10.1016/j.margen.2020.100740. URL: https://hal.inria.fr/hal-02866117.

[19] M. Folschette, V. Legagneux, A. Poret, L. Chebouba, C. Guziolowski and N. Théret. 'A pipeline to create predictive functional networks: application to the tumor progression of hepatocellular carcinoma'. In: *BMC Bioinformatics* 21.1 (14th Jan. 2020), p. 18. DOI: 10.1186/s12859-019-3316-1. URL: https://hal.archives-ouvertes.fr/hal-02095930.

[20] C. Frioux, S. Dittami and A. Siegel. 'Using automated reasoning to explore the metabolism of unconventional organisms: a first step to explore host–microbial interactions'. In: *Biochemical Society Transactions* (7th May 2020), pp. 1–19. DOI: 10.1042/BST20190667. URL: https://hal.archives-ouvertes.fr/hal-02569935.

[21] 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* (30th Oct. 2020). DOI: 10.1093/nar/gkaa958. URL: https://hal.archives-ouvertes.fr/hal-02988562.

[22] 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 (24th July 2020). DOI: 10.3389/fmicb.2020.01707. URL: https://hal.sorbonne-universite.fr/hal-02929424.

[23] 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). DOI: 10.1093/database/baab004. URL: https://hal.archives-ouvertes.fr/hal-03144306.

[24] 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* (25th Nov. 2020), jnnp-2020–324647. DOI: 10.1136/jnnp-2020-324647. URL: https://hal.inria.fr/hal-03046771.

[25] 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.

[26] 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. Guedon. 'Extracellular vesicles produced by human and animal Staphylococcus aureus strains share a highly conserved core proteome'. In: *Scientific Reports* 10.1 (24th Apr. 2020), pp. 1–13. DOI: 10.1038/s41598-020-64952-y. URL: https://hal.inrae.fr/hal-02638124.

International peer-reviewed conferences

[27] H. Talibart and F. Coste. 'ComPotts: Optimal alignment of coevolutionary models for protein sequences'. In: JOBIM 2020 - Journées Ouvertes Biologie, Informatique et Mathématiques. Montpellier, France, 30th June 2020, pp. 1–8. URL: https://hal.inria.fr/hal-02862213.

Scientific book chapters

[28] N. Théret, J. Feret, A. Hodgkinson, P. Boutillier, P. Vignet and O. Radulescu. 'Integrative models for TGF-β signaling and extracellular matrix'. In: *Extracellular Matrix Omics*. Vol. 7. Biology of Extracellular Matrix. https://v6ediss.universite-lyon.fr/sylvie-ricard-blum--32497.kjsp, Dec. 2020, p. 17. DOI: 10.1007/978-3-030-58330-9_10. URL: https://hal.inria.fr/hal-02458073.

Doctoral dissertations and habilitation theses

[29] M. Louarn. 'Analysis and integration of heterogeneous large-scale genomics data'. Université Rennes 1, 26th Nov. 2020. URL: https://hal.inria.fr/tel-03111759.

[30] M. Wery. 'Identification of causal pathologic signature by multi-omic data integration'. Université de Rennes 1, 16th Dec. 2020. URL: https://hal.inria.fr/tel-03115751.

Other scientific publications

[31] N. Buton. 'Rapport de stage Master 2 : Apprentissage et étude de Transformer pour la classification en familles protéiques'. Sorbonne université, 3rd Sept. 2020. URL: https://hal.inria.fr/hal-03103334.

[32] C. Frioux, A. Belcour, M. Aite, A. Bretaudeau, F. Hildebrand and A. Siegel. *Metage2Metabo: metabolic complementarity applied to genomes of large-scale microbiotas for the identification of keystone species*. Montpellier, France, 30th June 2020. DOI: 10.1101/803056. URL: https://hal.inria.fr/hal-03151934.

12.3 Cited publications

[33] 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.

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

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

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

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

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

[39] M. Boutet, L. Gauthier, M. Leclerc, G. Gros, V. De Montpreville, N. Théret, E. Donnadieu and F. Mami-Chouaib. 'TGF-β signaling intersects with CD103 integrin signaling to promote T lymphocyte accumulation and antitumor activity in the lung tumor microenvironment'. In: *Cancer Research* 76.7 (2016), pp. 1757–69. DOI: 10.1158/0008-5472.CAN-15-1545. URL: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01282442.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] M.-A. Laurent, D. Bonnier, N. Théret, P. Tufféry and G. Moroy. 'In silico characterization of the interaction between LSKL peptide, a LAP-TGF-beta derived peptide, and ADAMTS1'. In: *Computational Biology and Chemistry* 61 (Apr. 2016), pp. 155–161. DOI: 10.1016/j.compbiolchem.2016.01.012. URL: https://hal.archives-ouvertes.fr/hal-02394687.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.